

申 报	系列：教师系列
	教学科研并重型
	专业：畜牧学
	职称：副教授

业绩成果材料

（申报人的业绩成果材料包括论文、科研项目、获奖以及其他成果等）

单 位（二级单位） 动物科学学院

姓 名 高亚辉

材料核对人：

单位盖章：

核对时间：

华南农业大学制

目 录

一、科研项目

1. 主持：关于《整合多组学数据鉴定母猪繁殖性状的关键结构变异》项目的立项通知（合同）及有关佐证材料..... 5
2. 主持：关于《整合多组学的猪重要经济性状功能解析与分子机制挖掘》项目的立项通知（合同）及有关佐证材料.. 17
3. 主持：关于《影响猪繁殖性状功能基因及调控元件筛选及功能解析》项目的立项通知（合同）及有关佐证材料.... 66
4. 主持：关于《基于三代测序鉴定影响猪繁殖性状的结构变异》项目的立项通知（合同）及有关佐证材料..... 84
5. 主参：关于《省级畜禽核心育种场生产性能测定，资源保种场保护和种畜禽质量监测》项目的立项通知（合同）及有关佐证材料 94
6. 主参：关于《国家生猪产业技术体系岗位科学家》项目的立项通知（合同）及有关佐证材料 106
7. 主参：关于《种猪基因型检测及分子育种分析服务》项目的立项通知（合同）及有关佐证材料 122
8. 主参：关于《畜牧业生产统计分析预警》项目的立项通知（合同）及有关佐证材料 132

二、论文、著作等

1. 检索证明 146
2. 以第一作者发表本专业论文情况
- 2.1. Deciphering genetic characteristics of South China and North China indigenous pigs through selection signatures 152

2.2. Transcriptomic profiling of gastrointestinal tracts in dairy cattle during lactation reveals molecular adaptations for milk synthesis.....	169
2.3. PigBiobank: a valuable resource for understanding genetic and biological mechanisms of diverse complex traits in pigs	183
2.4. A compendium of genetic regulatory effects across pig tissues	193
2.5. Genome-wide association analysis of heifer livability and early first calving in Holstein cattle	227
2.6. Comparative transcriptome in large-scale human and cattle populations	236
2.7. A multi-tissue atlas of regulatory variants in cattle	260
2.8. Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods	287
2.9. Single-cell transcriptomic and chromatin accessibility analyses of dairy cattle peripheral blood mononuclear cells and their responses to lipopolysaccharide	295
2.10. Towards the detection of copy number variation from single sperm sequencing in cattle.....	309
2.11. Genome-wide recombination map construction from single sperm sequencing in cattle	318
2.12. Functional annotation of regulatory elements in	

cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning	327
2.13. Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning...	339
2.14. Genome-wide association study of Mycobacterium avium Subspecies Paratuberculosis infection in Chinese Holstein	350
2.15. Heritability estimates for susceptibility to Mycobacterium avium ssp. paratuberculosis infection in Chinese Holstein cattle	360
3. 以通讯作者发表本专业论文情况	
3.1. Benchmarking 24 combinations of genotype pre-phasing and imputation software for SNP arrays in pigs	366
3.2. 基于全基因组重测序检测中国地方猪的体型选择信号	376

【佐证材料切记与目录页所列页码对应, 不要用图片格式的材料进行打印。】

国家自然科学基金资助项目批准通知

(包干制项目)

高亚辉 先生/女士：

根据《国家自然科学基金条例》、相关项目管理办法规定和专家评审意见，国家自然科学基金委员会（以下简称自然科学基金委）决定资助您申请的项目。项目批准号：32402714，项目名称：整合多组学数据鉴定母猪繁殖性状的关键结构变异，资助经费：30.00万元，项目起止年月：2025年01月至2027年12月，有关项目的评审意见及修改意见附后。

请您尽快登录科学基金网络信息系统（<https://grants.nsfc.gov.cn>），**认真阅读《国家自然科学基金资助项目计划书填报说明》并按要求填写《国家自然科学基金资助项目计划书》（以下简称计划书）**。对于有修改意见的项目，请您按修改意见及时调整计划书相关内容；如您对修改意见有异议，须在电子版计划书报送截止日期前向相关科学处提出。

请您将电子版计划书通过科学基金网络信息系统（<https://grants.nsfc.gov.cn>）提交，由依托单位审核后提交至自然科学基金委。自然科学基金委审核未通过者，将退回的电子版计划书修改后再行提交；审核通过者，打印纸质版计划书（一式两份，双面打印）并在项目负责人承诺栏签字，由依托单位在承诺栏加盖依托单位公章，且将申请书纸质签字盖章页订在其中一份计划书之后，一并报送至自然科学基金委项目材料接收工作组。纸质版计划书应当保证与审核通过的电子版计划书内容一致。**自然科学基金委将对申请书纸质签字盖章页进行审核，对存在问题的，允许依托单位进行一次修改或补齐。**

向自然科学基金委提交电子版计划书、报送纸质版计划书并补交申请书纸质签字盖章页截止时间节点如下：

1. **2024年9月9日16点**：提交电子版计划书的截止时间；
2. **2024年9月16日16点**：提交修改后电子版计划书的截止时间；
3. **2024年9月23日**：报送纸质版计划书（一式两份，其中一份包含申请书纸质签字盖章页）的截止时间。
4. **2024年10月8日**：报送修改后的申请书纸质签字盖章页的截止时间。

请按照以上规定及时提交电子版计划书，并报送纸质版计划书和申请书纸质签字盖章页，逾期不报计划书或申请书纸质签字盖章页且未说明理由的，视为自动放弃接受资助；未按要求修改或逾期提交申请书纸质签字盖章页者，将视情况给予暂缓拨付经费等处理。

附件：项目评审意见及修改意见表

国家自然科学基金委员会

2024年8月23日

附件：项目评审意见及修改意见表

项目批准号	32402714	项目负责人	高亚辉	申请代码1	C1702
项目名称	整合多组学数据鉴定母猪繁殖性状的关键结构变异				
资助类别	青年科学基金项目	亚类说明			
附注说明					
依托单位	华南农业大学				
直接费用	30.00 万元	起止年月	2025年01月 至 2027年12月		
<p>通讯评审意见：</p> <p><1>具体评价意见：</p> <p>一、请评述该申请项目是否面向经济社会发展需要或国家需求背后的基础科学问题。请详细阐述判断理由。</p> <p>提高母猪繁殖力一直都是生猪育种重要目标之一，基于分子标记的基因组选择育种加速了遗传改良的进程。项目针对总产仔数和产活仔数等性状探讨SV影响繁殖性状的潜在机制，符合国家需求背后的科学问题。</p> <p>二、请评述申请项目所提出的科学问题的创新性与预期成果的科学价值。</p> <p>项目整合多组学筛选与繁殖性状相关的结构变异，可以应用于分子育种，具有较高的理论价值，预期成果可为实际生产积累科学数据，具有重要科学价值。</p> <p>三、请评述申请人的创新潜力与研究方案可行性；如有可能，请对完善研究方案提出建议。</p> <p>基于项目申请人的前期科学研究基础，以及项目的研究方案设计，针对总产仔数和产活仔数，结合基因、转录和蛋白等多个层面挖掘与繁殖性状相关的结构变异，探讨SV影响繁殖性状的潜在机制，申请人的研究思路具有较强的创新型，研究方案可行。申请者知识背景和前期研究基础较好，选题和技术手段创新，发表过高水平论文，鉴定的结构变异预期可用于分子育种。</p> <p>四、其他建议</p> <p><2>具体评价意见：</p> <p>一、请评述该申请项目是否面向经济社会发展需要或国家需求背后的基础科学问题。请详细阐述判断理由。</p> <p>本项目的选题能围绕猪产业技术的瓶颈问题进行，持续提高母猪繁殖力是生猪产业提质增效高质量发展的关键。母猪繁殖性状为低遗传力性状，传统遗传改良手段对其提高有限。应用新技术高效挖掘繁殖性状相关的分子标记并用于基因组选择，可有效加速繁殖性状的遗传进展。鉴于结构变异在分子标记中的重要性，并与单核苷酸多态性互为补充，申请者能够聚焦提出科学问题：结构变异如何影响母猪繁殖性状，落在基因组重复区域的结构变异是否与繁殖性状相关？</p> <p>二、请评述申请项目所提出的科学问题的创新性与预期成果的科学价值。</p> <p>项目提出的科学问题的创新性较强，项目针对目前广泛应用的大白猪总产仔数和产活仔数两个繁殖性状，充分利用三代测序、转录组测序和iTRAQ技术从基因组、转录组和蛋白组多层次挖掘与母猪繁殖性状相关的结构变异，解析结构变异—转录—蛋白—繁殖性状表型的过程，探讨SV影响母猪繁殖性状的潜在机制。预期其成果的价值能够深入探讨 SV 影响母猪繁殖性状的潜在机制，能为高繁殖力种猪培育提供分子素材和理论依据。</p> <p>三、请评述申请人的创新潜力与研究方案可行性；如有可能，请对完善研究方案提出建议。</p> <p>申请人具备良好的科学素养，工作专注且功底深厚，发表了高质量的论文，研究方案详实可行，技术路线的逻辑思路清晰。</p> <p>四、其他建议</p>					

<3>具体评价意见：

一、请评述该申请项目是否面向经济社会发展需要或国家需求背后的基础科学问题。请详细阐述判断理由。

种业振兴是十四五期间国家的重点任务，生猪产业更是重中之重，当前我国母猪的繁殖性能与世界养殖发达国家还存在差距，亟需进行改良。而生物技术的应用，尤其是全基因组选择技术的使用，将大大提高育种效率。而全基因组选择的基础是分子标记，通过全基因组SV鉴定，可以有效提供分子标记，同时从SV角度鉴定有望揭示与繁殖形状相关的遗传变异。因此，本研究的内容目标符合目标导向类基础研究。

二、请评述申请项目所提出的科学问题的创新性与预期成果的科学价值。

本项目针对我国母猪繁殖性能需要进一步提高的问题具有实际需要性，并据此提出充分利用现代生物技术，尤其是全基因选择技术进行解决的想法具有很强的实用性。作者创新性的从SV角度开发分析标记，以及探究SV对繁殖形状的影响，创新性强，预期研究成果对揭示与繁殖形状相关的遗传变异或对此领域的研究及应用开发具有较强的价值。

三、请评述申请人的创新潜力与研究方案可行性；如有可能，请对完善研究方案提出建议。

项目研究思路很好，团队研究基础强，个人参与的项目丰富，积累了大量研究经验。同时本项目研究方案设计合理，实施方法整体介绍较为清楚，研究方法科学，可行性强。建议在基因组数据分析SV步骤时增加个体，以充分发掘SV。

四、其他建议

基金撰写还有优化空间，前言中存在少量语言逻辑问题，对大群验证方法的描述有待细化。

修改意见：

生命科学部

2024年8月23日



项目批准号	32402714
申请代码	C1702
归口管理部门	
依托单位代码	51064208A0499-0932



324027141001451

国家自然科学基金 资助项目计划书 (包干制项目)

资助类别：青年科学基金项目

亚类说明：

附注说明：

项目名称：整合多组学数据鉴定母猪繁殖性状的关键结构变异

资助经费：30万元 执行年限：2025.01-2027.12

负责人：高亚辉 BRID：01602.00.68813

通讯地址：广州市天河区五山路483号

邮政编码：510642 电 话：020-85282019

电子邮件：gyhalvin@gmail.com

依托单位：华南农业大学

联系人：唐家林 电 话：020-85280070

填表日期：2024年08月26日

国家自然科学基金委员会制



国家自然科学基金资助项目计划书填报说明 （包干制项目）

- 一、项目负责人收到《国家自然科学基金资助项目批准通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办​​法和新修订的《国家自然科学基金资助项目资金管理办法》（以下简称《资金管理办法》，请查阅国家自然科学基金委员会官方网站首页“政策法规”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金资助项目计划书》（以下简称《计划书》）。
- 二、填写《计划书》时要科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行、检查和验收的依据。
- 三、《计划书》各部分填写要求如下：
 - （一）简表：由系统自动生成。
 - （二）摘要及关键词：各类获资助项目都应当填写中、英文摘要及关键词。
 - （三）正文：
 1. 青年科学基金项目、青年学生基础研究项目：如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中上述栏目明确要求调整研究期限或研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。
 2. 国家杰出青年科学基金项目和优秀青年科学基金项目按下列提纲撰写：
 - （1）研究方向；
 - （2）结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；
 - （3）研究内容、研究方案及预期目标（限两个页面）；
 - （4）年度研究计划；
- 四、资助经费相关要求：
 1. 资助经费批准时不再区分直接费用和间接费用。
 2. 项目负责人在提交计划书时需签署承诺书，承诺尊重科研规律，弘扬科学家精神，遵守科研伦理道德和作风学风诚信要求，认真开展科学研究工作；承诺项目经费全部用于与本项目研究工作相关的支出，不得用于与本项目研究无关的支出。
 3. 项目负责人提交计划书时，无需编制项目预算。项目资金由项目负责人自主决定使用，按照《资金管理办法》第九条规定的开支范围列支。有关管理费用的补助支出，由依托单位根据实际管理需要，在充分征求项目负责人意见基础上合理确定。绩效支出由项目负责人根据实际科研需要和相关薪酬标准自主确定，依托单位按照工资制度进行管理。对于青年学生基础研究项目，支付给项目负责人本人的劳务费用，应符合相关比例要求。其余用途经费无额度限制，由项目负责人根据实际需要自主决定使用。



4. 项目结题时，项目负责人根据实际使用情况编制项目经费决算，经依托单位财务、科研管理部门审核后，报自然科学基金委。依托单位应当在单位内部公开非涉密项目立项、主要研究人员、资金使用（重点是间接费用、外拨资金、结余资金使用等）、决算、大型仪器设备购置以及项目研究成果等情况，接受内部监督。
5. 自然科学基金委结合项目管理，对经费使用情况和依托单位管理情况定期开展抽查。



简表

项目负责人信息	姓 名	高亚辉	性 别	男	出生年月	1989年09月	民 族	汉族
	学 位	博士			职称	讲师		
	是否在站博士后	否			电子邮件	gyhalvin@gmail.com		
	电 话	020-85282019			个人网页			
	工 作 单 位	华南农业大学						
	所 在 院 系 所	动物科学学院						
依托单位信息	名 称	华南农业大学					代码	51064208A0499
	联 系 人	唐家林			电子邮件	kyc.jhk@scau.edu.cn		
	电 话	020-85280070			网站地址	http://kjc.scau.edu.cn/		
合作单位信息	单 位 名 称							
项目基本信息	项 目 名 称	整合多组学数据鉴定母猪繁殖性状的关键结构变异						
	资 助 类 别	青年科学基金项目				亚 类 说 明		
	附 注 说 明							
	申 请 代 码	C1702:家畜种质资源与遗传育种学						
	基 地 类 别							
	执 行 年 限	2025.01-2027.12						
	资 助 经 费	30万元						



项目摘要

中文摘要:

持续提高母猪繁殖力是生猪产业提质增效高质量发展的关键。母猪繁殖性状为低遗传力性状，传统遗传改良手段对其提高有限。应用新技术高效挖掘繁殖性状相关的分子标记并用于基因组选择，可有效加速繁殖性状的遗传进展。鉴于结构变异在分子标记中的重要性，并与单核苷酸多态性互为补充，提出科学问题：结构变异如何影响母猪繁殖性状，落在基因组重复区域的结构变异是否与繁殖性状相关？本项目拟针对大白猪总产仔数和产活仔数两个繁殖性状，利用三代测序、转录组测序和iTRAQ技术从基因组、转录组和蛋白组多层面挖掘与母猪繁殖性状相关的结构变异，解析结构变异→转录→蛋白→繁殖性状表型的过程，探讨SV影响母猪繁殖性状的潜在机制。本项目鉴定的结构变异可用于分子育种，为高繁殖力种猪培育提供分子素材和理论依据。

Abstract:

Continuously improving the reproductive capacity of sows is the key to improving the quality, efficiency and high-quality development of the pig industry. Sow reproductive traits are low heritability traits, and traditional genetic improvement methods can only improve them to a limited extent. The application of new technologies to efficiently mine molecular markers related to reproductive traits and apply them to genome selection can effectively accelerate the genetic progress of reproductive traits. Based on the premise that structural variation is an important molecular marker and complementary to single nucleotide polymorphisms, a scientific question is raised: How does structural variation affect sow reproductive traits and are structural variants falling in repetitive regions of the genome associated with reproductive traits? This project intends to target the two reproductive traits of Large White pigs: Total Number Born and Number Born Alive. By using third-generation sequencing, transcriptome sequencing and iTRAQ technology, we can fully mine structural variations related to sow reproductive traits from the genome, transcriptome and proteome levels, and explore the potential mechanism of SV affecting sow reproductive traits. The structural variations identified in this project can be used in molecular breeding to provide molecular materials and a theoretical basis for the breeding of high-reproductive pigs.

关键词(用分号分开): 猪; 数量性状; 繁殖性状; 结构变异; 三代测序

Keywords(用分号分开): Pig; Quantitative traits; Reproductive traits; Structural variation (SV); Third-generation sequencing



报告正文

研究内容和研究目标按照申请书执行。



国家自然科学基金项目负责人、依托单位承诺书

国家自然科学基金项目负责人承诺书

本人郑重承诺：我接受国家自然科学基金的资助，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等规定，和国家自然科学基金委员会关于资助项目管理、项目资金管理等各项规章，在《计划书》填写及项目执行过程中：

（一）按照《批准通知》《国家自然科学基金资助项目计划书填报说明》的要求填写《计划书》，未自行降低、更改目标任务或约定要求，或缩减研究（研制）内容；

（二）树立“红线”意识，严格履行科研合同义务，按照《计划书》负责实施本项目（批准号：32402714），切实保证研究工作时间，按时报送有关材料，及时报告重大情况变动，不违规将科研任务转包、分包他人，不以项目实施周期外或不相关成果充抵交差；

（三）遵守科研诚信、科技伦理规范和学术道德，认真开展研究工作，对资助项目发表的论著和取得的科研成果按规定进行标注，不在非本项目资助的成果或其他无关成果上标注本项目批准号，反对无实质学术贡献者“挂名”，不在成果署名、知识产权归属等方面侵占他人合法权益，并如实报告本人及项目组成员发生的违背科研诚信要求的任何行为；

（四）尊重科研规律，弘扬科学家精神，严谨求实，追求卓越，反对浮夸浮躁、投机取巧，不人为夸大学术或技术价值，不传播未经科学验证的现象和观点；

（五）将项目资金全部用于与本项目研究工作相关的支出，并结合科研活动需要，科学合理安排项目资金支出进度；

（六）做好项目组成员的教育和管理，确保遵守以上相关要求。

如违背上述承诺，本人愿接受国家自然科学基金委员会和相关部门做出的各项处理决定。

项目负责人（签字）：

年 月 日

国家自然科学基金项目依托单位承诺书

我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究队伍的稳定和研究项目实施所需的条件，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》《科技伦理审查办法（试行）》等规定，和国家自然科学基金委员会有关资助项目管理、项目资金管理、科研诚信管理和科技伦理管理等各项规定，并督促实施。

依托单位（公章）

年 月 日



国家自然科学基金资助项目签批审核表

本 栏 目 由 自 然 科 学 基 金 委 填 写	科学处审查意见：		负责人（签章）： 年 月 日
	科学部审查意见：		负责人（签章）： 年 月 日

课题编号：2024YFF1000101

密 级：公开

国家重点研发计划 课题任务书

课题名称：猪重要经济性状多组学解析及精准选育

所属项目：整合多组学信息解析畜禽重要经济性状及精准选育

所属专项：农业生物重要性状形成与环境适应性基础研究

项目牵头承担单位：中国农业科学院北京畜牧兽医研究所

课题承担单位：中国农业大学

课题负责人：申振才

执行期限：2024 年 12 月 至 2027 年 11 月

中华人民共和国科学技术部制

2024 年 12 月 09 日

0003YF 2024YFF1000101 2024-12-09 16:14:34



填写说明

- 一、任务书甲方即 项目牵头承担单位，乙方即 承担单位。
- 二、任务书 在“国家科技计划管理信息系统公共服务平台”，按照系统提示在线填写。
- 三、任务书中的单位名称， 按规范全称填写，并与单位公章一致。
- 四、任务书要求提供乙方与所有参加单位的合作协议， 对原件 行扫描后在线提交。
- 五、任务书中文字 用宋体小四号字填写。
- 六、凡不填写内容的栏目， 用“无”表示。
- 七、乙方完成任务书的在线填写，提交甲方审核确认后，用 A4 纸在线打印、装、签章。一式八份报 项目牵头承担单位签章，其中 承担单位一份，人一份，作为 项目任务书 附件六份。
- 八、如 项目下仅 一个， 任务书只 填报 算 分。
- 九、涉密 在“国家科技计划管理信息系统公共服务平台”下 任务书的电子版模板，按保密要求离线填写、报 。
- 十、《 项目申报书》和《 项目任务书》是本任务书填报的 要依据，任务书填报不得 低考核指标，不得自行对主要研究内容作大的 整。《 项目申报书》、《 项目任务书》和本任务书将共同作为 程管理、综合绩效评价（ 收）和监督 估的 要依据。



课题基本信息表

名称		猪 要经济性状多组学解析及精准”育							
编号		2024YFF1000101							
所属 目		整合多组学信息解析畜禽 要经济性状及精准”育							
所属专		农业生物 要性状形成与环境 应性基础研究							
密级		■公开 □秘密 □机密		单位总数		3			
类型		■基础前沿□ 大共性关 技术□应用示范研究□其他							
活动类型		■基础前沿□应用研究□ 发展							
研究所属学科		畜牧、兽医科学 畜牧学							
成果应用的主要国民经济行业		农、林、牧、渔业 畜牧业 牲畜 养 猪的 养							
的社会经济目标		农林牧渔业发展 畜牧业							
经 算		总 求 368.00 万元，其中中央 政专 求 368.00 万元							
周期节点		始时		2024 年 12 月		结束时		2027 年 11 月	
		实施周期		共 36 个月) 中期时 点		2026 年 05 月	
承担单位	单位名称		中国农业大学			单位法定 代表人姓名		孙其信	
	单位性		大专人校			组织机构代码		12100000400018162G	
	单位主管					属关系		中央	
	单位所属地区		北京市		北京市		海淀区		
	信地址		北京市海淀区圆明园西 2 号			政编码		100193	
	单位开户名称		中国农业大学						
	开户 行 (全称)		中国建 行北京上地支行			汇入地点		北京市 北京 市	



	行 号	11001045300053003131			行机构代码		105100005078
人	姓 名	申振才	性 别	<input checked="" type="checkbox"/> 男 <input type="checkbox"/> 女		出生日期	1980-12-19
	件类型	份	件号码	372926198012197731			
	所在单位	中国农业大学					
	最 学位	<input checked="" type="checkbox"/> 博士 <input type="checkbox"/> 硕士 <input type="checkbox"/> 学士 <input type="checkbox"/> 其他					
	职 称	<input type="checkbox"/> 正 级 <input checked="" type="checkbox"/> 副 级 <input type="checkbox"/> 中级 <input type="checkbox"/> 初级 <input type="checkbox"/> 其他				职 务	副主任
	电子 箱	zhencai688@sina.com			移动电	13466405191	
联系 人	姓 名	申振才	电子 箱	zhencai688@sina.com			
	固定电	01062736511	移动电	13466405191			
	件类型	份	件号码	372926198012197731			
务 人	姓 名	火晓	电子 箱	huoxx@cau.edu.cn			
	固定电	101-62737512	移动电	18514235893			
	件类型	份	件号码	620121198205270020			
其他 参与 单位	序号	单位名称		单位性		组织机构代码	
	1	中国农业科学人北京畜牧兽医研究所		事业型研究单位		12100000400882305E	
	2	华南农业大学		大专人校		124400004554165634	
参 加 人 数	<u>10</u> 人。其中：		级职称 <u>4</u> 人，中级职称 <u>0</u> 人，初级职称 <u>0</u> 人，其他 <u>6</u> 人；				
			博士学位 <u>4</u> 人，硕士学位 <u>3</u> 人，学士学位 <u>3</u> 人，其他 <u>0</u> 人。				
简介 (500 字以 内)	围绕猪主导品种，对“常规育种+生物技术+信息技术+人工智能”育种4.0时代的挑战，本对以下署研究内容：相关性状测定指标向智能化、目标化、多元化变；解决猪基因源利用效率和技术；开发基于多组学数据解析的全基因组精准”育技术。以提猪肉生产效率和肉品平衡育种为目标，拓展生产性能测定范围，构建猪生产效率与肉品相关性状的精细表型图，开展要表型的自动化、精准化、多元化、及数据清洗、标准化研究。为解析生产效率和肉品相关表型，整合多组学数据，基于先统)学方法或人工智能策略，对精细表型和分子表型行深度挖掘，筛”影响猪肉生产效率和肉品表型的关控变异、基因、代产物、微生物、生物学背景及其控机制。为提精准”育效率，研发基于线性模型的多组学模块化基因组”择算法，及基						



	于多组学信息的机器学习的整合多组学信息的方法，提升畜禽”择准确性，实现精准”育，加快育种 展。
--	---



一、目标及考核指标、考核方式/方法

课题目标、预期成果与考核指标表

课题目标 ¹	预期成果			考核指标 ²				考核方式 (方法)及 评价手段 ⁴
	预期成果名称		预期成果类型	指标 名称	立项时已 有指标值/ 状态	中期指标 值/状态 ³	完成时指标 值/状态	
围绕猪重要经济性状,拓展生产性能测定范围,获得与猪生产效率 and 肉质相关的精细表型图谱,实现重要表型的智能化、精准化采集及数据清洗和标准化;整合多组学数据,对精细表型和分子表型进行深度挖掘,筛选影响猪肉生产	主要成果	1	挖掘与肉类生产密切相关的关键遗传调控位点 8-10 个 <input type="checkbox"/> 新理论 <input type="checkbox"/> 新原理 <input type="checkbox"/> 新产品 <input checked="" type="checkbox"/> 新技术 <input type="checkbox"/> 新方法 <input type="checkbox"/> 关键部件 <input type="checkbox"/> 数据库 <input checked="" type="checkbox"/> 软件 <input type="checkbox"/> 应用解决方案 <input type="checkbox"/> 实验装置/系统 <input type="checkbox"/> 临床指南/规范 <input type="checkbox"/> 工程工艺 <input type="checkbox"/> 标准 <input checked="" type="checkbox"/> 论文 <input type="checkbox"/> 发明专利 <input type="checkbox"/> 其他	挖掘与肉类生产密切相关的关键遗传调控位点 8-10 个	无	4-5 个	8-10 个	论文见刊或录用通知、专利、科技报告或提交至国家共享数据库
		2	挖掘与肉类生产密切相关的候选基因 2-4 个 <input type="checkbox"/> 新理论 <input type="checkbox"/> 新原理 <input type="checkbox"/> 新产品 <input checked="" type="checkbox"/> 新技术 <input checked="" type="checkbox"/> 新方法 <input type="checkbox"/> 关键部件 <input type="checkbox"/> 数据库 <input type="checkbox"/> 软件 <input type="checkbox"/> 应用解决方案 <input type="checkbox"/> 实验装置/系统 <input type="checkbox"/> 临床指南/规范 <input type="checkbox"/> 工程工艺 <input type="checkbox"/> 标准 <input checked="" type="checkbox"/> 论文 <input type="checkbox"/> 发明专利 <input type="checkbox"/> 其他	挖掘与肉类生产密切相关的候选基因 2 个	无	1 个	2-4 个	论文见刊或录用通知、专利、科技报告或提交至国家共享数据库
		3	建立融合多组学信息的基因组选育模型和算法 1-2 个 <input type="checkbox"/> 新理论 <input type="checkbox"/> 新原理 <input type="checkbox"/> 新产品 <input checked="" type="checkbox"/> 新技术 <input checked="" type="checkbox"/> 新方法 <input type="checkbox"/> 关键部件 <input type="checkbox"/> 数据库 <input checked="" type="checkbox"/> 软件 <input type="checkbox"/> 应用解决方案 <input type="checkbox"/> 实验装置/系统 <input type="checkbox"/> 临床指南/规范 <input type="checkbox"/> 工程工艺 <input type="checkbox"/> 标准 <input checked="" type="checkbox"/> 论文 <input type="checkbox"/> 发明专利 <input type="checkbox"/> 其他	建立融合多组学信息的基因组选育模型和算法 2 个	无	1 个	2 个	论文见刊或录用通知、软件、行业专家评议、行业专家现场评估或第三方评价



效率和肉品质表型的关键调控变异、基因和高维分子表型，解析其生物学背景及调控机制；研发基于线性模型的多组学模块化基因组选择算法，研发整合多组学信息的机器学习模型和算法，提升基因组选择准确性，实现猪的精准选育。	4	发表高水平学术论文 3-4 篇	<input type="checkbox"/> 新理论 <input type="checkbox"/> 新原理 <input type="checkbox"/> 新产品 <input type="checkbox"/> 新技术 <input type="checkbox"/> 新方法 <input type="checkbox"/> 关键部件 <input type="checkbox"/> 数据库 <input type="checkbox"/> 软件 <input type="checkbox"/> 应用解决方案 <input type="checkbox"/> 实验装置/系统 <input type="checkbox"/> 临床指南/规范 <input type="checkbox"/> 工程技术 <input type="checkbox"/> 标准 <input checked="" type="checkbox"/> 论文 <input type="checkbox"/> 发明专利 <input type="checkbox"/> 其他	发表高水平学术论文 3-4 篇	无	1-2 篇	3-4 篇	论文见刊或录用通知
	5	为育种企业或专家提供重要经济性状基因位点 3 个及以上	<input type="checkbox"/> 新理论 <input type="checkbox"/> 新原理 <input type="checkbox"/> 新产品 <input checked="" type="checkbox"/> 新技术 <input checked="" type="checkbox"/> 新方法 <input type="checkbox"/> 关键部件 <input type="checkbox"/> 数据库 <input checked="" type="checkbox"/> 软件 <input type="checkbox"/> 应用解决方案 <input type="checkbox"/> 实验装置/系统 <input type="checkbox"/> 临床指南/规范 <input type="checkbox"/> 工程技术 <input type="checkbox"/> 标准 <input checked="" type="checkbox"/> 论文 <input checked="" type="checkbox"/> 发明专利 <input type="checkbox"/> 其他	为育种企业或专家提供重要经济性状基因位点 3 个及以上	无	1 个及以上	3 个及以上	行业专家评议、行业专家现场评估、第三方评价或育种企业证明
科技报告考核指标	序号		报告类型 ⁵	数量	提交时间		公开类别及时限 ⁶	
	1		年度技术进展报告	2	2025 年 11 月、2026 年 11 月		延期公开，3 年	
	2		中期技术进展报告	1	2026 年 5 月		延期公开，3 年	
	3		最终科技报告	1	2027 年 11 月		延期公开，3 年	
其他目标与考核指标：无								



备注：

1. **“课题目标”**，应从以下方面明确描述：（1）研发主要针对什么问题和需求；（2）将要解决哪些科学问题、突破哪些核心/共性/关键技术；（3）预期成果；（4）成果将以何种方式应用在哪些领域/行业/重大工程等，并拟在科技、经济、社会、环境或国防安全等方面发挥何种的作用和影响。（5）所列主要成果原则上不超过 5 项，如有其他重要成果放在“其他”成果中表述。
2. **“考核指标”**，指相应成果的数量指标、技术指标、质量指标、应用指标和产业化指标等，其中，数量指标可以为专利、产品等的数量，论文代表作应注重质量，不以数量作为评价标准；技术指标可以为关键技术、产品的性能参数等；质量指标可以为产品的耐震动、高低温、无故障运行时间等；应用指标可以为成果应用的对象、范围和效果等；产业化指标可以为成果产业化的数量、经济效益等。同时，对各项考核指标需填写立项时已有的指标值/状态以及课题完成时要到达的指标值/状态。同时，考核指标也应包括支撑和服务其他重大科研、经济、社会发展、生态环境、科学普及需求等方面的直接和间接效益。如对国家重大工程、社会民生发展等提供了关键技术支撑，成果转让并带动了环境改善、实现了销售收入等。若某项成果属于开创性的成果，立项时已有指标值/状态可填写“无”，若某项成果在立项时已有指标值/状态难以界定，则可填写“/”。
3. **“中期指标”**，各专项根据管理特点，确定是否填写，鼓励阶段目标明确的项目课题填写中期指标。
4. **“考核方式方法”**，应提出符合相关研究成果与指标的具体考核技术方法、测算方法等。
5. **“科技报告类型”**，包括项目综合绩效评价（验收）前撰写的全面描述研究过程和技术内容的最终科技报告、项目年度或中期检查时撰写的描述本年度研究过程和进展的年度技术进展报告以及在项目实施过程中撰写的包含科研活动细节及基础数据的专题科技报告（如实验报告、试验报告、调研报告、技术考察报告、设计报告、测试报告等）。其中，每个项目在综合绩效评价（验收）前应撰写一份最终科技报告；研究期限超过 2 年（含 2 年）的项目，应根据管理要求，每年撰写一份年度技术进展报告；每个项目可根据研究内容、期限和经费强度，撰写数量不等的专题科技报告。科技报告应按国家标准规定的格式撰写。
6. **“公开类别及时限”**，公开项目科技报告分为公开或延期公开，内容需要发表论文、申请专利、出版专著或涉及技术诀窍的，可标注为“延期公开”。需要发表论文的，延期公开时限原则上在 2 年（含 2 年）以内；需要申请专利、出版专著的，延期公开时限原则上在 3 年（含 3 年）以内；涉及技术诀窍的，延期公开时限原则上在 5 年（含 5 年）以内。涉密项目科技报告按照有关规定管理。





二、课题研究内容、研究方法及技术路线

（一）课题的主要研究内容

拟解决的关键科学问题、关键技术问题，针对这些问题拟开展的主要研究内容，限1000字以内。

1. 拟解决的关键科学问题或关键技术问题：

（1）以提高猪肉生产效率与肉品质的平衡育种为目标，拓展现有生产性能测定范围，构建智能化、精细化的猪生产效率与肉品质相关表型采集测定体系；

（2）针对猪分子表型库不完善的问题，通过多组学整合分析方法，构建猪多模态分子表型图谱，为遗传机制解析与精准选育提供重要基础；

（3）针对猪经济性状功能位点与基因缺乏的问题，开发和优化整合方法，筛选与猪生产效率和肉品质性状相关的关键位点、基因及分子表型；

（4）针对目前基因组遗传变异无法完全解释部分性状遗传力的问题，开发充分利用多组学信息的选育新模型与算法，为畜禽选育提供更精准、高效的手段。

2. 本课题的主要研究内容如下：

（1）研究猪重要经济性状的智能化和精细化测定方法，结合传统测定方法，利用现代化表型数据采集设备和基于人工智能的数据分析方法，获得准确的表型数据，完善源头性能测定数据，实现猪生产效率和肉品质相关表型的智能化和精细化测定及新表型挖掘与评估。

（2）基于多组学数据，挖掘转录组（如基因表达量、可变剪接等）、代谢组及宏基因组（微生物组成与丰度）的多模态分子表型，构建多模态分子表型图谱，识别重要经济性状的关键分子表型。

（3）对猪重要经济性状进行多组学解析，例如鉴定影响猪生产效率和肉品质相关性状的关键基因组变异，结合二代和三代转录组精确构建基因和转录本表达图谱，鉴定相关蛋白和代谢产物，整合多组学数据筛选关键调控位点、基因和机制，明确其生物学效应并评估育种价值等。

（4）采用各种方法构建融合多组学信息的精准选育体系，例如在 BLUP 框架下，开发多组学模块化基因组选择的线性模型和算法，鉴于线性模型难以捕捉复杂的非线性关系，进一步考虑融合多组学信息的非线性模型和算法开发。



（二）课题采取的研究方法

针对课题研究拟解决的问题，拟采用的方法、原理、机理、算法、模型等限 1000 字以内。

本课题拟采用的方法、原理、机理、算法、模型等具体包括：

1. 猪生产效率和肉品质相关表型的智能化和精细化测定及新表型挖掘与评估

结合传统测定方法，获得准确的表型数据。背膘厚、眼肌面积及肌内脂肪酸含量等使用超声波扫描仪等测定；对难测（如胴体、日增重）或人工测量效率低的性状（如总产仔数），采用深度相机等设备监测，引入 YOLO、mask-RCNN、UNet 等深度学习方法，提升效率和准确性；利用高效液相色谱技术，精细测定猪肉脂肪酸及风味物质含量；采用弹性网络、组 Lasso 等对高维表型进行特征提取，挖掘新表型，综合评估其科学性。

2. 多模态分子表型图谱构建

基于多组学数据，挖掘转录组、蛋白组、代谢组及宏基因组的多模态分子表型。利用 GCTA 等方法计算各分子表型的遗传力，并运用 LDSC、MTAG 等方法评估分子表型间的遗传相关性。采用统计或机器学习方法（如 RFE、mRMR、CCA、互信息法等）进行特征选择，识别重要经济性状的关键分子表型。

3. 猪重要经济性状的多组学解析

（1）鉴定影响猪生产效率和肉品质相关性状的关键基因组变异：利用 PacBio HiFi 构建图形泛基因组，精准检测基因组单碱基变异和结构变异。结合全基因组关联分析，定位重要变异并进行功能注释与验证，揭示遗传突变和结构变异对猪生产效率和肉品质性状的调控效应及其相对贡献。

（2）结合二代和三代转录组精确构建基因和转录本表达图谱：利用二代高通量转录组测序生成短读段数据，获得全基因组范围的基因和转录本表达量。通过三代全长转录组测序产生的全长读段，准确捕获完整转录本结构，精确解析剪接、基因融合及复杂转录事件。

（3）鉴定相关蛋白和代谢产物：利用蛋白质组学技术（如质谱分析）识别和定量蛋白质。通过 MaxQuant 和 Proteome Discoverer 比对蛋白质数据库明确其特征，解析其功



能和作用途径。利用代谢组学技术（如 LC-MS、GC-MS）测定样本中代谢物，揭示其丰度和代谢通路的动态变化。通过 MetaboAnalyst 提供代谢产物的注释和功能解析。

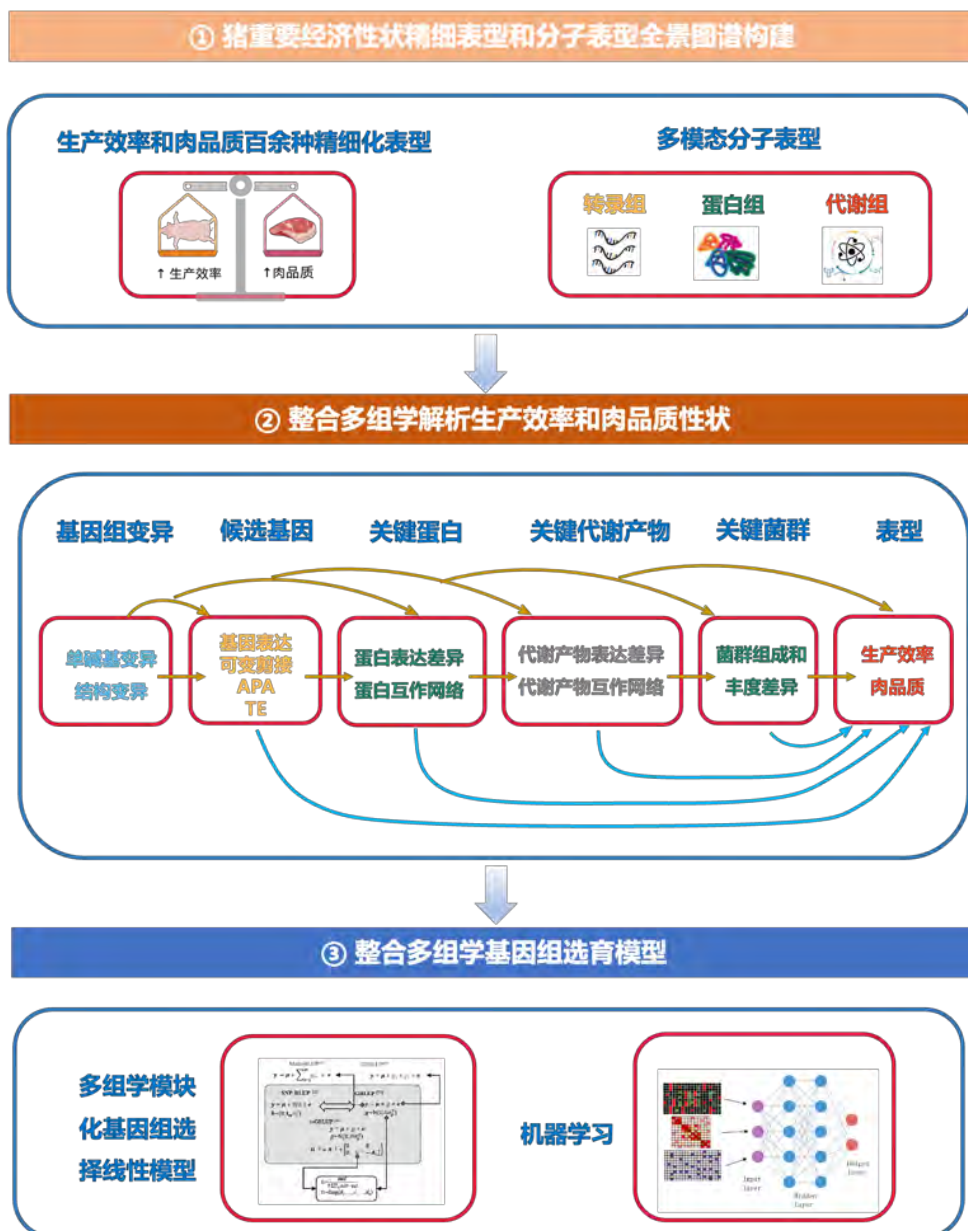
（4）整合多组学数据筛选关键调控位点、基因和机制：基于转录组数据，构建基因共表达网络，识别核心调控基因或模块，揭示基因间相互作用及其对性状形成的影响。结合蛋白质组和代谢组数据，解析蛋白质与代谢物之间的关系，识别与性状相关的关键代谢途径及其调控蛋白质。结合代谢组和肠道宏基因组数据，解析微生物组成和代谢产物的关系。利用 FastQTL 进行结构变异 molQTL 定位，并比较其与 SNP molQTL 定位结果。借鉴多组织 TWAS 分析策略并结合 GWAS 进行关联分析，揭示从基因组变异到基因分子表型，再到表型的全景图，从而筛选影响猪重要经济性状的關鍵调控位点、基因及调控机制。

（5）开展功能验证，明确其生物学效应并评估育种价值：通过 CRISPR/Cas9 基因编辑技术或 RNA 干扰技术对核心候选突变或基因进行体内体外功能验证，明确其在猪重要经济性状中的具体作用，对其调控效应进行验证和评估。

4. 融合多组学信息的精准选育体系的构建

（1）在 BLUP 框架下，整合各组学信息，并将鉴定到的关键基因及其调控网络融入统计模型中，通过矩阵降维、并行计算构建融合多组学信息的基因组选育模型和算法；

（2）鉴于线性模型难以捕捉复杂的非线性关系，采用卷积神经网络技术，捕获多组学数据的高维特征。利用自编码器将特征映射到不同低维空间，运用 SVR、RR、BNN 以及 XGBoost、Adaboost 等多种机器学习方法，构建稳定性及泛化性高的融合多组学信息的基因组选育模型和算法。



课题技术路线图

三、主要创新点

围绕基础前沿、共性关键技术或应用示范等层面，简述课题的主要创新点。具体内容应包括该项创新的基本形态及其前沿性、时效性等，并说明是否具备方法、理论和知识产权特征。每项创新点的描述限 500 字以内。

1. 创新点 1：表型数据的测定和分析方法创新——采用创新的智能设备、方法拓展生产性能测定范围，获得与猪生产效率和肉品质相关的精细表型图谱，实现重要表型的智能化、精准化采集及数据清洗和标准化，为研究表型机制打下良好基础；

2. 创新点 2：多组学解析创新——整合多组学数据，对精细表型和分子表型进行深度挖掘，筛选影响猪肉生产效率和肉品质表型的关键调控变异、基因和高维分子表型，更好地解析其生物学背景及调控机制；

3. 创新点 3：利用多组学信息的模型创新——研发基于线性模型的多组学模块化基因组选择算法，研发整合多组学信息的机器学习模型和算法，提升基因组选择准确性，服务于猪的精准选育体系的构建。



四、预期经济社会效益

课题的科学、技术、产业预期指标及科学价值、社会、经济、生态效益。限 500 字以内。

1. 课题的科学、技术、产业预期指标

预期科学指标：本课题围绕猪的生产效率与肉品质相关性状开展研究，着眼于智能化、精准化测定表型数据并据此构建多模态分子表型图谱，对猪重要经济性状进行多组学解析，挖掘其遗传机制并进一步挖掘出目标性状的关键调控位点。最终挖掘与猪重要经济性状相关的关键调控位点 8-10 个，重要候选基因 2-4 个。

预期技术指标：本课题将对猪重要经济性状进行多组学解析，开发能够利用多组学信息的稳定高效基因组选育模型和算法 1-2 个，较传统选育算法预测准确性提升 3% 以上，并据此构建能够整合多组学信息的猪精准选育体系。

预期产业指标：为育种企业或专家提供重要基因的调控位点 3 个及以上，为我国猪种质创新和新品种培养提供优质基因资源。

2. 课题的社会、经济、生态效益

社会效益方面：本课题拟构建一个能整合多组学信息的猪精准选育体系，为政府决策者、企业、科研机构以及育种专家提供强而有效的数据支撑和决策辅助。

经济效益方面：本课题研发基因组选择新方法，大大提高选种准确性，预计为猪养殖业带来显著的经济回报。

生态效益方面：本课题将显著提高猪品种质量，从而有效提高生产效率、节省资源并减少污染排放，从而提升生态效益。



五、课题年度计划

按每 6 个月制定形成课题的计划进度，应将课题的考核指标分解落实到年度计划中。

1. 年度：2024 年 12 月—2025 年 5 月

任务：猪重要常规性状表型和智能化测定表型等数据收集，并收集相关组织，进行多组学测定，构建多组学数据，并进行初步分析。

考核指标：挖掘与猪重要经济性状相关的关键调控位点 1 个。

成果形式：关键遗传调控位点以发表论文、科技报告或提交到国家共享数据库的形式体现。

2. 年度：2025 年 6 月—2025 年 11 月

任务：分子表型图谱构建；多组学数据标准化方法研究，多组学数据高维特征提取，以及各组学数据模块化构建方法研究。

考核指标：挖掘与猪重要经济性状相关的关键调控位点 2-3 个，发表高水平学术论文 1-2 篇。

成果形式：关键遗传调控位点以发表论文、科技报告或提交到国家共享数据库的形式体现。

3. 年度：2025 年 12 月—2026 年 5 月

任务：构建猪智能化表型采集新方法；多组学数据特征间的连接与叠加关系分析，不同组学之间相关矩阵的构建以及整合多组学信息的线性模型快速求解方法的开发。

考核指标：挖掘与肉类生产密切相关的关键遗传调控位点 2-3 个；建立基因组选育模型和算法 1-2 个，完成高水平论文 1-2 篇。

成果形式：关键遗传调控位点和候选基因以发表论文、科技报告或提交到国家共享数据库的形式体现；模型和算法以软著或共享数据库体现。

4. 年度：2026 年 6 月—2026 年 11 月

任务：构建猪重要经济性状表型数据清洗处理方法；对不同的表型尤其是分子表型进行降维处理及特征分析等，筛选具代表性的特征表型，构建快速表型测定系统；构建猪重要经济性状精细表型图谱。

考核指标：挖掘与肉类生产密切相关的关键遗传调控位点 1-2 个；候选基因 1-2 个；发表高水平论文 1-2 篇；为育种企业或专家提供重要经济性状基因位点 1 个及以上。



成果形式：关键遗传调控位点以发表论文、科技报告或提交到国家共享数据库的形式体现；重要经济性状基因位点由行业专家评议、行业专家现场评估、第三方评价或育种企业证明。

5. 年度：2026 年 12 月—2027 年 5 月

任务：识别对猪重要经济性状起关键作用的候选代谢产物；整合多组学信息的新线性模型的效果评估，以及不同性状适用模型的研究。

考核指标：挖掘与猪重要经济性状相关的关键遗传调控位点 1-2 个，候选基因 1-2 个，完成高水平论文 1-2 篇，获得融合基因组信息的新算法和模型 1 个，为育种企业或专家提供重要经济性状基因位点 1 个及以上。

成果形式：关键遗传调控位点和候选基因以发表论文、科技报告或提交到国家共享数据库的形式体现；重要经济性状基因位点由行业专家评议、行业专家现场评估、第三方评价或育种企业证明；模型和算法以软著或共享数据库体现。

6. 年度：2027 年 6 月—2027 年 11 月

任务：构建机器学习的多组学基因组选择新方法，并评估基于机器学习的多组学基因组选择新方法的效果，针对不同性状，确定最优的多组学组合模型。撰写总结报告，完成总结和项目结题验收与鉴定。

考核指标：挖掘与肉类型生产性状相关遗传调控位点 1-2 个，发表高水平论文 1-2 篇，获得基因组新算法和模型 1 个；为育种企业或专家提供重要经济性状基因位点 1 个及以上。

成果形式：关键遗传调控位点和候选基因以发表论文、专利、科技报告或提交到国家共享数据库的形式体现；重要经济性状基因位点由行业专家评议、行业专家现场评估、第三方评价或育种企业证明；模型和算法以软著或共享数据库体现。



六、课题组织实施机制及保障措施

1、 的内 组织管理方式、协 机制等， 500 字以内。

根据《国家 点研发）划管理办法》规定， 人对 全 ，加强 的规范化、科学化管理，建立科学、合理的组织管理体系，建立、健全各 管理制度， 以保 的 利完成。

人的主要职 包括根据实施情况，召 而 专家组会 ，并对实施 程中出现的 大 提出建 ，每个季度检查 执行 度和完成情况；组织和协 各个子 的交流与合作，优势互补； 国内外同行对 的实施情况 行 估及动态 整；组织 的中期检查、结 收和 务审）等工作；接受科技 和 目依托 的检查。

组是执行科研合同、完成科研任务的实体，由 组 及研究 干组成。明确各 的研究任务和分工， 用 组 制。 组 的主要职 包括协 各成员组的科研合作及 源共享，推动 的整体研究 度， 目管理协 组及时沟 和汇报研究成果，确保完成研究任务，确保 目目标的实现。

单独 立专 经 户，实行专款专用。要求各子 承担单位主管 每年 行 自查。按 目任务分工，各参与单位团 因自己的原因导致 目成果未能 到 目任务书约定的考核指标的，相应团 应当 取措施尽快推 执行，并各自承担由此而增加的相应 用。积极 合 目 人、 目专家组和 组的协 指导，将保 的 利实施， 成各 任务目标。

2、 实施的相关政策，已有的组织、技术基础，支撑保 条件， 500 字以内。

1) 政策保

中央一号文件《中共中央国务院人关于全 推 乡村振兴加快农业农村现代化的意见》均明确强 种业的关键作用；2021 年提出农业现代化，种子是基础；2022 年提出全 实施种业振兴行动方案，2023 年再次 申实施种业振兴行动的要性，要求全 实施生物育种 大 目。在此期 ，国家多个主管 对种业的发展均做出 要指示。2021 年，中央全 深化改 委员会第二十次会 审 了《种业振兴行动方案》，明确指出必 把民族种业搞上去，把种源安全提升到关系国家安全的战略 度。本 目聚焦指南目标，紧紧围绕农业动物优 形状的决定机制开展研究工作，符合国家对生物育种基础性



研究给予 期稳定支持的政策。

2) 管理保

严格按照国家和农业农村 相关组织要求和保 条件 立。本 单位中国农业大学拥有完整的 组织机构和完备的管理制度，可以保 按照国家政策履行 实施的管理职 。其他参加单位 是在此 域具有明显优势的 校，具备参与国家级科研)划 目的丰富经 ，为 实施提供了基本组织保 。 在管理上制定了严 的措施，包括年度考核、绩效考核和动态管理等机制，确保 的 效实施； 目牵头单位组织成立了 目管理协 组和咨 专家组以强化 目的日常管理；本 建立合同协约，明确任务、 任和权力，各单位之 互相 合支持， 为 和 目的 利实施提供了有力的保 。

3、对实现 目总目标的支撑作用，及与 目内其他 的协同机制， 500 字以内。

中国农业大学拥有畜禽生物育种全国 点实 室、畜禽营养与 养全国 点实 室、国家畜禽种 源库等 6 个国家级平台。 有大型仪器 备共享平台，拥有大型仪器 160 多台（套），以“仪器共享、样品检测、数据分析、信息挖掘”为手段，为 行提供全方位支撑。 各参与单位包括中国农业科学人北京畜牧兽医研究所、华南农业大学两个优 新基因挖掘与育种价值 价 域的技术优势单位，可完全满 和 目实施的硬件和技术平台的要求。各参加单位 期从事动物优 性状的挖掘和 传改良创新工作，在优 品种培育方 积累了丰富的经 ，可保 本 和 目的 利实施。



七、知识产权对策、成果管理及合作权益分配

限 500 字以内。

本项目由管理办公室严格按照《科技成果登记办法》《知识产权认证管理办法》等国家科技成果相关规定，制定知识产权对策、成果管理办法及合作权益分配方案，保障各单位之间的知识产权分配合规合法。

1. 知识产权对策

在项目执行过程中，牵头和参与单位拥有独立申请专利、发表论文、成果鉴定的权利。由单一单位完成的技术成果及其形式的知识产权归该单位所有；而由多方共同完成的成果及知识产权则归共有，相关的权益分配和专利申报排名根据各方的贡献大小，并通过共同协商签订合同来确定。在此基础上，项目组要积极申报尤其是发明专利，加快进度以保护过程中产生的原创性技术或方法。

2. 成果管理

所有项目产出的成果，无论是论文、专著、专利、软件、数据库还是奖项，都必须注明国家重点研发计划项目资助及课题编号。在专利申请前，相关成果不得发表、公布或泄露给他人。如果未经批准而发表或泄露信息，导致研究成果无法获得专利保护，将对相关责任人追究责任。

3. 合作权益分配

本项目及其课题实施过程中产生的知识产权，如果涉及国家安全、国家利益和重大社会公共利益，属于国家所有，但项目责任单位享有免费使用权利。各参与单位对于自己承担攻关内容的成果及专利权都有所有权。如果欲将研究成果申报奖项，需要征得其他合作单位的同意，并保持公开透明，不能瞒过其他单位独立申报。对于由某一单位或团队独立完成的技术成果，成果归属该单位或团队，但仍需纳入项目成果中。对于多个单位或团队合作完成的成果，成果共有，贡献大小决定了各单位或团队的排名和独立使用权。



八、需要约定的其他内容

限 500 字以内。

（1）项目承担单位将按照科技计划项目科学数据汇交的有关要求，制定科技资源汇交方案，将科学数据汇交到有关方面认可的科学数据中心并出具汇交凭证。（2）项目承担单位将按照国家重点研发计划项目安全管理的有关要求，切实履行项目安全管理职责，加强人员培训教育，强化科研过程安全管理。



九、课题参加人员基本情况表

填表说明： 1. 专业技术职称：A、正高级 B、副高级 C、中级 D、初级 E、其他； 2. 投入本课题的全时工作时间（人月）是指在课题实施期间该人总共为课题工作的满月度工作量；累计是指课题组所有人员投入人月之和； 3. 课题固定研究人员需填写人员明细； 4. 是否有工资性收入：Y、是 N、否； 5. 人员分类代码：B、课题负责人 C、项目/课题骨干 D、其他研究人员； 6. 工作单位：填写单位全称，其中高校要具体填写到所在院系。														
序号	姓名	性别	出生日期	证件类型	证件号码	专业技术职称	职务	最高学位	专业	投入本课题的全时工作时间（人月）	人员分类代码	在课题中分担的任务	是否有工资性收入	工作单位
1	申振才	男	1980-12-19	身份证	372926198012197731	副高级	副主任	博士	数学，计算机科学	18	课题负责人	研发融合多组学信息的基因组选育模型和算法	是	中国农业大学理学院
2	王涵	女	1994-03-26	身份证	370782199403264308	副高级	无	博士	数学，统计学	12	课题骨干	猪重要经济性状多组学解析及精准选育	是	中国农业大学理学院
3	赵福平	男	1981-02-08	身份证	430421198102081850	正高级	无	博士	动物遗传育种	9	课题骨干	猪重要经济性状多组学解析及精准选育	是	中国农业科学院北京畜牧兽医研究所
4	高亚辉	男	1989-09-01	身份证	140428198909010434	副高级	无	博士	动物遗传育种	12	课题骨干	猪重要经济性状多组学解析	是	华南农业大学动物科学学院
5	曲阳	男	1997-07-23	身份证	230103199707233931	其他	无	硕士	计算机科	12	其他研究	数据分析	否	中国农业大学信电学院



									学技术		人员			
6	孙雨悦	女	2002-05-05	身份证	150302200205050529	其他	无	学士	统计学	12	其他研究 人员	数据分析	否	中国农业大学理学院
7	廖天翊	男	2001-08-13	身份证	430481200108139159	其他	无	学士	计算机科 学技术	12	其他研究 人员	分析数据	否	中国农业大学理学院
8	蔡晓钿	女	1997-08-04	身份证	445102199708040020	其他	无	硕士	动物遗传 育种与繁 殖	12	其他研究 人员	数据分析	否	华南农业大学动物科学学 院
9	钟展明	男	1998-03-16	身份证	441424199803163070	其他	无	硕士	动物遗传 育种与繁 殖	30	其他研究 人员	数据分析	否	华南农业大学动物科学学 院
10	周天如	女	2001-05-16	身份证	370902200105160927	其他	无	学士	动物遗传 育种与繁 殖	30	其他研究 人员	数据分析	否	华南农业大学动物科学学 院
固定研究人员合计										159	/	/	/	/
流动人员或临时聘用人员合计										0	/	/	/	/
累计										159	/	/	/	/



课题预算表

表B1 课题编号： 2024YFF1000101 课题名称： 猪重要经济性状多组学解析及精准选育 金额单位： 万元

序号	预算科目名称	金额
	(1)	(2)
1	一、中央财政专项资金	368.00
2	（一）直接费用	296.00
3	1. 设备费	
4	其中：购置设备费	
5	2. 业务费	226.00
6	3. 劳务费	70.00
7	（二）间接费用	72.00
8	二、其他来源资金	
9	三、合计	368.00

注：1. 间接费用无需编制预算说明；2. 绩效支出在间接费用中无比例限制。承担单位在统筹安排间接费用时，要处理好合理分摊间接成本和对科研人员激励的关系，绩效支出安排与科研人员在课题工作中的实际贡献挂钩。



设备费——购置/试制设备预算明细表

表B2

课题编号：2024YFF1000101

课题名称：猪重要经济性状多组学解析及精准选育

金额单位：万元

填表说明：1.设备分类：购置、试制； 2.购置设备类型：通用、专用； 3.试制设备不需填列本表（9）列、（10）列、（11）列、（12）列； 4.设备单价的单位为万元/台套，设备数量的单位为台套； 5.单价50万元以下的设备不用填写； 6.本表只填写中央财政资金购置（试制）的设备。												
序号	设备名称	设备分类	功能和技术指标	单价	数量	金额	购置或试制单位	安置单位	购置设备类型	主要生产厂家及国别	规格型号	拟开放共享范围
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
无记录												
单价50万元以上购置设备合计							/	/	/	/	/	/
单价50万元以上试制设备合计							/	/	/	/	/	/
累计							/	/	/	/	/	/



课题单位经费预算明细表

表B3 课题编号：2024YFF1000101

课题名称：猪重要经济性状多组学解析及精准选育

金额单位：万元

填表说明：1.单位类型分课题承担单位、课题参与单位； 2.组织机构代码指企事业单位国家标准代码，单位若已三证合一请填写单位统一社会信用代码，无组织机构代码的单位填写“000000000”。										
序号	单位名称	组织机构代码-统一社会信用代码		单位类型	任务分工	研究任务负责人	合计	中央财政专项资金		其他来源资金
								小计	其中：间接费用	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	中国农业大学	统一社会信用代码	12100000400018162G	课题承担单位	猪重要经济性状多组学解析及精准选育	申振才	169.00	169.00	32.50	
2	中国农业科学院北京畜牧兽医研究所	统一社会信用代码	12100000400882305E	课题参与单位	猪重要经济性状多组学解析及精准选育	赵福平	109.50	109.50	21.90	
3	华南农业大学	统一社会信用代码	124400004554165634	课题参与单位	猪重要经济性状多组学解析及精准选育	高亚辉	89.50	89.50	17.60	
累计							368.00	368.00	72.00	



预算说明

一、中央财政资金

预算的编制要坚持任务相关性、政策相符性和经济合理性，实事求是编制提出课题预算。填报时，直接费用应按设备费、业务费、劳务费三个类别填报，每个类别结合科研任务按支出用途进行说明。除 50 万元以上的设备外，其他费用只提供基本测算说明，不需要提供明细。

1. 设备费（是指项目实施过程中购置或试制专用仪器设备，对现有仪器设备进行升级改造，以及租赁外单位仪器设备而发生的费用等。计算类仪器设备和软件工具可在设备费科目编列。填报时，50 万元以上的设备详细说明，50 万元以下的设备费用分类说明）

无

2. 业务费（是指在项目实施过程中消耗的各种材料、低值易耗品等、发生的测试化验加工、燃料动力、出版文献、信息传播、知识产权事务、会议、差旅、国际合作与交流以及其他与项目实施直接相关的各项费用。编报时，对单笔大额支出、对外委托支出重点说明）

本课题预算业务费 226.00 万元，其中中央财政资金 226.00 万元，其他来源资金 0 万元。具体明细如下：

(1) 材料费：48.04 万元

①实验动物样本采集及提取检测试剂费：24.50 万元

本研究需要对 200 头个体进行组织或者血样的采集，采集后的样本进行后续基因组、转录组等多组学测序。以上工作均涉及组织和血液等样本 DNA 和 RNA 提取、纯化及检测相关试剂，包括动物基因组提取试剂盒、血液 RNA 提取试剂盒、DNA ladder、RNA 提取试剂盒和 Trizol 等，共计 24.50 万元，见表 1-1。

表 1-1 实验动物样本采集及提取检测试剂费用明细

名称	规格	单价 (元)	数量	金额 (万元)
动物组织基因组提取试剂盒	盒	980	50	4.90
血液 DNA 提取试剂盒	盒	1200	50	6.00



组织 RNA 提取试剂盒	盒	1800	40	7.20
血液 RNA 提取试剂盒	盒	1600	40	6.40
合计				24.50

②细胞培养所用试剂：14.90 万元

主要用于细胞水平的功能基因验证及调控网络解析等实验所需试剂，包括细胞培养、细胞分离传代、细胞冻存和细胞转染等所需试剂，主要包括 DMEM 基础培养基、RPMI 1640 基础培养基、DMEM-F12 培养基、胰蛋白酶、进口胎牛血清和 Lipofectamine 2000 转染试剂等，共计 14.90 万元，见表 1-2。

表1-2 常用细胞培养所用试剂费用明细

名称	规格	单价（元）	数量	金额（万元）
胰蛋白酶 (Invitrogen)	100 g	3000	10	3.00
Collagenase 胶原酶	支	3500	10	3.50
bFGF 生长因子 (Sigma)	支	2000	10	2.00
胎牛血清	瓶	6000	5	3.00
青霉素-链霉素液	瓶	400	5	0.20
嘌呤霉素	瓶	1200	5	0.60
Lipofectamine 2000	支	3000	5	1.50
Lipofectamine 3000	支	4000	2	0.80
通用细胞冻存液	瓶	100	30	0.30
合计				14.90

③分子生物学操作常用化学试剂：8.64 万元

分子生物学实验中常用化学试剂，主要包括无水乙醇、苯酚、异丙醇、氯仿、葡萄糖、亚牛磺酸、丙酮、盐酸、氢氟酸、氯化钾、高氯酸、氢氧化钠、氯化钠、氯化钙、磷酸氢二钠和异戊醇等，共计 8.64 万元，见表 2-3。

表1-3 分子生物学操作常用化学试剂

名称	规格	单价（元）	数量	金额（万元）
----	----	-------	----	--------



课题拟召开课题启动中期以及总结会议和相关的审计会议，每次会议邀请3名副高级以上职称的专家，每人每天2000元，小计为 $4 \times 2000 \text{ 元} \times 3 \text{ 人} = 2.40 \text{ 万元}$ 。

用于项目技术交流会等相关会议邀请专家进行技术指导，预计邀请副高级以上职称的专家27人次，2000元/人天，小计为5.40万元。

二、其他来源资金

对其他来源资金主要用途、支出预算做简要说明。

无



十一、 相关附件

1. 乙方与参加单位有关协议（须加盖乙方与参加单位公章、法人签字签章；协议文件须扫描上传。如无参加单位，则不填）；

课题组织实施协议

课题名称：猪重要经济性状多组学解析及精准选育

课题承担单位（甲方）：中国农业大学

课题参与单位（乙方）：中国农业大学

甲乙双方就联合实施国家重点研发计划“农业生物重要性状形成与环境适应性基础研究”重点专项2024年度项目“整合多组学信息解析畜禽重要经济性状及精准选育（基础研究类）”中“课题一：猪重要经济性状多组学解析及精准选育”工作，甲乙双方友好协商，达成以下组织实施协议：

- 1. 甲乙双方承诺本着集成优势、相互支持、责任明确、知识产权合理管理、合作共赢的精神，共同完成课题的实施工作。
- 2. 甲方作为课题承担单位，负责课题的组织协调及任务实施。乙方作为课题参加单位，积极配合并协助甲方。
- 3. 双方承诺课题实施过程产生的所有科学数据以及总结、验收等环节所需的相关材料，按期递交到专项指定平台，在专项约定的条件下对专项各承担单位乃至今后面向全社会所有的科技工作者和公众开放共享。
- 4. 乙方单位的任务分工、考核指标，经费分配如下表所示。甲方根据实际批复的经费额度，与乙方协商任务和经费分配方案。甲乙双方知晓应按国家重点研发计划管理办法及相应的经费管理规定等国家相关法规开展相关研究和经费管理，严格履行相应的义务。

乙方承担的主要任务及经费分配

参与课题	任务分工	考核指标	任务责任人	中央财政专项资金（万元）
课题一	猪重要经济性状精准表型图谱及	挖掘与肉类生产密切相关的关键遗传调控位点3个，候选性状1个；建立融合	申振才	169



	采集方法的构建	多组学信息的基因组选育模型和算法 1-2 个；发表高水平学术论文 2 篇；为育种企业或专家提供重要经济性状基因位点 1 个以上。		
--	---------	--	--	--

- 课题实施过程中知识产权归属和使用严格按照中华人民共和国和项目承担单位科技成果管理办法执行。双方在本课题实施之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。
- 合作双方因履行本协议而发生的争议，应协商解决，任何一方不得自行处理。
- 本协议经双方签字盖章后生效，一式 2 份，双方各执 1 份，具有同等法律效力。
- 其它未尽事宜由双方协商解决。

甲方（课题承担单位盖章）：

乙方（课题参加单位盖章）：

法人（签章）：

法人（签章）：

课题负责人（签字）：

课题参加人（签字）：

年 月 日

年 月 日



课题组织实施协议

课题名称: 猪重要经济性状多组学解析及精准选育

课题承担单位 (甲方): 中国农业大学

课题参与单位(乙方): 中国农业科学院北京畜牧兽医研究所

甲乙双方就联合实施国家重点研发计划“农业生物重要性状形成与环境适应性基础研究”重点专项2024年度项目“整合多组学信息解析畜禽重要经济性状及精准选育（基础研究类）”中“课题一：猪重要经济性状多组学解析及精准选育”工作，甲乙双方友好协商，达成以下组织实施协议：

1. 甲乙双方承诺本着集成优势、相互支持、责任明确、知识产权合理管理、合作共赢的精神，共同完成课题的实施工作。
2. 甲方作为课题承担单位，负责课题的组织协调及任务实施。乙方作为课题参加单位，积极配合并协助甲方。
3. 双方承诺课题实施过程产生的所有科学数据以及总结、验收等环节所需的相关材料，按期递交到专项指定平台，在专项约定的条件下对专项各承担单位乃至今后面向全社会所有的科技工作者和公众开放共享。
4. 乙方单位的任务分工、考核指标，经费分配如下表所示。甲方根据实际批复的经费额度，与乙方协商任务和经费分配方案。甲乙双方知晓应按国家重点研发计划管理办法及相应的经费管理规定等国家相关法规开展相关研究和经费管理，严格履行相应的义务。

乙方承担的主要任务及经费分配

参与课题	任务分工	考核指标	任务责任人	中央财政专项资金（万元）
课题一	猪重要经济性状精准表型图谱及采集方法的构建	挖掘与肉类生产密切相关的关键遗传调控位点 3 个，候选性状 1 个；建立融合多组学信息的基因组选育模型和算法 1 个；发表高水平学术论文 1 篇；为育种企业或专家提供重要经济性状基因位点 1 个以上。	赵福平	109.5



5. 课题实施过程中知识产权归属和使用严格按照中华人民共和国和项目承担单位科技成果管理办法执行。双方在本课题实施之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。
6. 合作双方因履行本协议而发生的争议，应协商解决，任何一方不得自行处理。
7. 本协议经双方签字盖章后生效，一式 2 份，双方各执 1 份，具有同等法律效力。
8. 其它未尽事宜由双方协商解决。

甲方（课题承担单位盖章）：

乙方（课题参加单位盖章）：

法人（签章）：

法人（签章）：

课题负责人（签字）：

课题参加人（签字）：

年 月 日

年 月 日



课题组织实施协议

课题名称：猪重要经济性状多组学解析及精准选育

课题承担单位(甲方): 中国农业大学

课题参与单位(乙方): 华南农业大学

甲乙双方就联合实施国家重点研发计划“农业生物重要性状形成与环境适应性基础研究”重点专项2024年度项目“整合多组学信息解析畜禽重要经济性状及精准选育（基础研究类）”中“课题一：猪重要经济性状多组学解析及精准选育”工作，甲乙双方友好协商，达成以下组织实施协议：

1. 甲乙双方承诺本着集成优势、相互支持、责任明确、知识产权合理管理、合作共赢的精神，共同完成课题的实施工作。
2. 甲方作为课题承担单位，负责课题的组织协调及任务实施。乙方作为课题参加单位，积极配合并协助甲方。
3. 双方承诺课题实施过程产生的所有科学数据以及总结、验收等环节所需的相关材料，按期递交到专项指定平台，在专项约定的条件下对专项各承担单位乃至今后面向全社会所有的科技工作者和公众开放共享。
4. 乙方单位的任务分工、考核指标，经费分配如下表所示。甲方根据实际批复的经费额度，与乙方协商任务和经费分配方案。甲乙双方知晓应按国家重点研发计划管理办法及相应的经费管理规定等国家相关法规开展相关研究和经费管理，严格履行相应的义务。

乙方承担的主要任务及经费分配

参与课题	任务分工	考核指标	任务责任人	中央财政专项资金（万元）
课题一	猪重要经济性状精准表型图谱及采集方法的构建	挖掘与肉类生产密切相关的关键遗传调控位点 2-3 个，候选性状 1-2 个；发表高水平学术论文 1 篇；为育种企业或专家提供重要经济性状基因位点 1 个以上。	高亚辉	89.5

5. 课题实施过程中知识产权归属和使用严格按照中华人民共和国和项目承



担单位科技成果管理办法执行。双方在本课题实施之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。

6. 合作双方因履行本协议而发生的争议，应协商解决，任何一方不得自行处理。

7. 本协议经双方签字盖章后生效，一式 2 份，双方各执 1 份，具有同等法律效力。

8. 其它未尽事宜由双方协商解决。

甲方（课题承担单位盖章）：

乙方（课题参加单位盖章）：

法人（签章）：

法人（签章）：

课题负责人（签字）：

课题参加人（签字）：

年 月 日

年 月 日



2. 申报指南规定的其他附件。



课题组织实施协议

课题名称：猪重要经济性状多组学解析及精准选育

课题承担单位（甲方）：中国农业大学

课题参与单位（乙方）：中国农业大学

甲乙双方就联合实施国家重点研发计划“农业生物重要性状形成与环境适应性基础研究”重点专项 2024 年度项目“整合多组学信息解析畜禽重要经济性状及精准选育（基础研究类）”中“课题一：猪重要经济性状多组学解析及精准选育”工作，甲乙双方友好协商，达成以下组织实施协议：

- 1. 甲乙双方承诺本着集成优势、相互支持、责任明确、知识产权合理管理、合作共赢的精神，共同完成课题的实施工作。
- 2. 甲方作为课题承担单位，负责课题的组织协调及任务实施。乙方作为课题参加单位，积极配合并协助甲方。
- 3. 双方承诺课题实施过程产生的所有科学数据以及总结、验收等环节所需的相关材料，按期递交到专项指定平台，在专项约定的条件下对专项各承担单位乃至今后面向全社会所有的科技工作者和公众开放共享。
- 4. 乙方单位的任务分工、考核指标，经费分配如下表所示。甲方根据实际批复的经费额度，与乙方协商任务和经费分配方案。甲乙双方知晓应按国家重点研发计划管理办法及相应的经费管理规定等国家相关法规开展相关研究和经费管理，严格履行相应的义务。

乙方承担的主要任务及经费分配

参与课题	任务分工	考核指标	任务责任人	中央财政专项资金（万元）
课题一	猪重要经济性状精准表型图谱及采集方法的构建	挖掘与肉类生产密切相关的关键遗传调控位点 3 个，候选性状 1 个；建立融合多组学信息的基因组选育模型和算法 1-2 个；发表高水平学术论文 2 篇；为育种企业或专家提供重要经济性状基因	申振才	169

		位点 1 个以上。		
--	--	-----------	--	--

5. 课题实施过程中知识产权归属和使用严格按照中华人民共和国和项目承担单位科技成果管理办法执行。双方在本课题实施之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。
6. 合作双方因履行本协议而发生的争议，应协商解决，任何一方不得自行处理。
7. 本协议经双方签字盖章后生效，一式 2 份，双方各执 1 份，具有同等法律效力。
8. 其它未尽事宜由双方协商解决。

甲方（课题承担单位盖章）：



法人（签章）：

孙其信

课题负责人（签字）：申振才

2024年12月9日

乙方（课题参加单位盖章）：



法人（签章）：

孙其信

课题参加人（签字）：申振才
王涵

2024年12月9日

课题组织实施协议

课题名称：猪重要经济性状多组学解析及精准选育

课题承担单位（甲方）：中国农业大学

课题参与单位（乙方）：中国农业科学院北京畜牧兽医研究所

甲乙双方就联合实施国家重点研发计划“农业生物重要性状形成与环境适应性基础研究”重点专项 2024 年度项目“整合多组学信息解析畜禽重要经济性状及精准选育（基础研究类）”中“课题一：猪重要经济性状多组学解析及精准选育”工作，甲乙双方友好协商，达成以下组织实施协议：

1. 甲乙双方承诺本着集成优势、相互支持、责任明确、知识产权合理管理、合作共赢的精神，共同完成课题的实施工作。
2. 甲方作为课题承担单位，负责课题的组织协调及任务实施。乙方作为课题参加单位，积极配合并协助甲方。
3. 双方承诺课题实施过程产生的所有科学数据以及总结、验收等环节所需的相关材料，按期递交到专项指定平台，在专项约定的条件下对专项各承担单位乃至今后面向全社会所有的科技工作者和公众开放共享。
4. 乙方单位的任务分工、考核指标，经费分配如下表所示。甲方根据实际批复的经费额度，与乙方协商任务和经费分配方案。甲乙双方知晓应按国家重点研发计划管理办法及相应的经费管理规定等国家相关法规开展相关研究和经费管理，严格履行相应的义务。

乙方承担的主要任务及经费分配

参与课题	任务分工	考核指标	任务责任人	中央财政专项资金（万元）
课题一	猪重要经济性状精准表型图谱及采集方法的构建	挖掘与肉类生产密切相关的关键遗传调控位点 3 个，候选性状 1 个；建立融合多组学信息的基因组选育模型和算法 1 个；发表高水平学术论文 1 篇；为育种企业或专家提供重要经济性状基因位点 1 个以上。	赵福平	109.5

5. 课题实施过程中知识产权归属和使用严格按照中华人民共和国和项目承担单位科技成果管理办法执行。双方在本课题实施之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。

6. 合作双方因履行本协议而发生的争议，应协商解决，任何一方不得自行处理。

7. 本协议经双方签字盖章后生效，一式 2 份，双方各执 1 份，具有同等法律效力。

8. 其它未尽事宜由双方协商解决。

甲方（课题承担单位盖章）：



法人（签章）：

孙其信

课题负责人（签字）：申振才

2024年12月9日

乙方（课题参加单位盖章）：



法人（签章）：

张军民

课题参加人（签字）：

李瑞平

2024年12月9日

课题组织实施协议

课题名称：猪重要经济性状多组学解析及精准选育

课题承担单位（甲方）：中国农业大学

课题参与单位（乙方）：华南农业大学

甲乙双方就联合实施国家重点研发计划“农业生物重要性状形成与环境适应性基础研究”重点专项 2024 年度项目“整合多组学信息解析畜禽重要经济性状及精准选育（基础研究类）”中“课题一：猪重要经济性状多组学解析及精准选育”工作，甲乙双方友好协商，达成以下组织实施协议：

1. 甲乙双方承诺本着集成优势、相互支持、责任明确、知识产权合理管理、合作共赢的精神，共同完成课题的实施工作。
2. 甲方作为课题承担单位，负责课题的组织协调及任务实施。乙方作为课题参加单位，积极配合并协助甲方。
3. 双方承诺课题实施过程产生的所有科学数据以及总结、验收等环节所需的相关材料，按期递交到专项指定平台，在专项约定的条件下对专项各承担单位乃至今后面向全社会所有的科技工作者和公众开放共享。
4. 乙方单位的任务分工、考核指标，经费分配如下表所示。甲方根据实际批复的经费额度，与乙方协商任务和经费分配方案。甲乙双方知晓应按国家重点研发计划管理办法及相应的经费管理规定等国家相关法规开展相关研究和经费管理，严格履行相应的义务。

乙方承担的主要任务及经费分配

参与课题	任务分工	考核指标	任务责任人	中央财政专项资金（万元）
课题一	猪重要经济性状精准表型图谱及采集方法的构建	挖掘与肉类生产密切相关的关键遗传调控位点 2-3 个，候选性状 1-2 个；发表高水平学术论文 1 篇；为育种企业或专家提供重要经济性状基因位点 1 个以上。	高亚辉	89.5

5. 课题实施过程中知识产权归属和使用严格按照中华人民共和国和项目承担单位科技成果管理办法执行。双方在本课题实施之前各自所获得的知识产权及相应权益均归各自所有，不因共同申请本课题而改变。

6. 合作双方因履行本协议而发生的争议，应协商解决，任何一方不得自行处理。

7. 本协议经双方签字盖章后生效，一式 2 份，双方各执 1 份，具有同等法律效力。

8. 其它未尽事宜由双方协商解决。

甲方（课题承担单位盖章）：



法人（签章）：

孙其信

课题负责人（签字）：申振才

2024年12月9日

乙方（课题参加单位盖章）：



法人（签章）：

高翔

课题参加人（签字）：

2024年12月9日

任务书签署

甲乙双方根据《国务院印发关于深化中央财政科技计划（专项、基金）管理改革方案的通知》（国发〔2014〕64号）、《国务院关于优化科研管理提升科研绩效若干措施的通知》（国发〔2018〕25号）、《国务院办公厅关于改革完善中央财政科研经费管理的若干意见》（国办发〔2021〕32号）、《科技部 财政部关于印发<国家重点研发计划管理暂行办法>的通知》（国科发资〔2017〕152号）、《财政部 科技部关于印发<国家重点研发计划资金管理办法>的通知》（财教〔2021〕178号）、《科学技术活动违规行为处理暂行规定》（科学技术部令第19号）、《科技部财政部关于印发<中央财政科技计划（专项、基金等）监督工作暂行规定>的通知》（国科发政〔2015〕471号）、《科技部 自然科学基金委关于进一步压实国家科技计划（专项、基金等）任务承担单位科研作风学风和科研诚信主体责任的通知》（国科发监〔2020〕203号）等有关文件规定，以及有关法律、政策和管理要求，依据项目立项通知，签署本任务书。

同时，本单位和项目负责人**郑重承诺**：对本项目所有成果产出（包括但不限于新产品、新技术、标准、论文、专利等）的真实性、与项目的关联性负责，将按要求落实科研作风学风和科研诚信主体责任；项目经费全部用于与本项目研究工作相关的支出，不截留、挪用、侵占，不用于与科学研究无关的支出；接受并积极配合相关部门的监督检查。如有违反，本单位和项目负责人以及相关成果产出者愿接受项目管理专业机构和相关部门做出的各项处理决定，包括但不限于终止项目执行、追回项目（课题）经费，取



消一定期限国家科技计划项目申报资格，记入科研诚信严重失信行为数据库以及主要负责人接受相应党纪政纪处理等。

项目牵头承担单位（甲方）：

法定代表人签字（签章）：

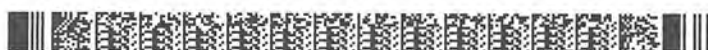
张军民



项目负责人签字（签章）：

侯明

2024年12月12日



课题承担单位（乙方）：

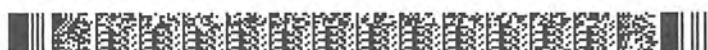
法定代表人签字（签章）：孙其信



2024年12月12日

课题负责人签字（签章）：申振才

2024年12月12日



任务书签署

甲乙双方根据《国务院印发关于深化中央财政科技计划（专项、基金）管理改革方案的通知》（国发〔2014〕64号）、《国务院关于优化科研管理提升科研绩效若干措施的通知》（国发〔2018〕25号）、《国务院办公厅关于改革完善中央财政科研经费管理的若干意见》（国办发〔2021〕32号）、《科技部 财政部关于印发〈国家重点研发计划管理暂行办法〉的通知》（国科发资〔2017〕152号）、《财政部 科技部关于印发〈国家重点研发计划资金管理办法〉的通知》（财教〔2021〕178号）、《科学技术活动违规行为处理暂行规定》（科学技术部令第19号）、《科技部财政部关于印发〈中央财政科技计划（专项、基金等）监督工作暂行规定〉的通知》（国科发政〔2015〕471号）、《科技部 自然科学基金委关于进一步压实国家科技计划（专项、基金等）任务承担单位科研作风学风和科研诚信主体责任的通知》（国科发监〔2020〕203号）等有关文件规定，以及有关法律、政策和管理要求，依据项目立项通知，签署本任务书。

同时，本单位和项目负责人**郑重承诺**：对本项目所有成果产出（包括但不限于新产品、新技术、标准、论文、专利等）的真实性、与项目的关联性等负责，将按要求落实科研作风学风和科研诚信主体责任；项目经费全部用于与本项目研究工作相关的支出，不截留、挪用、侵占，不用于与科学研究无关的支出；严格按照政府采购和保密法律法规规定开展政府采购活动，规范信息公开工作；接受并积极配合相关部门的监督检查。如有违反，本单位和项目负责人以及相关成果产出者愿接受项目管理专业机构和相关部门做出的各项处理决定，包括但不限于终止项目执行、追回项目（课题）经费，取消一定期限国家科技计划项目申报资格，记入科研诚信严重失信行为数据库以及主要负责人接受相应党纪政纪处理等。



项目牵头承担单位（甲方）：

法定代表人签字（签章）：

张军民



项目负责人签字（签章）：

侯明

2024 年 12 月 12 日

课题承担单位（乙方）：

法定代表人签字（签章）：

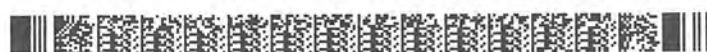
孙其信



课题负责人签字（签章）：

申振才

2024 年 12 月 12 日



国家重点研发计划 子课题任务书

子课题名称：	影响猪繁殖性状功能基因及调控元件筛选及 功能解析
课题牵头单位（甲方）：	四川农业大学
所属项目名称：	畜禽重要经济性状功能基因和调控元件筛选 及功能解析
所属专项：	畜禽新品种培育与现代牧场科技创新
项目管理专业机构：	中国农村技术开发中心
子课题承担单位（乙方）：	华南农业大学
子课题负责人：	高亚辉
执行期限：	2023 年 12 月 至 2028 年 11 月

中华人民共和国科学技术部制

2023 年 9 月

填 写 说 明

一、任务书甲方即课题牵头承担单位，乙方即子课题承担单位。

二、任务书中的单位名称，请按规范全称填写，并与单位公章一致。

三、任务书要求提供乙方与所有参加单位的合作协议。

四、任务书中文字须用宋体小四号字填写。

五、凡不填写内容的栏目，请用“无”表示。

六、乙方完成任务书的填写完成，提交甲方审核确认后，用 A4 纸在线打印、签章后上传电子扫描件。

七、如课题下仅设一个子课题，子课题任务书只需填报课题基本信息表与课题预算部分。

八、涉密课题请在“国家科技计划管理信息系统公共服务平台”下载任务书的电子版模板，按保密要求离线填写、报送。一式八份报项目课题承担单位签章，其中课题承担单位两份，子课题负责人四份，作为项目任务书附件两份。

九、《项目申报书》和《项目任务书》是本任务书填报的重要依据，任务书填报不得降低考核指标，不得自行对主要研究内容作大的调整。《项目申报书》和《项目任务书》和本任务书将共同作为课题过程管理、综合绩效评价（验收）和监督评估的重要依据。

课题基本信息表

子课题名称		影响猪繁殖性状功能基因及调控元件筛选及功能解析								
子课题编号		2023YFD1300403-03								
所属课题名称		畜禽重要经济性状功能基因和调控元件筛选及功能解析								
课题编号		2023YFD1300403								
所属项目		畜禽育种大数据与新一代全基因组选择技术体系研究								
密级		<input checked="" type="checkbox"/> 公开 <input type="checkbox"/> 秘密 <input type="checkbox"/> 机密		单位总数		1				
课题成果技术就绪度		<input checked="" type="checkbox"/> 1. 发现基本原理 <input type="checkbox"/> 2. 形成技术方案 <input type="checkbox"/> 3. 方案通过验证 <input type="checkbox"/> 4. 形成单元并验证 <input type="checkbox"/> 5. 形成分系统并验证 <input type="checkbox"/> 6. 形成原型并验证 <input type="checkbox"/> 7. 现实环境的应用验证 <input type="checkbox"/> 8. 用户验证认可 <input type="checkbox"/> 9. 得到推广应用								
课题成果应用的主要国民经济行业		科学研究和技术服务业 研究和试验发展 自然科学研究和试验发展								
课题的社会经济目标		一级目标：农林牧渔业发展 二级目标：畜牧业								
经费预算		总需求 80.00 万元，其中中央财政专项资金需求 80.00 万元								
课题周期节点		起始时间		2023 年 12 月		结束时间		2028 年 11 月		
		实施周期		共 60 个月		预计中期时间点		2026 年 6 月		
子课题承担单位	单位名称		华南农业大学			单位法定代表人姓名		薛红卫		
	单位性质		大专院校			组织机构代码		124400004554165634		
	单位主管部门		广东省教育厅			隶属关系		地方		
	单位所属地区		广东省			地市（市、自治州、盟）		广州市		
	通信地址		广东省广州市天河区五山路 483 号			邮政编码		510642		
	单位开户名称		华南农业大学							
	开户银行（全称）		中国工商银行广州五山支行			汇入地点		广东省广州市		
	银行账号		3602002609000310520			银行机构代码		102581000546		
子课题负责人	姓 名	高亚辉		性 别	<input checked="" type="checkbox"/> 男 <input type="checkbox"/> 女		出生日期		1989. 09. 01	
	证件类型	身份证		证件号码	140428198909010434					

	所在单位	华南农业大学		
	最高学位	<input checked="" type="checkbox"/> 博士 <input type="checkbox"/> 硕士 <input type="checkbox"/> 学士 <input type="checkbox"/> 其他		
	职 称	<input type="checkbox"/> 正高级 <input checked="" type="checkbox"/> 副高级 <input type="checkbox"/> 中级 <input type="checkbox"/> 初级 <input type="checkbox"/> 其他		职务 无
	电子邮箱	yahui.gao@scau.edu.cn	移动电话	15652937882
子课题 联系 人	姓 名	陈丹霞	电子邮箱	cdx1415336962@163.com
	固定电话	13632269242	移动电话	13632269242
	证件类型	身份证	证件号码	440982199501194066
子课题 参加人 数	<u>4</u> 人。其中：	高级职称 <u>1</u> 人，中级职称 <u>0</u> 人，初级职称 <u>1</u> 人，其他 <u>2</u> 人；		
		博士学位 <u>2</u> 人，硕士学位 <u>1</u> 人，学士学位 <u>1</u> 人，其他 <u>0</u> 人。		
课题 简介 (限 500 字以内)	猪繁殖性状是重要的经济性状，其性状形成的分子调控机制一直被视为猪分子育种改良的基石和突破口。本项目针对获得的繁殖性状存在表型差异的猪群体，拟利用二代和三代基因组重测序技术准确鉴定差异表型猪的基因组变异；利用转录组、单细胞核转录组等多组学技术解析猪繁殖性状的分子基础，鉴定影响猪繁殖的候选基因组变异。最后，对关键基因进行分子功能的机制解析及育种价值评估。			

填表说明：1. 组织机构代码指企事业单位国家标准代码，单位若已三证合一请填写单位统一社会信用代码，无组织机构代码的单位填写“000000000”；
2. 单位公章名称必须与单位名称一致；
3. 单位开户名称应与单位名称一致，如有开户名称不一致等特殊情况，必须提供证明文件。

一、子课题目标及考核指标、评测方式/方法

子课题目标：构建猪繁殖性状多组学图谱

考核指标：构建猪繁殖结构变异及转录组表达图谱；筛选影响猪繁殖性状的候选基因 1-2 个；发表高水平论文 1-2 篇。

评测方式/方法：公共数据库检索或 SCI 检索

二、子课题研究内容及研究方法

（一）子课题的主要研究内容

拟解决的关键科学问题、关键技术问题，针对这些问题拟开展的主要研究内容，限1000字以内。

猪的繁殖性状是重要的经济性状，其性状形成的分子调控机制一直被视为猪分子育种改良的基石和突破口。本项目拟采用领域内重大共性组学关键技术，着眼于基因组变异，聚焦于繁殖性状候选基因鉴定，落脚于关键基因分子机制解析和育种价值评估，最终将构建猪多组学数据库，涵盖基因组序列变异、细胞注释等信息，为新一代猪全基因组选择模型开发提供多组学特征信息。具体将从2个方面开展研究：（1）鉴定影响猪繁殖的候选基因组变异；（2）开展繁殖性状相关基因的功能验证，明确生物学效应并评估育种价值。

（二）子课题采取的研究方法

针对课题研究拟解决的问题，拟采用的方法、原理、机理、算法、模型等，限1000字以内。

针对当前我国猪繁殖性状功能基因挖掘不足、多组学数据整合效率低、资源利用低效、分析方法单一的问题，拟以繁殖性状存在表型差异的畜禽群体为研究对象，整合多组学技术，鉴定繁殖强相关育种基因或调控元件。

任务的具体研究方法如下：

基于三代测序技术对高、低繁殖力的两个群体进行测序，采用“基于读长”和“基于组装”两种方式鉴定结构变异，预测潜在影响猪繁殖性状的结构变异；整合转录组数据剖析结构变异影响性状表型的遗传机制，筛选候选基因；与已有的猪繁殖性状相关的数量性状基因座定位结果、全基因组关联分析结果和染色质状态进行比较，整合多种信息分析影响猪繁殖性状的遗传机制；大规模收集猪表型组数据，进一步解析猪繁殖性状的形成和调控规律。

三、预期经济效益

课题的科学、技术、产业预期指标及科学价值、社会、经济、生态效益。限500字以内。

猪繁殖性状结构变异检测在畜牧业科学技术领域占有重要地位，对提高猪只遗传改良效率、促进畜牧业可持续发展具有重要意义。该技术旨在通过三代测序分析手段，识别影响猪只繁殖性状的关键基因或基因组区段，从而指导猪只的遗传改良和育种选择。

科学与技术预期指标。精准检测：利用三代测序手段准确识别影响猪繁殖性状的基因变异；数据分析：准确解读基因变异与猪繁殖性状之间的关系；应用研究：将研究成果应用于实际育种，提高猪种的繁殖效率和生产性能。

科学价值。促进遗传学和基因组学理论的发展，为猪只遗传改良提供科学依据。提高对猪繁殖生物学机制的理解，推动生命科学研究的深入。

社会与经济效益。提高生产效率：通过优化猪只的遗传性状，提高繁殖效率和生产性能，增加畜牧业产出。促进产业升级：推动畜牧业向更高效率、更环保的方向发展，促进相关产业链的技术升级和产业结构调整。增加农民收入：提升猪只的经济性状，增加养殖效益，直接促进农民增收。

生态效益。资源高效利用：通过改善猪只的繁殖性能，降低资源消耗，提高饲料转化率，减少畜牧业对环境的压力。可持续发展：促进畜牧业向环境友好型、可持续发展方向转变，有助于生态平衡的维护。

四、子课题年度计划

按每年度制定形成课题的计划进度，应将课题的考核指标分解落实到年度计划中。

1、年度：2023 年 12 月-2024 年 11 月

任务：筛选繁殖性状存在表型差异的猪试验群体，相关组织样品的采集及多组学数据的测定。

考核指标：确定采集组织及组学数据质量。

成果形式：子课题的研究进展。

2、年度：2024 年 12 月—2025 年 11 月

任务：差异表型猪的基因组 SNPs、InDels 和 SVs 变异位点的分析鉴定。

考核指标：差异变异位点筛选

成果形式：子课题的研究进展，文章。

3、年度：2025 年 12 月—2026 年 11 月

任务：完成猪重要经济性状的多组学数据的测定工作，研究差异表型猪的单细胞转录组差异。

考核指标：完成组学数据测定，初步构建单细胞转录图谱。

成果形式：子课题的研究进展，文章或专利。

4、年度：2026 年 12 月—2027 年 11 月

任务：完善单细胞时空转录图谱，研究繁殖强相关基因。

考核指标：绘制单细胞时空转录图谱，鉴定影响猪繁殖性状的候选基因

成果形式：文章或专利。

5、年度：2027 年 12 月—2028 年 11 月

任务：基于获得的多组学数据，探究关键变异影响猪重要经济性状形成的分子机理，在细胞和活体水平开展相关基因的功能验证。

考核指标：明确关键变异影响猪重要经济性状形成的分子机理和相关基因功能。

成果形式：文章或专利。

五、课题组织实施机制及保障措施

1、课题的内部组织管理方式、协调机制等，限 500 字以内。

本子课题由华南农业大学承担。研究团队长期致力于畜禽遗传育种和分子遗传学

的研究，在国内外具有较高的学术影响力，在畜禽复杂性状的遗传机制解析、功能基因组学和全基因组选择等领域具有扎实的研究基础和技术储备。由课题承担单位四川农业大学成立课题管理办公室，负责开展相关会议组织、信息发布、成果管理等项目日常管理工作。将邀请国内生物育种和数字种业领域的知名专家对课题实施进行技术咨询和指导，制定科学的评价指标体系，实时动态追踪课题完成情况，及时解决出现的问题，确保按计划高质量完成目标任务。课题负责人和项目资深专家顾问负责制定科学详细的实施方案、技术路线和工作计划，整体把握课题进度和完成质量，处理课题执行过程中的各种具体问题，协调人员安排，定期汇报课题的进展情况及阶段性成果。课题和子课题负责人及课题骨干负责提出本课题及项目共性问题及其解决方案，协调课题实施方案的技术要求，制定课题内部或跨课题的技术、数据统一规范，为项目最终考核指标育种技术体系的集成及综合示范提供技术保障。

2、课题实施的相关政策，已有的组织、技术基础，支撑保障条件，限 500 字以内。

2020 年中央经济工作会议和 2021 年中央一号文件都明确要“打好种业翻身仗”：对育种基础性研究以及重点育种项目给予长期稳定支持；深入实施农作物和畜禽良种联合攻关；实施新一轮畜禽遗传改良计划和现代种业提升工程。同时，国家高度重视种业领域的科技发展，2021 年出台了《种业振兴行动方案》。本课题研究内容符合政策相关要求，具有较好的政策支撑条件。本课题将严格按照国家相关管理规定进行科学的项目管理，建立健全组织协调机制，注重完善协同创新管理体系和相应规章制度。本课题参加单位包括国内在该领域的主要相关研究队伍，与国内畜禽育种优势企业具有长期深入的产学研合作，形成了从基础研究、技术开发、产业应用、技术推广构成的人才团队和雄厚的技术力量，可为本课题实施提供良好的组织保障和实施平台。课题团队由四川农业大学和华南农业大学组成，团队拥有猪禽种业全国重点实验室和国家生猪种业工程技术研究中心等国家级科研平台，在数字育种、全基因组选择、功能基因解析等方面拥有国内国际领先的科研条件和软硬件设施装置。相关技术成果和技术储备为本课题开展提供了重要的研发经验和前期条件，可为本课题的顺利实施提供资源技术支撑。

3、对实现项目总目标的支撑作用，及与项目内其他课题的协同机制，限 500 字以内。

在研究内容方面：本项目的总目标是创建新一代畜禽全基因组选择育种技术体系，实现遗传评估准确性、育种效率显著提升。该课题的目标是整合畜禽经济性状多组学数

据筛选关键基因及调控元件，是总项目的一个分支，从基础的调控原件到关键基因的研究，为总课题的全面性和系统性研究提供支持。本项目着眼于关键基因分子机制解析和育种价值评估，最终将构建 2 个主要畜禽多组学数据库，为新一代畜禽全基因组选择模型开发提供多组学特征信息。

在平台方面：团队拥有猪禽种业全国重点实验室和国家生猪种业工程技术研究中心等国家级科研平台，在数字育种、全基因组选择、功能基因解析等方面拥有领先的科研条件和软硬件设施装置。相关技术成果为本课题开展提供了重要的研发经验和前期条件，可为本课题的顺利实施提供资源技术支撑。

项目内其他课题的协同机制：项目组具有教授、讲师、博士研究生、硕士研究生和实验技术人员多层次研究梯队，研究力量搭配合理，具备优秀的联合攻关团队，各个梯队在生物学、生物信息学、蛋白质组学等方面均有涉及，各有优势，相互探讨、相互协助与优势互补能够保证课题的顺利完成。

4、生物安全管理与科普宣传

课题需严格遵守《农业转基因生物安全管理条例》及其配套规章制度，按要求进行试验审批或报备。课题单位需建立生物安全管理责任制，做到分工明确、责任到人，对违反相关规定的，接受相关责任追究和取消相关人员及团队承担有关农业科研项目资格等处罚。

为营造生物育种产业化良好的社会氛围，课题单位需积极参与科普宣传工作，制定年度科普宣传工作计划，并报农业农村部科学技术司审核。课题单位通过举办专题报告会、讲座、培训、开放日，发表科普文章，参与集体科普活动等方式积极开展生物技术科普宣传，让大众更好了解生物技术，科学引导大众认识生物技术研发与应用。本课题拟推荐 1 名正式工作人员担任网络科普员，每月在网上发布不少于 5 篇生物技术科普信息（含转发和评论），并在农业农村部科学技术司统一组织下开展科普宣传活动。

六、知识产权对策、成果管理及合作权益分配

（一）知识产权对策

本子课题参加单位在本课题实施之前所获得的知识产权及相应权益均归各自所有。课题执行过程中，各参加单位取得的研究成果和相关的知识产权归各单位所有，但课题主持单位有权因非商业目的和课题研究需求（如：以政府性会议、报告、技术

文件、统计资料、原始数据等)使用参加单位课题相关信息和数据资料,在课题申报和执行期间进行知识产权共享。

(二)成果管理

知识产权与科研成果应当严格遵照国家相关的法律和规定实施管理。因课题实施需要与本课题相关的技术资料、数据等所有信息,未经提供方同意,不得提供给第三方或对外公开。课题主持单位与课题参加单位及其有关人员均应遵照相关法律政策的要求,承担保密责任,并应采取相应的保密措施。

(三)合作权益分配

参加单位合作产生的研究成果和相关知识产权归双方或多方共有,依各方在成果中的实际分工和贡献大小署名和分配权益。

七、需要约定的其他内容

无

八、子课题参加人员基本情况表

填表说明：1、专业技术职称：A、正高级 B、副高级 C、中级 D、初级 E、其他；
2、投入本课题的全时工作时间（人月）是指在课题实施期间该人总共为课题工作的满月度工作量；累计是指课题组所有人员投入人月之和；
3、课题固定研究人员需填写人员明细；
4、是否有工资性收入：Y、是 N、否；
5、人员分类代码：B、课题负责人 C、项目/课题骨干 D、其他研究人员；
6、工作单位：填写单位全称，其中高校要具体填写到所在院系。

序号	姓名	性别	出生日期	证件类型	证件号码	专业技术职称	职务	最高学位	专业	投入本课题的全时工作时间（人月）	人员分类代码	在课题中分担的任务	是否有工资性收入	工作单位
1	高亚辉	男	1989.09.01	身份证	140428198909010434	B	无	博士	动物遗传育种与繁殖	40	C	课题负责人，组织实施	是	华南农业大学动物科学学院
2	魏趁	女	1991.10.11	身份证	652101199110110426	D	无	博士	动物遗传育种与繁殖	50	D	负责数据分析，具体实验实施	是	华南农业大学动物科学学院
3	龚文滔	男	1996.09.29	身份证	440981199609293916	E	无	硕士	动物遗传育种与繁殖	50	D	负责数据分析，具体实验实施	否	华南农业大学动物科学学院
4	黎柏朗	男	2000.05.27	身份证	440681200005274211	E	无	学士	动物遗传育种与繁殖	50	D	负责数据分析，具体实验实施	否	华南农业大学动物科学学院
固定研究人员合计														
流动人员或临时聘用人员合计														
累计														
										190	/	/	/	/
										0	/	/	/	/
										190	/	/	/	/

九、子课题预算表（2023-2024 年）

序号	预算科目名称	金额：单位万元
	(1)	(2)
1	一、中央财政专项资金	80
2	（一）直接费用	62.19
3	1. 设备费	0
4	其中：购置设备费	0
5	2. 业务费	49.94
6	3. 劳务费	12.25
7	（二）间接费用	17.81
8	二、其他来源资金	0
9	三、合计	80

注：1、间接费用无需编制预算说明；2、绩效支出在间接费用中无比例限制。承担单位在统筹安排间接费用时，要处理好合理分摊间接成本和对科研人员激励的关系，绩效支出安排与科研人员在课题工作中的实际贡献挂钩。

预算说明

一、中央财政资金

预算的编制要坚持任务相关性、政策相符性和经济合理性，实事求是编制提出课题预算。填报时，直接费用应按设备费、业务费、劳务费三个类别填报，每个类别结合科研任务按支出用途进行说明。除 50 万元以上的设备外，其他费用只提供基本测算说明，不需要提供明细。

子课题总经费 80.00 万元，其中直接经费 62.19.00 万元，间接经费 17.81 万元。

1. 设备费：合计支出 0.00 万元。

2. 业务费：合计支出 49.94 万元。

2.1 材料费：12.44 万元

(1) 根据子课题任务需求，需购买种猪用于屠宰测定相关性能，购买种猪与屠宰费用：3000.00 元/头×15 头=4.50 万元；

(2) 本子课题执行过程中，需要采集种猪的组织样品，并对采集的样品进行实验，需要消耗低值易耗品（手套、口罩、干冰等）及试验试剂，预计支出 7.94 万元。

2.2 测试化验加工费：27.00 万元

(1) 根据本子课题任务要求，需要对种猪采集的组织样品进行转录组测序，需对 50 个样本进行转录组测序，每个样本测序费用为 1,000 元，50 个样本测序费用为：1,000 元/样本×50 样本，小计 5.00 万元；

(2) 在大规模视频数据计算及分析的过程中，由于运算强度大、负荷高、耗时长，需依托天河二号、太湖之光等国家超级计算平台完成相关计算任务，预计需要购买机时 220 万核时，单价 0.10 元/核时，小计 22.00 万元。

2.3 燃料动力费：0.00 万元

2.4 出版/文献/信息传播/知识产权事务费：5.50 万元

(1) 论文版面费 4.50 万元：预计发表部分国际权威刊物论文 3 篇，版面费 4.50 万元；

(2) 知识产权事务费 0.60 万元：预计申请专利 1 项，每项 6000 元，小计 0.60 万元；

(3) 其他费用 0.40 万元：项目相关资料打印等支出 0.40 万元。

2.5 会议/差旅/国际合作交流费：5.00 万元

会议/差旅/国际合作交流费预算支出 5.00 万元，未超过直接费用的 10%，因此不再对预算内容和资金安排进行说明，也不再提供预算测算依据。本项目中差旅费主要用于课题组成员参加国内学术会议、调研及派遣学生前往合作猪场的采样差旅费。

3. 劳务费：合计支出 12.25 万元

3.1 劳务费 12.25 万元

(1) 参与本子课题博士研究生 1 人，每人补助 2000 元/月，执行期内预计工作 30 个月/人，1 人小计支出 6.00 万元；博士研究生的主要工作内容：负责本子课题的具体实施，管理测序数据和整合、学术成果（文章和专利等）整合及撰写等；

(2) 参与本子课题硕士研究生 1 人，补助 1000 元/月，执行期内预计工作 30 个月/人，1 人小计支出 3.00 万元。硕士研究生的主要工作内容：本子课题具体的实验操作，整理实验数据并进行分析，开展具体任务的研发等；

(3) 本子课题聘用科研财务助理 1 名用于课题资料整理及财务管理，工资 6000 元/月，子课题执行期间工作 5 个月，预计支出 3.00 万元

(4) 临时聘用屠宰实验人员进行屠宰及采样，预计支出 0.25 万元。

(二) 间接费用

本课题预算间接费用 17.81 万元，主要用于在项目组织实施过程中发生的无法由直接费用中列支的相关费用。主要包括承担单位为项目研究提供的现有仪器设备及房屋占有费、日常水电气暖消耗费、有关管理费用及科研人员绩效支出等。

二、其他来源资金

无。

十、相关附件

1. 乙方与参加单位有关协议（须加盖乙方与参加单位公章、法人签字签章；协议文件须扫描上传。如无参加单位，则不填）；
2. 申报指南规定的其他附件。

任务书签署

甲乙双方根据《国务院印发关于深化中央财政科技计划（专项、基金）管理改革方案的通知》（国发〔2014〕64号）、《国务院关于优化科研管理提升科研绩效若干措施的通知》（国发〔2018〕25号）、《国务院办公厅关于改革完善中央财政科研经费管理的若干意见》（国办发〔2021〕32号）、《科学技术活动违规行为处理暂行规定》（科学技术部令第19号）、

《科技部 财政部关于印发<中央财政科技计划（专项、基金等）监督工作暂行规定>的通知》（国科发政〔2015〕471号）、《科技部 自然科学基金委关于进一步压实国家科技计划（专项、基金等）任务承担单位科研作风学风和科研诚信主体责任的通知》（国科发监〔2020〕203号）、《科技部、财政部、自然科学基金委关于进一步加强统筹国家科技计划项目立项管理工作的通知》（国科办资〔2022〕107号）等有关文件规定，以及有关法律、政策和管理要求，依据项目立项通知，签署本任务书。

同时，本单位和课题负责人郑重承诺：对本课题所有成果产出（包括但不限于新产品、新技术、标准、论文、专利等）的真实性、与项目（课题）的关联性等负责，将按要求落实科研作风学风和科研诚信主体责任；课题经费全部用于与本课题研究工作相关的支出，不截留、挪用、侵占，不用于与科学研究无关的支出；接受并积极配合相关部门的监督检查。如有违反，本单位和课题负责人以及相关成果产出者愿接受项目管理专业机构和相关部门做出的各项处理决定，包括但不限于终止课题执行、追回课题经费，取消一定期限国家科技计划项目（课题）申报资格，记入科研诚信严重失信行为数据库以及主要负责人接受相应党纪政纪处理等。

课题牵头承担单位（甲方）：

法定代表人签字（签章）：

吴建



课题负责人签字（签章）：

陆

年 月 日

子课题牵头承担单位（乙方）：

法定代表人签字（签章）：

薛红已



子课题负责人签字（签章）：

高

年 月 日

任务书编号：2024A04J3806

广州市科技计划项目 任务书

项目名称：	基于三代测序鉴定影响猪繁殖性状的结构变异
承担单位：	华南农业大学
项目负责人：	高亚辉
计划类别：	基础研究计划
专题名称：	2024年度基础与应用基础研究专题
支持方向：	青年博士“启航”项目
组织单位：	华南农业大学
起止时间：	2024-01-01 至 2025-12-31
主管处室：	基础研究处

广州市科学技术局制

二〇二四年

填写说明

1. 任务书甲方为广州市科学技术局；乙方为项目承担单位；丙方为项目组织单位。

2. 任务书基于项目申报书转换而成，请按照“广州科技大脑”提示在线填写核实，若存在不填写内容的栏目，请用“无”表示；任务书中的单位名称应为规范全称，并与单位公章一致。

3. 乙方与合作单位的合作协议自动从项目申报书中读取，如需变化调整，须待任务书签订后，按要求及时办理重大变更。

4. 乙方完成项目任务书在线填写，依次提交丙方和甲方审核确认后，按要求登录“穗好办”APP完成电子签章。不具备电子签章条件的单位，经与业务主管处室沟通对接后，可下载电子版项目任务书用A4纸双面打印装订签章；一式六份报甲方和丙方签章，其中甲方两份丙方两份，项目承担单位和项目负责人各一份。

5. 涉密项目请在“广州科技大脑”下载项目任务书模板，按保密要求离线填写报送。

6. 项目申报书是项目任务书填报的重要依据，未经甲方许可，乙方不得修改考核指标，调整主要研究内容。项目任务书将作为项目实施管理、验收结题和监督评估的重要依据。

7. 项目任务书中的“备注”，包括重要的必须补充的内容。

8. “广州科技大脑”是项目管理过程中重要通知和文书的电子送达平台。为确保电子送达渠道畅通，乙方和项目负责人应及时更新维护“广州科技大脑”的单位和个人信息。

9. 根据相关要求，项目涉及人体临床研究的，项目需经医学伦理委员会审查通过并在任务书附件栏上传相关佐证材料。

一、项目基本信息

项目 基本 信息	项目名称	基于三代测序鉴定影响猪繁殖性状的结构变异
	申请市财政科技经费	5(万元)
	研究期限	2(年)
项目 摘要	繁殖效率可影响猪养殖业的经济效益，挖掘繁殖性能相关的分子标记和基因并应用于基因组选择，可加快繁殖性状的遗传进展。作为基因组特征之一的结构变异对畜禽重要性状的表型起重要作用。本项目以猪繁殖力为目标性状，利用三代测序对繁殖性状表现极端的两组个体进行测序，构建全基因组结构变异。通过组间比较，鉴定影响猪繁殖性状的结构变异和基因。本项目将为猪育种的基因组选择提供理论依据，为推动猪产业可持续发展提供指导意义。	

二、项目单位情况

项目承担单位	单位名称	华南农业大学	统一社会信用代码	124400004554165634
	注册时间	1952-01-01	单位类型	高等院校
	注册地址	广东省广州市天河区五山路483号		
	办公地址	广东省广州市天河区五山路483号		
	联系人	姓名	倪慧群	
		手机号码	13711345768	
		电子邮箱	kjcgxk@scau.edu.cn	
	开户银行	广东广州工行五山支行		
	开户户名	华南农业大学		
银行账号	3602002609000310520			

三、项目负责人信息

姓名	高亚辉	证件类型	身份证
证件号码	140428198909010434	性别	男
出生日期	1989-09-01	民族	汉族
国籍	中国	学历	博士研究生
学位	博士	学位授予国家 (或地区)	中国
职务	无	职称	无
所学专业	动物遗传育种与繁殖	手机号码	15652937882
办公电话	020-85282019	电子邮箱	yahui.gao@scau.edu.cn

四、项目经费信息

本项目总投入：¥（5）万元，其中，市财政科技经费：¥（5）万元，自筹经费：¥（0）万元。

经费下达计划			
资金来源	小计	市财政科技经费	自筹经费
2024	5	5	0
总计	5	5	0

（单位：万元）

注：本专题纳入“包干制”，市财政科技经费按市科技计划项目经费“包干制”相关规定执行。

五、预期代表性成果

项目负责人在项目实施期内，以该项目作为资助项目获得以下5种情形之一且经费使用符合规定的，由组织单位审核后通过验收。

（一）项目实施期内，以第一作者/通讯作者发表论文1篇或以上（须标注资助项目编号）；

（二）项目实施期内，以第一完成人申请或授权专利、软件著作权1项或以上；

（三）项目实施期内，获省级以上科技计划项目或人才项目支持1项或以上；

（四）项目实施期内，获省级以上科技奖励（含列入获奖团队成员名单）1项或以上；

（五）项目实施期内，获得职称晋升。

六、备注

专题补充约定条款：

甲方对未履行勤勉尽责义务的相关责任主体，自作出处理结论之日起，依照法律法规规定或任务书约定实施惩戒5年，取消相关责任主体申报市科技计划项目、申领市科技计划项目经费的资格。

预期代表性成果需在实施期内获得。

项目承担单位（乙方）及项目负责人承诺书

承诺书

本单位/本人作为广州市科技计划项目承担单位/项目负责人，将严格遵守广州市科技计划管理相关规定，严格履行自身责任，加强对项目组人员及合作单位的管理，在此郑重承诺：

（一）确保与本项目有关的全部材料真实、合法、有效，未侵犯其他方知识产权等权利，不存在多头申报、重复申报行为；

（二）严格遵守《广州市科技创新条例》《广州市科技计划项目管理办法》《广州市科技计划项目经费管理办法》《广州市科技计划科技报告管理办法》等相关规定，实施项目和经费管理；

（三）严格遵守国家、省、市关于科研诚信和科技伦理的有关法律、法规，相关政策以及各项规定，加强项目实施过程中的科研诚信及科技伦理管理，恪守科研道德准则。

如有违反，本单位/本人愿意接受相关部门做出的各项处理决定，包括但不限于终止项目、停拨经费、核减经费、追回经费，取消一定期限广州市科技计划项目申报资格，记入科研失信行为数据库，将不良行为向社会公开等。

项目承担单位：华南农业大学

日期：年 月 日

项目负责人：高亚辉

日期：2023年12月17日

任务书签署

甲乙丙三方根据《广州市科技计划项目管理办法》《广州市科技计划项目经费管理办法》《广州市科技计划科技报告管理办法》等有关文件规定，以及有关法律、政策和管理要求，签署本任务书。

签订地点：广州市越秀区

广州市科学技术局（甲方）：广州市科学技术局
局项目经办人：联系电话：
责任处室负责人：

项目承担单位（乙方）：华南农业大学
二级部门：华南农业大学动物科学学院
项目负责人：高亚辉
项目经费汇入账号
账户名：华南农业大学 账号：3602002609000310520
开户银行：广东广州工行五山支行
财务负责人：肖斐

组织单位（丙方）：华南农业大学
项目经办人：

合同编号: NJTG 20250168

2024年省级乡村振兴战略专项 种业振兴行动项目

合 同 书

项目名称: 省级畜禽核心育种场生产性能测定、资源保种场保护和种畜禽质量监测

项目管理单位(甲方): 广东省农业技术推广中心

项目承担单位(乙方): 华南农业大学

项目负责人: 张 哲 联系电话: 18825084398

项目联系人: 陈丹霞 联系电话: 13632269243

广东省农业农村厅制

第一条 为保障2024年省级乡村振兴战略专项种业振兴行动项目顺利实施，按时保质保量完成项目任务，根据《中华人民共和国民法典》、《广东省省级财政专项资金管理办法（修订）》（粤府〔2023〕34号）、《广东省农业农村厅种业振兴行动专项资金管理办法（试行）》等文件有关规定，经甲、乙双方协商一致，签署本合同书。

第二条 甲方的权利义务：本合同履行过程中，甲方有权对乙方项目的实施情况和资金到位、使用情况进行监督、检查，提出改进要求。

第三条 乙方的权利义务：

1. 按财政资金管理规定，对甲方核拨的资金做到专款专用，单独列账，并随时配合甲方进行监督检查；

2. 认真填写本合同书《项目任务书》，《项目任务书》的内容应与《联合申报项目协议书》保持一致；

3. 严格按照本合同书及合同书《项目任务书》的要求及时完成项目建设内容，项目实施完成后，按照甲方要求提交验收报告；

4. 按照《广东省农业农村厅种业振兴行动专项资金管理办法（试行）》规定，按期（每年6月30日、12月31日）向甲方书面报告项目实施进展及资金使用情况等内容；

5. 乙方需保留与项目实施相关的协议和财务凭证，并向甲方备案。

第四条 本项目资金不得用于以下方向：1. 行政事业单位基本支出；2. 各项奖金、津贴和福利补助；3. 企业担保金和弥补企业亏损；4. 修缮楼堂馆所以及建造职工住宅；5. 弥补单位预算支出缺口和偿还债务；6. 购买交通工具及通讯设备；7. 形成地方政府债务的支出；8. 购买理财产品、发放借款及平衡预算等。

第五条 项目验收。项目验收及结果处理严格执行《广东省农业农村厅专项资金项目验收管理办法（试行）》（粤农农办〔2023〕73号）的规定。乙方应在项目完成后3个月内，提出验收申请。申请验收除了规定材料外，还应该提交项目审计报告或者经费决算表，其中财政专项资金50万元以下的项目，需提交由项目承担单位财务部门出具的经费决算表，财政专项资金50万元（含）以上的项目，需提交由项目承担单位委托会计师事务所出具的审计报告。

第六条 在履行本合同的过程中，如出现相关政策法规重大改变等不可抗力情况，甲方有权对所核拨经费的数量和时间进行相应调整。因非不可抗力因素导致的项目未履行或未履行完毕，或因乙方责任造成项目不能继续开展的，甲方有权终止项目合同，收回尚未使用和使用不符合规定的财政经费。

第七条 在履行本合同的过程中，乙方发现可能导致项目整体或部分失败的情形时，应及时通知甲方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第八条 实施项目所获得的科技成果(项目成果)归属、成果转让和实施技术成果所产生的经济利益的分享，按照国家和广东省有关规定执行。项目研究成果应向省农业农村厅进行登记、备案，对外发布前应征求省农业农村厅的意见。

第九条 本合同在履行过程中发生的任何争议，由甲乙双方友好协商解决。

第十条 本合同未尽事宜，双方同意按照《广东省省级财政专项资金

管理办法（修订）》（粤府〔2023〕34号）、《广东省农业农村厅种业振兴行动专项资金管理办法（试行）》履行。

项目管理单位（甲方）（盖章）：广东省农业技术推广中心

甲方代表（签字）：



签订日期：2025年3月21日



项目承担单位（乙方）（盖章）：华南农业大学

乙方代表（签字）：



银行账户名：华南农业大学

开户银行：中国工商银行广州五山支行

银行账号：3602002609000310520



签订日期：2025年3月21日

项目任务书

填写说明

一、本项目任务书由乙方填写。

二、本项目任务书所列内容应实事求是填写，表达要明确、严谨。对填写不符合要求的，或填报内容出现虚报夸大、不切实际的，将退回项目承担单位修改。

三、项目任务书规定的项目考核指标、建设内容和绩效目标必须依据《联合申报项目协议书》填写，应遵循明确、量化、可考核的原则，其中技术指标应明确项目完成时达到的关键技术参数及预期可以形成的发明专利、标准、新技术、新产品、新装置、论文、专著等的数量。

四、《联合申报项目协议书》及申报单位基本情况表是本项目任务书填报的重要依据，项目任务书填报不得修改考核指标、绩效目标、资金预算等内容。《联合申报项目协议书》、申报单位基本情况表和本项目任务书将共同作为项目过程管理、综合绩效评价（验收）和监督评估的重要依据。

五、省财政资金支出的预算计划应按照国家及省相关规定执行。

六、表格栏目不够可自行增加。

一、目的及意义

主要说明项目的建设目的、研究价值和意义。

广东是畜牧业生产和消费大省。202 年，广东畜牧业总产值 1696.8 亿元，比 2012 年增加 507 亿元，位居全国第 10 位，占农林牧渔业总产值 20.3%。2023 年全省肉类产量 498.91 万吨，同比增长 4.9%，排名全国第六；禽蛋产量 49.91 万吨，同比增长 5.7%；奶类产量 20.25 万吨，同比增长 2.2%。肉蛋奶产量除低于山东外，肉类产量大幅高于江苏、浙江、福建，禽蛋、奶类产量与浙江、福建基本持平。广东生猪年出栏量常年保持在 3500 万头左右，生猪自给率达到国家要求的 70%以上。2023 年，广东积极推进生猪产能恢复，新增产能持续释放，生猪存出栏增幅明显，生猪年末存栏 2049.2 万头，比 2020 年末增加 281.9 万头，增长 15.95%，基本恢复到正常年份水平。全年生猪出栏 3794.01 万头，同比增长 8.5%，全国排名第七。“十八大”以来，广东家禽年出栏基本保持增长趋势，家禽年出栏从 2012 年的 11.3 亿只，先微跌到 2014-2016 年的 10 亿只左右，再上涨到 2020-2023 年的 13 亿只左右。2023 年，广东家禽出栏 13.74 亿只，同比增长 2.8%，全国排名第二。省委省政府高度重视畜禽育种工作，《关于加快推进生猪家禽产业转型升级的意见》（粤府办〔2019〕25 号）提出，要实施畜禽种业振兴行动，加强新品种（配套系）选育和扩繁，开展畜禽良种联合攻关，提升主要品种核心种源自给率和育种企业国际竞争力，打造一批“粤字号”特色优质畜禽产业品牌。为深入贯彻落实省委、省政府的决策部署，进一步规范种畜禽核心场管理，推动畜禽遗传改良计划实施，增强畜禽种业自主创新能力和企业核心竞争力，实现畜禽种业高质量发展，省农业农村厅印发了《广东省省级种畜禽核心场管理办法》（粤农农规〔2021〕5 号），要求省级畜禽核心育种场按要

求完成省农业农村厅部署的种畜禽性能测定任务，并将测定数据报送广东省生猪和家禽测定中心。广东省生猪和家禽测定中心对测定数据统一进行遗传评估，出具性能测定报告、撰写形势分析报告、基因组选择评估结果等，在开展种畜禽质量监督工作的同时，指导种畜禽生产企业开展选种选配工作，提高种畜禽育种效率。

综上所述，开展种畜禽生产性能测定与种畜禽质量监测非常必要。

二、项目建设内容

详细说明项目建设内容（项目需求或项目建设任务）。

提升种猪育种自主创新能力，整合企业、科研力量，开展区域性联合育种，完善遗传评估体系，加快基因组选择等育种新技术推广和应用，持续提高种猪选育遗传进展。一是对省级种猪核心育种场采集的基因组测序数据进行质量评估，基因组育种值计算，对符合联合评估要求的种猪群体开展联合遗传评估，并按季度撰写基因组育种效果评估报告；二是不断完善猪基因组遗传评估核心算法，并开展算法效果评测。

备注：项目建设内容（项目需求或项目建设任务）按《联合申报项目协议书》内容

三、项目绩效目标

主要说明项目实施后，预期达到的目标和产生的效果，相关表述应量化。

对省级种猪核心育种场采集的基因组测序数据进行质量评估，基因组育种值计算，对符合联合评估要求的种猪群体开展联合遗传评估，种猪基因组测序数据质量可靠，且在种猪选育中得到较好应用。编写种猪基因组育种效果评估报告 4 期。

四、项目进度安排

详细说明各阶段的工作内容和时间安排情况。

2024 年 12 月-2025 年 3 月，检测省级种猪核心场育种场采集的基因组测序数据的质量，并进行质控；

2025 年 3 月-2025 年 6 月，对基因组测序数据进行育种值及相关遗传参数估计；

2025 年 6 月-2025 年 12 月，对符合联合评估要求的种猪群体开展联合遗传评估；并不断完善猪基因组遗传评估核心算法及算法效果测评。

备注：项目绩效目标按《联合申报项目协议书》内容填写。

五、项目组主要成员(含项目负责人)

姓名	性别	身份证号	单位	职称/职务	电话
张哲	男	411329198403051332	华南农业大学	教授/副院长	18825084398
高亚辉	男	140428198909010434	华南农业大学	副教授	15652937882
滕金言	男	450802199707123639	华南农业大学	副教授	15521314172
袁晓龙	男	341621198904152331	华南农业大学	副教授	13560369611
李加琪	男	440106196509221918	华南农业大学	教授	13609709797

六、资金使用预算

主要说明资金使用的范围或方向及资金使用进度安排。

本项目省级财政经费共 30 万元，主要用于样品测序支出测试化验工费 18.2 万元，数据分析的学生劳务费 10.00 万元，间接费用支出 1.80 万元。

资金使用预计 2025 年 6 月前支出 60%，2025 年 10 月前支出 100%。

七、保障措施

说明围绕完成项目任务、目标所要采取的具体措施。

一是实行项目负责人管理制度，以加强对项目方案设计、组织实施、资金使用、进度检查、总结验收等方面的全面领导与协调。二是项目建设严格执行落实省相关财务制度，全面实行财务审批制度和报账制度，专款专用，专账管理，强化进度管理，确保项目资金使用规范。三是项目建设及运行阶段，加强与中心的合作，优化资源配置，强化项目实施过程的监督与评估，确保项目高质量完成并取得预期成效。

编号：CARS-35

现代农业产业技术体系 2024 年度任务书

岗位名称： 配套系育种

岗位科学家： 张 哲

岗位科学家依托单位： 华南农业大学

依托单位法定代表人： 薛红卫

农业农村部科学技术司

二〇二四年四月

填 写 说 明

1. 本任务书由首席科学家、产业技术研发中心依托单位、岗位科学家及岗位科学家依托联合签订。
2. 本任务书要求按照已给的格式，5 号宋体字填写，单倍行间距，段落间无间距，A4 纸双面打印。
3. 本任务书封面不签字盖章，仅在签约方页签字盖章。
4. 本任务书可视填报内容自行增加页码。
5. 本任务书一式 4 份，生猪产业技术研发中心依托单位 1 份，首席科学家 1 份，岗位科学家 1 份，岗位科学家依托单位 1 份。

体 系 编 号

CARS-01	水稻产业技术体系	CARS-29	葡萄产业技术体系
CARS-02	玉米产业技术体系	CARS-30	桃产业技术体系
CARS-03	小麦产业技术体系	CARS-31	香蕉产业技术体系
CARS-04	大豆产业技术体系	CARS-32	荔枝龙眼产业技术体系
CARS-05	大麦青稞产业技术体系	CARS-33	天然橡胶产业技术体系
CARS-06	谷子高粱产业技术体系	CARS-34	牧草产业技术体系
CARS-07	燕麦荞麦产业技术体系	CARS-35	生猪产业技术体系
CARS-08	食用豆产业技术体系	CARS-36	奶牛产业技术体系
CARS-09	马铃薯产业技术体系	CARS-37	肉牛牦牛产业技术体系
CARS-10	甘薯产业技术体系	CARS-38	肉羊产业技术体系
CARS-11	木薯产业技术体系	CARS-39	绒毛用羊产业技术体系
CARS-12	油菜产业技术体系	CARS-40	蛋鸡产业技术体系
CARS-13	花生产业技术体系	CARS-41	肉鸡产业技术体系
CARS-14	特色油料产业技术体系	CARS-42	水禽产业技术体系
CARS-15	棉花产业技术体系	CARS-43	兔产业技术体系
CARS-16	麻类产业技术体系	CARS-44	蜂产业技术体系
CARS-17	糖料产业技术体系	CARS-45	大宗淡水鱼产业技术体系
CARS-18	蚕桑产业技术体系	CARS-46	特色淡水鱼产业技术体系
CARS-19	茶叶产业技术体系	CARS-47	海水鱼产业技术体系
CARS-20	食用菌产业技术体系	CARS-48	虾蟹产业技术体系
CARS-21	中药材产业技术体系	CARS-49	贝类产业技术体系
CARS-22	绿肥产业技术体系	CARS-50	藻类产业技术体系
CARS-23	大宗蔬菜产业技术体系	CARS-51	种业创新共性技术体系
CARS-24	特色蔬菜产业技术体系	CARS-52	耕地资源利用与保护共性技术创新团队
CARS-25	西甜瓜产业技术体系	CARS-53	绿色低碳共性技术创新团队
CARS-26	柑橘产业技术体系	CARS-54	智慧农业共性技术创新团队
CARS-27	苹果产业技术体系	CARS-55	产业经济共性技术创新团队
CARS-28	梨产业技术体系		

一、基本情况

(一) 岗位科学家情况					
岗位名称	配套系育种				
岗位科学家	张哲	性别	男	出生年月	1984. 03
职称	教授	学历	博士研究生	行政职务	副院长
工作单位	华南农业大学 动物科学学院				
通讯地址/邮编	广州市天河区五山路 483 号 510642				
电话/电子信箱	020-85282019 / zhezhang@scau. edu. cn				
所属功能研究室	遗传改良研究室				
功能研究室主任					
(二) 团队成员情况					
姓名	学历/职称	出生年月	性别	工作单位	电话/邮箱
袁晓龙	博士研究生/ 副教授	1989.04	男	华南农业大学	13560369611/ yxl@scau.edu.cn
高亚辉	博士研究生/ 副教授	1989.09	男	华南农业大学	15652937882/ yahui.gao@scau.edu.cn
李加琪	博士研究生/ 教授	1965.09	男	华南农业大学	13609709797/ jqli@scau.edu.cn
滕金言	博士研究生/ 副教授	1997.07	男	华南农业大学	15521314172/ jinyan.teng@scau.edu.cn

二、重点任务

（一）产业重大关键技术攻关

CARS-35-01A：国家种猪繁育体系优化与新品种培育

1、重要意义（500 字以内）				
<p>（1）产业中存在的问题（需明确产业面临的难点、堵点、卡点、痛点等问题）</p> <p>近年来，我国生猪产业连续受到蓝耳病、非洲猪瘟等的严重影响，生猪种业也发生巨大变革。短期内优质种源供给能力下降，重引进、轻选育，导致核心种源缺乏持续性能改进的源动力。为提升我国生猪种业自主创新能力，减少对外依赖，确保种源供给和安全，面向未来生猪种业发展趋势，重构和优化种猪繁育体系与新品种培育，持续支撑生猪产业高质量发展。</p> <p>（2）重要意义</p> <p>我国每年消耗 5600 万吨猪肉，占全球猪肉消费的一半，庞大的消费能力不能指望国际市场，必须牢牢端在中国人自己的手上。农以种为先，生猪种业是国家基础性、战略性核心产业，是现代生猪产业技术水平的集中体现，是生猪产业科技进步的重要标志，也是猪全产业链中技术含量高、集成度最密集的环节。习近平总书记明确指出“把民族种业搞上去、加快培养具有自主知识产权的优良品种”</p>				
2、研究内容（500 字以内）				
<p>（1）基于种公猪精液共享构建新型种猪繁育体系</p> <p>构建“祖代+父母代”内部循环的种群更新方式，建立基于种公猪站精液传递为主的新型良种猪繁育体系，强化育种场与种公猪站的遗传交流与联系，推动开展区域性联合育种。</p> <p>（2）基于种猪大数据共享平台建设区域性联合育种体系</p> <p>在优秀种公猪遗传资源共享体系建设基础上，搭建种猪育种大数据平台，建立区域性跨场间核心育种群持续、稳定、双向的遗传和数据交流，确保种猪遗传评估准确性和选育效率提高，持续开展种猪联合育种，推动区域联合育种体系的建立。</p> <p>（3）国家核心育种群结构优化与品种（配套系）培育</p> <p>通过实施全国生猪遗传改良计划，结合改良计划生产指标具体要求（包括生长及繁殖性状），优化国家核心育种群结构，持续开展大规模种猪性能测定，推动优质猪大规模产业化应用，培育快长、节粮、高繁等满足市场需求的新品种（配套系）</p> <p>（4）全基因组选择及分子育种技术研发及应用</p> <p>建立高质量的猪分子育种参考群，挖掘猪复杂性状关键基因和调控位点，整合大规模生物学先验信息，开发基因组选择新方法，建立基于基因组信息联合遗传评估体系，推广实施全基因组选择技术。</p>				
3、技术路线（200 字以内）				
<p>基于全产业链，构建基于种公猪站为纽带（基因流）、育种群、生产群与屠宰消费端一体化数据共享（数据流）的种猪繁育体系，收集全产业链各环节育种数据，实现数据与资源的共享。应用杂交育种、全基因组选择等关键技术，培育高繁殖、节粮、快长、抗逆等不同特色新品种（配套系）。</p>				
4、任务分工				
序号	专家姓名	任务角色	岗位名称	任务分工

1	张哲	成员	配套系育种	任务技术路线指定，全基因组选择技术研发与推广应用
2	袁晓龙	成员	配套系育种	关键基因功能挖掘
3	高亚辉	成员	配套系育种	全基因组选择技术研发
4	李加琪	成员	配套系育种	育种共享体系组织实施
5	滕金言	成员	配套系育种	分子设计育种
5、预期结果（300 字以内）				
(1) 基于种公猪站为纽带的种猪繁育体系初步形成； (2) 建立种猪育种大数据、种公猪站联合共享网络体系，核心场间可持续遗传联系机制初步形成； (3) 全基因组选择技术全面应用，核心场育种能力持续增强，新品种（配套系）培育能力显著增强。				
6、考核指标（200 字以内）				
(1) 国家核心育种场种公猪测定量达到3000头，构建全基因组选择参考群5000头以上； (2) 新增繁殖性能记录1万条、性能测定2万条、芯片测定5000条； (3) 筛选种猪配套组合1个，有效提高生产性能； (4) 发表高水平论文3篇，申请专利3项。				

（二）服务县域经济发展

CARS-35-01B：河南省南阳市内乡县

1、产业分析（300 字以内）
<p>内乡县位于河南省西南部，属秦巴片区特困县。全县总面积2465平方公里，辖16个乡镇，288个行政村，总人口73万人。因为人均耕地面积不足1亩，需要从内乡县实际情况出发，因地制宜，探索符合内乡县的乡村振兴发展的新模式和新路径。内乡县在生猪养殖方面，结合自身区域优势，依托南阳综合试验站建设依托单位牧原食品股份有限公司，成立了全国最大的养猪专业合作社。自2018年我国发生第一起非洲猪瘟疫情以来，对我国生猪产业造成了巨大的冲击，通过各类产业资源的整合，区域性生产组织、繁育体系和生猪屠宰体系将重塑，现代猪舍设计、智能养殖与行业大数据将进行深度挖掘与运用。</p>
2、任务内容（500 字以内）
<p>(1) 在推进内乡县乡村振兴过程中，依托南阳综合试验站建设依托单位的发展规划和当地乡村振兴部署，探索适合内乡县发展的“5+”模式，即“基层组织+龙头企业+ 金融机构+ 合作社+ 养殖户”五位一体的畜牧业资源整合模式。</p> <p>(2) 构建“养殖-沼肥-生态农业”的循环经济模式，实现生猪养殖污染治理全覆盖，将内乡县建成国家畜牧业绿色发展示范县。</p> <p>(3) 积极推动农牧装备产业园、牧原智慧物流园和肉食品产业园等在内的 15 个生猪稳产保供建</p>

<p>设项目发展，确保全县生猪养殖项目新增产能 240 万头，生猪屠宰新增产能 400 万头，同时通过复工复产，多方扶持，拉动社会资金新建、改扩建规模化猪场 38 个，全面带动生猪产业转型升级。</p> <p>(4) 积极推动生猪稳产保供政策的落地，全力服务好生猪扩能为核心全生猪全产业链综合体的建设。</p>				
<p>3、工作机制（100 字以内）</p>				
<p>实行任务责任人负责制，张哲教授团队负责品种遗传改良任务，利用全基因组选择育种等技术手段，加快种猪遗传改良进程，并配合疫病防控体系构建、精准营养与豆粕减量技术、粪污处理与资源化利用、生猪屠宰加工与全产业链标准综合体、示范基地建设等任务，促进河南省南阳市内乡县经济发展。</p>				
<p>4、任务分工</p>				
<p>生猪体系参加人员任务分工</p>				
序号	功能研究室名称	岗位名称	专家姓名	任务分工
1	遗传改良	配套系育种	张哲	品种遗传改良任务
2	遗传改良	配套系育种	袁晓龙	品种遗传改良任务
3	遗传改良	配套系育种	高亚辉	品种遗传改良任务
4	遗传改良	配套系育种	李加琪	品种遗传改良任务
5	遗传改良	配套系育种	滕金言	品种遗传改良任务
<p>对接机构</p>				
序号	对接机构	联系人	联系方式	任务分工
1	内乡县畜牧局	王金遂	13949352739	示范推广
2	牧原食品股份有限公司 (南阳综合试验站)	李彦朋	18272763827	成果应用与示范推广
<p>5、考核指标（100 字以内）</p>				
<p>(1) 完成技术指导 5 次以上，培训技术人员和养殖户 100 人次以上；</p> <p>(2) 制定规模化养殖场非洲猪瘟疫情综合防控技术规范 1 项；</p> <p>(3) 完善现代化良种猪繁育模式，创建不同规模生猪养殖场现代化案例 1 个；</p>				

CARS-35-02B: 广东省湛江市遂溪县

<p>1、产业分析（300 字以内）</p>
<p>遂溪县位于广东省西南部的雷州半岛上，林牧资源丰富，是全国生猪调猪大县，2023年，全县生猪养殖场近800家（其中，规模猪场370多家），能繁母猪存栏6万多头，生猪出栏120万头；</p>

<p>遂溪县作为湛江市生猪产业重点发展区，重点发展规模化、标准化生猪养殖场，不仅推动了瘦肉猪的生产，同时加大力度提升广东小耳花猪、黑猪等地方猪资源的开发利用，满足差异化市场需求，为该县脱贫攻坚、发展农村经济、壮大县域经济具有重要作用。根据广东省生猪产业发展规划，目前该县正在创建“正大集团生猪产业园”、“壹号土猪产业园”和“地方猪种业产业园”，三个产业园将为遂溪县带来百万头的生猪产能。</p>				
<p>2、任务内容（500 字以内）</p>				
<p>(1) 结合广东省国家生猪产业体系岗位专家的专业特长以及壹号食品企业发展规划和当地脱贫攻坚部署，帮助遂溪县优化地方猪产业规划和布局。</p> <p>(2) 服务遂溪县精准扶贫，围绕地方猪保种与开发利用、地方猪精准营养、地方猪疫病防治等技术研发与应用，加快广东小耳花猪选育及产业化，大幅提升广东小耳花猪生产效率。</p> <p>(3) 实施“公司+基地+农户”的经营模式，持续完善地方猪繁育体系及养殖技术规程，探索产业链的联合与合作，助力乡村振兴战略。</p> <p>(4) 积极努力就生猪产业科技创新、推动产业提质增效、推进产业绿色发展、促进农民增产增收。</p>				
<p>3、工作机制（100 字以内）</p>				
<p>实行任务责任人负责制，张哲教授团队负责配套组合筛选，利用全基因组选择育种，分子遗传育种等先进育种技术手段，加快新品种（配套系）培育。</p>				
<p>4、任务分工</p>				
<p>**体系参加人员任务分工</p>				
序号	功能研究室名称	岗位名称	专家姓名	任务分工
1	遗传改良	配套系育种	张哲	配套组合筛选
2	遗传改良	配套系育种	袁晓龙	配套组合筛选
3	遗传改良	配套系育种	高亚辉	配套组合筛选
4	遗传改良	配套系育种	李加琪	配套组合筛选
5	遗传改良	配套系育种	滕金言	配套组合筛选
<p>对接机构</p>				
序号	对接机构	联系人	联系方式	任务分工
1	遂溪县农业农村局	杜润清	13822555901	示范推广
2	广东壹号食品股份有限公司（湛江试验站）	曾检华	13929527885	成果应用与示范推广
<p>5、考核指标（100 字以内）</p>				
<p>(1) 带动农户 50 户以上；</p> <p>(2) 推广体系成果 1 项，在地方猪种选育等环节应用；</p> <p>(3) 完成技术指导 5 次以上，培训技术人员和养殖户 100 人次以上。</p>				

(三) 重大突发性事件应急和咨询服务

- 1.监测本产业生产和市场的异常变化，及时向农业农村部上报情况。
- 2.组织开展应急性技术指导和培训工作。
- 3.发生重大自然灾害或重大突发性事件，及时制订分区域的应急预案与技术指导方案，建立专家组，明确工作机制，并以体系的名义上报农业农村部科技教育司。
- 4.加大与大型地方龙头企业的对接力度，开展交流活动，促进科技与产业经济紧密结合。

生猪体系科企对接情况表

序号	企业名称 (全称)	专家 姓名	企业地址	企业类型	服务内容	企业荣誉	企业联络人及 联系方式
1	福建永诚农牧科技集团有限公司	遗传改良研究室李加琪	福建省福清市	生猪育种、养殖	1 育种技术指导	福建省农业产业化省级重点龙头企业	薛永钦， 13859062888
2	广西农垦永新畜牧集团有限公司	遗传改良研究室李加琪	广西省南宁市	生猪育种、养殖	1 育种技术指导 2.种猪精准饲养和管理技术 3.屠宰加工	农业产业化国家重点龙头企业、农业部无公害生猪质量安全追溯试点单位	吴细波， 13737009410， 覃燕灵 15978130817， 姚若存 15877148878
3	广西农垦永新畜牧集团有限公司良圻原种猪场	遗传改良研究室李加琪	广西省南宁市	养殖企业	生猪精准营养需求	农业产业化国家重点龙头企业	覃燕灵 15978130817， 吴先华 1517775660
4	深圳市农牧实业有限公司	遗传改良研究室李加琪	深圳市福田区	养殖屠宰加工企业	国家生猪育种群联合育种体系建立与持续选育	郑华： 13510362488	深圳市农牧实业有限公司

5.国家乡村振兴重点帮扶县科技特派团工作

序号	服务区县	技术顾问/ 团长/组长	专家姓名	工作内容	考核指标
1	会泽县	组长：张哲	李加琪、陈赞谋、袁晓龙、高亚辉、滕金言	组成会泽县养猪产业技术顾问组，优化家庭农场非洲猪瘟的防控措施，高效运行环保及安全体系，优化创新合作模式，培养新型职业农民，促进当地经济发展，实现会	1、非洲猪瘟防控方案及措施1套； 2、优化创新模式2个； 3、技术培训指导10次。

				泽县乡村振兴。	
--	--	--	--	---------	--

6.完成农业农村部各相关司局临时交办的任务

<p>(1) 监测本产业生产和市场的异常变化，及时向产业技术研发中心上报情况。</p> <p>(2) 组织开展应急性技术指导和培训工作。</p> <p>(3) 发生重大自然灾害或重大突发性事件，及时制订分区域的应急预案与技术指导方案，建立专家组，明确工作机制，并以体系的名义上报农业农村部科技教育司。</p>
--

(四) 产业基础数据平台建设

1. 种猪生产企业数据库
负责机构：遗传改良研究室
考核指标（数据量）： 收集全国 20 家省级以上种猪场基本信息数据。
2. 种猪育种数据库
负责机构：遗传改良研究室
考核指标（数据量）： 收集国家生猪核心育种场的系谱登记信息、生产性能测定、繁殖性能记录等。
3. 种猪资源数据库
负责机构：遗传改良研究室
考核指标（数据量）： 完成全国主要饲养引进猪种、地方猪种的基础信息，预计数据量 50 条。

(五) 体系宣传报道计划

无。

二、2024 年经费预算表

科目名称	主 要 用 途	经 费 (万元)
1.设备费	用于在研究开发和试验示范过程中，购置或试制专用仪器设备、购置计算类仪器设备和软件工具、对现有仪器设备进行升级改造、以及租赁外单位仪器设备或购买专用软件授权而发生的费用。	12.00
2.业务费	用于在研究开发和试验示范过程中，消耗的材料、辅助材料等低值易耗品的采购、运输、装卸、整理等费用，发生的测试化验加工、燃料动力、会议/差旅费、出版/文献/信息传播/知识产权事务等费用，以及其他相关支出。	32.50
3.劳务费	用于在研究开发和试验示范过程中，支付给参与体系任务的研究生、博士后、访问学者和科研辅助人员等劳务性费用，其开支标准参照当地科学研究和技术服务业从业人员平均工资水平，根据其在体系研发中承担的工作任务确定，其由单位缴纳的社会保险补助、住房公积金等可纳入劳务费科目支出。	21.00
4.管理费	用于在研究开发和试验示范过程中对使用依托单位现有仪器设备及房屋、试验田，日常水、电、气、暖消耗，以及其他有关管理费用的补助支出。管理费比例不超过扣除设备费后的 8%，由依托单位管理和使用。	4.50
合 计	70.00（万元）	

三、2024 年绩效目标表

名称		生猪现代农业产业技术体系		
总经费		70.00 万元		
年度总体目标	完成任务书中分配的岗位重点任务、体系任务，参与遂溪、内乡2个县域建设科技推进，持续完善种猪场基本信息数据收集，为云南省会泽县提供生猪产业技术帮扶，在生猪产能恢复、种业振兴、非洲猪瘟防控等方面开展相关技术研发与成果推广应用。			
	一级指标	二级指标	三级指标	指标值
绩效指标	产出指标	数量指标	获得新品种（新产品、新设备、新技术新规程等）等技术成果数量	
		数量指标	申报专利数量	2 件
		数量指标	支持担任首席科学家、岗位科学家、综合试验站站长的数量	1 人
		数量指标	科技成果试验示范情况	
		数量指标	科技成果推广应用情况	
		数量指标	培训基层农技推广人员和农民等	500 人
		数量指标	中央级媒体报道次数	
		时效指标	在主产区开展产业应急技术服务反应时间	
	效益指标	社会效益指标	向中央和地方政府、企事业单位提供各类产业分析报告和政策建议等	3 份
		生态效益指标	研发的绿色增产技术或产品在主产区示范应用可实现节水或节肥或节药或省工	
满意度指标	服务对象满意度指标	上级部门满意度	98%以上	
		技术用户满意度	99%以上	

四、共同条款

签约各方共同遵守现代农业产业技术体系建设实施方案（试行）和现代农业产业技术体系管理办法及其它有关规定。

1. 现代农业产业技术体系建设经费要专账管理，专款专用。严格按照《现代农业产业技术体系管理办法》的有关规定和体系经费预算执行。若经费超支，由首席科学家自筹解决，不得影响体系任务执行。

2. 任务执行过程中，如需要修改原有任务和相关指标，须报农业农村部科学技术司审定同意。

3. 首席科学家因不可抗力不能履行任务职责时，应及时以产业技术研发中心正式文件形式报农业农村部科学技术司，并出具不能履行合同的证明材料。

4. 在履行任务职责过程中，由于人为因素导致任务无法完成，视情况追究有关人员责任。

5. 首席科学家依托单位应确保聘任人员和团队成员的稳定，不得随意调换。确需调换，须正式报农业农村部科学技术司同意。

6. 首席科学家要严格履行本任务书的各项指标，每年年底前，须提交体系年度任务执行情况总结报告、经费决算及下年度工作计划。

7. 首席科学家应细化任务书规定的各项指标，并与研究室主任、岗位科学家、综合试验站站长签订任务委托协议。

8. 体系任务书中的重点任务和数据库研发形成的知识产权及成果归国家所有，其管理及使用参照国家有关规定执行。“现代农业产业技术体系”英文名称为“China Agriculture Research System”，缩写为CARS。各体系的名称为“国家**产业技术体系”，英文名称为“CARS—产业英文名称”。形成的知识产权及成果统一标注“Supported by the earmarked fund for CARS”，并注明体系编号。

9. 在体系建设过程中，如有从国外引进的新品种或种质，必须交国家种质资源库统一登记。

10. 体系在执行任务过程中，须按照国家科技保密有关法律法规进行管理。

11. 本任务书经各方签字、盖章后生效。在执行过程中如发生争议、纠纷时，由各方协商解决，或通过法律程序裁决。

五、签约方

国家生猪产业技术体系首席科学家（签字）：陈瑶芝	
2024 年 11 月 26 日	
国家生猪产业技术研发中心依托单位：（公章） 2024 年 11 月 26 日	
依托单位法定代表人（签字）：	
2024 年 11 月 26 日	
功能研究室主任（签字）：	
2024 年 月 日	
岗位科学家（签字）： 张超	
2024 年 月 日	
岗位依托单位：	（公章）
依托单位法定代表人（签字）：	
2024 年 月 日	

合同编号：

技术服务合同

项目名称：种猪基因型检测及分子育种分析服务

委 托 方：福清市永诚畜牧有限公司

（甲 方）福清市永诚畜牧有限公司

受 托 方：华南农业大学

（乙 方）华南农业大学

签订时间：2024 年 03 月 10 日

签订地点：广东省广州市

有效期限：2024 年 1 月 1 日-2024 年 12 月 31 日

中华人民共和国科学技术部印制

填 写 说 明

一、本合同为中华人民共和国科学技术部印制的技术服务合同示范文本，各技术合同认定登记机构可推介技术合同当事人参照使用。

二、本合同书适用于一方当事人（受托方）以技术知识为另一方（委托方）解决特定技术问题所订立的合同。

三、签约一方为多个当事人的，可按各自在合同关系中的作用等，在“委托方”、“受托方”项下（增页）分别排列为共同委托人或共同受托人。

四、本合同书未尽事项，可由当事人附页另行约定，并作为本合同的组成部分。

五、当事人使用本合同书时约定无需填写的条款，应在该条款处注明“无”等字样。

技术服务合同

委托方（甲方）：福清市永诚畜牧有限公司

住 所 地：福建省福清市高山镇薛港村坑北

法定代表人：郭有良

项目联系人：邱定杰

联系方式：13850110641

通讯地址：福建省福清市福人路融商大厦 A 区 1503

电 话： / 传真： /

电子信箱：qiudingjie45@163.com

受托方（乙方）：华南农业大学

住所地：广州市天河区五山街道 483 号

法定代表人：薛红卫

项目联系人：张哲

联系方式：18825084398

通讯地址：广州市天河区五山路 483 号华南农业大学动科新院楼

电 话：020-85295159 传真：

电子信箱：zhezhang@scau.edu.cn

本合同甲方委托乙方就种猪基因型检测及分子育种分析服务项目进行的专项技术服务，并支付相应的技术服务报酬。双方经过平等协商，在真实、充分地表达各自意愿的基础上，根据《中华人民共和国合同法》的规定，达成如下协议，并由双方共同恪守。

第一条：甲方委托乙方进行技术服务的内容如下：

1. 技术服务的目标：开展种猪及后备猪基因型检测及分子育种分析服务。

2. 技术服务的内容：（1）乙方完成甲方送检的核心群种猪及后备猪的基因型检测 1000 头；（2）对送检个体开展品种纯度鉴定、亲缘关系分析，并出具分析报告；（3）对乙方群体开展年度基因组育种分析，并出具分析报告。

3. 技术服务的方式：现场指导和实验室检测及基因型数据分析。

第二条：乙方应按下列要求完成技术服务工作：

1. 技术服务地点：广州市华农农业大学动物科学学院

2. 技术服务期限：2024 年 1 月-2024 年 12 月

3. 技术服务进度：2024 年 1 月-12 月，制定基因型检测方案；参考群样品 DNA 抽提及基因型检测，对检测的基因型进行纯种鉴定，亲子鉴定及亲缘关系分析；对后备公母猪通过与已检公猪进行亲子鉴定分析，鉴定出与原系谱是否匹配。

4. 技术服务质量要求：制定详细样品基因型检测方案；提供技术报告、分析报告、基因组选择评估报告，发表论文 1 篇。

5. 技术服务质量期限要求：合同期限内

第三条：为保证乙方有效进行技术服务工作，甲方应当向乙方提供下列工作条件和协作事项：

1. 提供技术资料：

(1) 参考群后备公猪、后备母猪及其他组织样品等相关材料；

(2) 对采集的样品进行保存、运输。

2. 提供工作条件：

(1) 无。

3. 其他：无。

4. 甲方提供上述工作条件和协作事项的时间及方式：2024年1月-12月，分批次。

第四条：甲方向乙方支付技术服务报酬及支付方式为：

1. 技术服务费总额为：贰拾万元整（¥20.00 万元整）

2. 技术服务费由甲方一次性支付乙方。

具体支付方式和时间如下：

(1) 自签订合同日 30 日内，支付全部技术开发费用（20.00 万元）。（乙方需在收款后 10 个工作日内提供相应金额的广东省行政事业单位资金往来结算票据（电子）给甲方）。

乙方开户银行名称、地址和账号为：

开户银行：中国工商银行广州五山支行

地址：华南农业大学

账号：3602002609000310520

第五条：双方确定因履行本合同应遵守的保密义务如下：

甲方：

1. 保密内容（包括技术信息和经营信息）：研究结果及技术资料。

2. 涉密人员范围: 参与项目的所有人员。

3. 保密期限: 3 年。

4. 泄密责任: 赔偿乙方总研发经费的三倍。

乙方:

1. 保密内容 (包括技术信息和经营信息): 研究结果及技术资料。

2. 涉密人员范围: 参与项目的所有人员。

3. 保密期限: 3 年。

4. 泄密责任: 赔偿甲方总研发经费的三倍。

第六条: 本合同的变更必须由双方协商一致, 并以书面形式确定。但有下列情形之一的, 一方可以向另一方提出变更合同权利与义务的请求, 另一方应当在 15 日内予以答复; 逾期未予答复的, 视为同意:

1. 甲方要求乙方提前完成技术开发项目或者赶工的。

2. 甲方在承诺对乙方已经发生的费用给予补偿的前提下提出终止本合同的;

3. 乙方因甲方未按照合同约定提供技术资料或者工作条件、协作事项而推迟服务日期的;

4. 一方向对方提出书面索赔请求的。

第七条: 双方确定以下列标准和方式对乙方的技术服务工作成果进行验收:

1. 乙方完成技术服务工作的形式: 提供基因型检测报告及技术分析报告。

2. 技术服务工作成果的验收标准：基因组检测报告 1 份，技术分析报告 1 份，基因组选择评估报告 1 份，合作发表论文 1 篇（署名共同单位）。

3. 技术服务工作成果的验收方法：合同约定的内容，由甲方进行验收。

4. 验收的时间和地点：2024 年 12 月 31 日，由甲方指定验收地点。

第八条：双方确定：

1. 在本合同有效期内，甲方利用乙方提交的技术服务工作成果所完成的新的技术成果，归甲方所有。

2. 在本合同有效期内，乙方利用甲方提供的技术资料和工作条件所完成的新的技术成果，归甲乙双方所有。

第九条：双方确定，按以下约定承担各自的违约责任：

1. 甲方违反本合同第四条约定，应当每迟延 1 日，按照合同价款的 2%向乙方支付违约金，迟延超过 30 日时，乙方有权利单方面解除本合同并就本方受到的损失进行索赔。

2. 乙方违反本合同第七条约定，应当每迟延 1 日，按照合同价款的 2%向乙方支付违约金，迟延超过 30 日时，甲方有权利单方面解除本合同并就本方受到的损失进行索赔。

第十条：双方确定，在本合同有效期内，甲方指定邱定杰为甲方项目联系人，乙方指定张哲为乙方项目联系人。项目联系人承担以下责任：

1. 负责项目的管理、执行、信息交换、总结，协调双方协作事项。

一方变更项目联系人的，应当及时以书面形式通知另一方，未及时通知并影响本合同履行或造成损失的，应承担相应的责任。

第十一条：双方确定，出现下列情形，致使本合同的履行成为不必要或不可能的，可以解除本合同：

1. 发生不可抗力；

2. 因甲方安全生产、技术改进等相关原因需要终止的。

第十二条：双方因履行本合同而发生的争议，应协商、调解解决。协商、调解不成的，确定按以下第1种方式处理：

1. 提交广州仲裁委员会仲裁；

2. 依法向人民法院起诉。

第十三条：双方确定：本合同及相关附件中所涉及的有关名词和技术术语，其定义和解释如下：

1. 无

第十四条：与履行本合同有关的下列技术文件，经双方确认后，1为本合同的组成部分：

1. 技术背景资料：基因型检测方案。

第十五条：双方约定本合同其他相关事项为：无

第十六条：本合同一式肆份，具有同等法律效力。

第十七条：本合同经双方签字盖章后生效。

甲方：_____ (盖章)
法定代表人 / 委托代理人：_____ (签名)
年 月 日



乙方：_____ (盖章)
法定代表人 / 委托代理人：_____ (签名)
年 月 日



印花税票粘贴处：

(以下由技术合同登记机构填写)

合同登记编号：

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

1. 申请登记人：_____
2. 登记材料：(1) _____
(2) _____
(3) _____
3. 合同类型：_____
4. 合同交易额：_____
5. 技术交易额：_____

技术合同登记机构（印章）

经办人：

年 月 日

广东省农业科研类及技术推广示范类 项目（专项）实施方案

项目（课题）名称：广东省现代畜牧业高质量发展示范推广

承 担 单 位：广东省农业技术推广中心

项 目 负 责 人：刘建营

起 止 年 限：2024年1月1日至2024年12月31日

广东省农业农村厅制

广东省农业农村厅农业科研类及技术推广示范类 项目（课题）信息表

项目（课题）名称		广东省现代畜牧业高质量发展示范推广			
密 级		<input type="checkbox"/> 绝密 <input type="checkbox"/> 机密 <input type="checkbox"/> 秘密 <input checked="" type="checkbox"/> 公开			
承担单位	名 称	广东省农业技术推广中心			
	单位所在地	广东省天河区		代码	12440000455858409J
	通讯地址	广东省天河区柯木塱南路 28 号		邮编	
	开户银行	广州银行沙河支行			
	银行帐号	2988 0001 7265			
	单位性质	<input type="checkbox"/> 事业型研究单位 <input checked="" type="checkbox"/> 其他事业单位 <input type="checkbox"/> 大专院校 <input type="checkbox"/> 转制为企业的科研院所 <input type="checkbox"/> 国有企业 <input type="checkbox"/> 集体所有制企业 <input type="checkbox"/> 合资企业 <input type="checkbox"/> 外商投资企业 <input type="checkbox"/> 港、澳、台投资企业 <input type="checkbox"/> 其他企业		代码	12440000455858409J
	上级行政主管部门	广东省农业农村厅		代码	
	国务院国资委企业	<input type="checkbox"/> 是 <input checked="" type="checkbox"/> 否		“双一流”大学	<input type="checkbox"/> 是 <input checked="" type="checkbox"/> 否
参与单位	序号	单位名称			
	1	广东省农业科学院动物科学研究所			
	2	华南农业大学			
	3	广东省养猪行业协会			
项目（课题）负责人	姓 名	刘建营		性 别	<input checked="" type="checkbox"/> 男 <input type="checkbox"/> 女
	学 位	<input type="checkbox"/> 博士 <input checked="" type="checkbox"/> 硕士 <input type="checkbox"/> 学士 <input type="checkbox"/> 其他		出生日期	1980
	职 称	<input checked="" type="checkbox"/> 高级 <input type="checkbox"/> 中级 <input type="checkbox"/> 初级 <input type="checkbox"/> 其他		专 业	畜牧兽医
	所在单位	广东省农业技术推广中心			
	身份证件	居民身份证	证件号码	412829198009161232	
	联系电话	13760845058		E-mail	59069847@qq.com
项目（课题）主要内容		1. 现代化美丽牧场与畜禽养殖标准化示范创建； 2. 开展奶牛生产性能测定； 3. 高海拔地区适养型黄羽肉鸡优选及示范； 4. 开展广东生猪产能调控调研相关工作，开展畜牧业统计分析预警。			

一、目标与任务

(①项目(课题)研究/推广示范内容及任务分解:要解决的主要技术难点和问题,技术方案和创新点等;②项目(课题)研究/推广示范目标;③主要示范和产业化内容及相关技术路线)

1 现代化美丽牧场与畜禽养殖标准化示范创建

1.1 主要内容:制定《2024年广东省畜禽养殖标准化示范创建活动工作方案》、《2024年广东省现代化美丽牧场示范创建活动工作方案》、《关于开展2024年国家标准化示范场创建活动的通知》,通过组织申报、专家现场验收、网上公示等,遴选省现代化美丽牧场和国家畜禽养殖标准化示范场,示范带动全省畜牧业转型升级高质量发展。

1.2 技术方案:通过制定2024省现代化美丽牧场、国家和省级畜禽养殖标准化示范场创建、验收方案,组织行业专家开展现场验收,对创建结果进行报道宣传。

1.3 主要解决的技术难点和问题:制定实施方案组织专家现场评审,受非洲猪瘟疫情影响,现场评审时间存在不确定性。

1.4 创新点:推进广东畜牧业转型升级高质量发展,在全国范围内率先在政策制度上创新,鼓励企业进行现代化美丽、省级标准化示范场创建,大力推进了质量兴牧、绿色兴牧和我省畜牧业现代化,全面提升了我省畜牧业质量效益竞争力。

1.5 示范目标:创建一批(最少10个以上)质量高、示范性强的现代化美丽牧场,100个以上畜禽标准化示范场,示范带动全省畜牧业高质量发展。

2. 开展奶牛生产性能测定

2.1. 主要内容:开展2000头奶牛生产性能测定工作,主要测定指标:乳蛋白、乳脂、乳糖、尿素氮、体细胞等并形成参测牧场DHI报告;购置自动洗瓶机器人1台,提高实验室效率,让实验室更多更好的服务牧场;举办1期奶牛相关技术培训,培训人数50人以上。

1.2 技术方案:(1)DHI测定:通过每月统一采集一次每头泌乳牛的奶样,分析奶牛泌乳性能、乳成分以及体细胞计数等。结合收集牛群的饲养管理与奶牛品种登记资料,利用专用的分析软件分析资料,形成反映奶牛场配种繁殖、饲养管理、生产性能、乳房保健、疾病防治等方面准确信息,即DHI报告。形成牧场存在的问题分析与改进措施、牧场未来发展的建议,牛场管理人员可充分理解并利用生产性能测定报告和反馈建议,科学有效地加强管理、充分发挥牛群的生产潜力,进而提高生产效率和经济效益。(2)提高DHI实验室效率:主要购置1台自动清洗机器人及其他设备,提高实验室测定效率以及测定及时性,提高中心的质量检测能力。

1.3 主要解决的技术难点和问题:本项目基于目前DHI技术快速发展过程中存在的一系列问题,如奶牛场的管理及技术人员对DHI测定报告的读解能力欠缺,DHI测定工作与奶牛群体改良和奶牛科学管理尚未有机结合起来等问题,构建多方协同的DHI一体化服务模式,精细解读DHI测定报告,及时地发现牛群及个体在繁殖育种、营养状况、饲料配方、生产性能及奶牛疾病和乳品质等方面存在的问题,并在专家和技术人员的指导下,制定合理方案,为奶牛场提出指导性建议,并通过相关技能培训和现场指导,使

整个奶牛群体达到最优化的生产和控制状态、最优化资源配置，提高生产效率和经济效益。

1.4 创新点：在开展奶牛生产性能测定，保障测定效率的基础上，开展精细解读 DHI 测定报告，通过将 DHI 分析报告和生产建议，场技术服务和示范推广有机结合，将 DHI 一体化服务模式应用于全省奶牛饲养管理中心，确保奶牛生产性能测定技术在奶牛养殖领域的推广和落实，对促进我省奶牛养殖行业的健康发展具有重要意义，同时也促进了产业需要和科技创新。

1.5 示范目标：完成我省 2000 头奶牛生产性能测定工作任务，主要测定指标：乳蛋白、乳脂、乳糖、尿素氮、体细胞等并形成参测牧场 DHI 报告；购置自动洗瓶机器人一台，提高实验室效率，让实验室更多更好的服务牧场；举办 1 期奶牛相关技术培训，培训人数 50 人以上。

3. 高海拔地区适养型黄羽肉鸡优选及示范

3.1 主要内容：开展高海拔地区黄羽肉鸡产业发展状况调研，落实用于品种（配套系）筛选所需的养殖场区，搭建肉鸡养殖所需的软硬件条件。结合当地情况，针对试养品种（配套系）提供指导意见，并进行定期调研和分析。在高海拔适宜区域开展良种良法的示范推广。

3.2 技术方案：本项目将从项目承担单位已收集（培育）的丰富的黄羽肉鸡种质资源入手，通过初步的筛选 3-5 个黄羽肉鸡品种（配套系）开展饲养试验，通过系统的高原适应性研究与评估，优选高海拔地区适养型肉鸡，并形成健康养殖技术，开展良种良法的示范推广。

3.3 主要解决的技术难点和问题：通过项目的实施，可以进一步扩大黄羽肉鸡品种（配套系）良种辐射半径、丰富高海拔地区黄羽肉鸡种类，满足日益增长的消费需要，助力高海拔地区肉鸡产业提质增效。

3.4 创新点：本项目的顺利实施，有望优选出高海拔地区适养型黄羽肉鸡品种（配套系），并形成配套健康养殖技术，通过良种良法的示范推广，丰富当地肉鸡种类、扩大黄羽肉鸡良种辐射半径。

3.5 示范目标：初选黄羽肉鸡品种（配套系）3-5 个；完成 3-5 个肉鸡品种（配套系）的适应性研究与评价；优选高海拔地区适养型黄羽肉鸡品种（配套系）1 个；研究并建立配套健康养殖技术 1 套。

4. 生猪产能调控，畜牧业统计分析

4.1 主要内容：广东生猪产能调控调研相关工作，开展全省畜牧业统计分析预警，做好直联直报系统填报、工作督导和数据审核，编写畜牧业生产形势分析报告，组织开展生猪产能形势分析座谈和深度调研，开展生猪生产大数据挖掘及数据平台建设，对生猪生产形势进行研判，形成生猪产能预警分析报告。

4.2 技术方案：省农业技术推广中心负责全省生猪生产固定监测县、价格监测县以及生猪规模养殖场数据的采集、审核及实地核查工作，并将每月采集的数据提供给华南农业大学进行统计分析，形成《广东畜牧业简报》。广东省养猪行业协会通过组织生猪产能调控联盟单位、行业专家开展生猪产能形势分析座谈和调研，对生猪生产形势进行研判，形成生猪产能预警分析报告。在生猪价格大幅波动期间，负责调查价格形成的原因，并形成专报，供行业主管部门参考。

4.3 主要解决的技术难点和问题：构建上下联动、响应及时的生猪生产逆周期调控

机制，促进生猪产业持续健康发展，不断提升猪肉供应安全保障能力。

4.4 创新点：生猪生产分析报告内容要涉及生产效益指标，生产指标要涵盖生猪存栏、能繁母猪存栏、生猪出栏等关键指标，效益指标要涵盖养殖成本、出栏价格、出栏体重等关键指标。

4.5 示范目标：完成全省畜牧业调查统计与分析预警，形成《广东畜牧业简报》12期；建设广东省生猪产能调控机制，组织召开4次生猪产能调控分析会，每次参会人数不少于60人。

二、预期成果及考核指标

（①主要技术指标：如形成的知识产权、技术标准、新技术、新产品、新装置、论文专著等数量、指标及其水平等；②主要经济指标：如技术及产品应用所形成的市场规模、效益等；③项目（课题）实施中形成的示范基地、中试线、生产线及其规模等；④人才队伍建设；⑤其他应考核的指标。）

（1）现代化美丽牧场与畜禽养殖标准化示范创建：创建一批（最少10个以上）质量高、示范性强的现代化美丽牧场，100个以上畜禽标准化示范场，示范带动全省畜牧业高质量发展；

（2）开展奶牛生产性能测定：完成我省2000头奶牛生产性能测定工作任务，主要测定指标：乳蛋白、乳脂、乳糖、尿素氮、体细胞等并形成参测牧场DHI报告；购置自动洗瓶机器人1台，提高实验室效率，让实验室更多更好的服务牧场；举办1期奶牛相关技术培训，培训人数50人以上。

（3）初选黄羽肉鸡品种（配套系）3-5个；在高海拔地区完成3-5个肉鸡品种（配套系）的适应性研究与评价；优选高海拔地区适养型黄羽肉鸡品种（配套系）1个；研究并建立配套健康养殖技术1套。

（4）完成全省畜牧业调查统计与分析预警，形成《广东畜牧业简报》12期；建设广东省生猪产能调控机制，组织召开4次生猪产能调控分析会，每次参会人数不少于60人。

三、项目（课题）年度计划及年度目标

1 现代化美丽牧场与畜禽养殖标准化示范创建

1.1 创建筹备阶段（2024年1月1日-2024年5月30日），开展调研，收集资料，广泛征求专家和各地畜牧部门意见。

1.2 组织申报阶段（2024年6月1日-2024年8月31日），省农业农村厅发印发工作方案，组织各地积极开展申报，组织开展技术培训。

1.3 专家评选阶段（2024年9月1日-2024年10月31日），组织专家按创建验收标准验收，对经申报材料审核合格的企业（养殖场）开展现场审核验收。

1.4 总结推广阶段（2024年11月1日-2024年12月31日），充分利用现场会、报刊杂志和图片视频等，广泛宣传推广，辐射带动发展。

2 开展奶牛生产性能测定

2.1 2024年01月-2024年12月：每月进行DHI采样及测定，参测牧场报告编制及解读。

2.2 2024年01月-2024年02月：仪器设备询价及技术咨询；

2.3 2024年03月-2024年04月：仪器设备购置申请；

2.4 2024年05月-2024年10月：完成设备购置。

3 高海拔地区适养型黄羽肉鸡优选及示范

2024年1月~2024年6月

（1）完成高海拔地区黄羽肉鸡产业发展状况调研，落实用于品种（配套系）筛选所需的养殖场区，搭建肉鸡养殖所需的软硬件条件。

（2）根据市场需求，针对试养品种（配套系）提供指导意见，对黄羽肉鸡品种（配套系）进行初选。

2024年7月~2024年12月

（1）完成3-5个黄羽肉鸡品种（配套系）适应性研究和系统评估。

（2）综合适应性研究结果与上市肉鸡市场认可情况、经济效益分析等优选出高海拔地区适养型黄羽肉鸡品种（配套系）1个。

（3）开展良种良法的示范推广，建设示范基地1个。

4 生猪产能调控，畜牧业统计分析

筹备准备阶段（2023年12月1日-2023年12月31日），组织项目

参与单位制定项目推进时间计划表。

资料编写与发布阶段（2024年1月1日-2024年12月31日），组织开展全省畜牧业统计调研调查，收集汇总数据，获得畜牧业生产第一手资料，召开畜牧业生产形势分析会，编写畜牧业生产形势分析报告。

四、任务分工情况

任务名称	承担单位	任务负责人	研究/推广示范内容	项目资金 (万元)	其中：财政 资金(万元)
现代化美丽牧场与畜禽养殖标准化示范创建；奶牛生产性能测定。	广东省农业技术推广中心	刘建营	作为牵头单位，统筹项目的实施；开展好现代化美丽牧场与畜禽养殖标准化示范创建；奶牛生产性能测定。	80	80
高海拔地区适养型黄羽肉鸡优选及示范	广东省农业科学院动物科学研究所	李莹	高海拔地区适养型黄羽肉鸡优选及示范。	60	60
畜牧业生产统计分析预警	华南农业大学	李加琪	畜牧业生产统计分析预警等。	30	30
生猪产能调控，生产形势分析	广东省养猪行业协会	刘小红	生猪产能调控，生产形势分析。	20	20

五、项目（课题）资金支出预算

总投入经费：（单位：万元）	
专项资金： 190	自筹资金：

（一）项目（课题）省级财政专项资金预算

单位：万元（保留两位小数）

序号	预算科目名称	合计	省级财政专项资金	其他来源资金
	(1)	(2)	(3)	(4)
1	一、资金支出	190	190	
2	（一）直接费用	190	190	
3	1. 设备费	43	43	
4	2. 材料费	4	4	
5	3. 测试化验加工费			
6	4. 燃料动力费			
7	5. 出版/文献/信息传播/知识产权事务费			
8	6. 会议/差旅/国际合作交流费	12	12	
9	7. 培训费	5	5	
10	8. 劳务费			
11	9. 专家咨询费	6	6	
12	10. 维修维护费	8	8	
13	11. 其他支出	2	2	
14	12. 委托业务费	110	110	
14	（二）间接费用			
	二、项目资金来源			
15	（一）省级财政专项资金	190	190	
16	（二）其他来源资金			
17	1. 单位自筹资金			
18	2. 其他资金			

(二) 项目(课题)单位自筹资金预算

序号	预算科目名称	合计	用途说明	其中：承担单位	其中：参与单位
1	一、资金支出				
2	(一) 直接费用				
3	1. 设备费				
4	2. 材料费				
5	3. 测试化验加工费				
6	4. 燃料动力费				
7	5. 出版/文献/信息传播/知识产权事务费				
8	6. 会议/差旅/国际合作交流费				
9	7. 培训费				
10	8. 劳务费				
11	10. 专家咨询费				
12	11. 基本建设费				
13	(1) 房屋建筑物构建				
14	(2) 专用设备购置				
15	(3) 基础设施建设				
16	(4) 大型修缮				
17	(5) 信息网络建设				
18	(6) 其他基本建设支出				
19	12. 其他费用				

六、项目（课题）单位提供的技术与条件保障

（包括现有技术基础和承诺提供的支撑条件，如仪器设备、水电、燃料、环保等条件）

项目牵头单位（广东省农业技术推广中心）：广东省农业技术推广中心是广东省农业农村厅直属副厅级公益一类事业单位，现内设 6 个副处级机构，分别是综合部、计划财务与资产管理部、种植业技术与种业推广部、畜牧技术推广部、渔业技术推广部、农业机械化技术与鉴定部，事业编制 149 名。设主任 1 名，专职副书记 1 名、副主任 3 名；其中高级职称人数占 60% 以上。

中心共有四个基地，其中种植业基地约 750 亩和畜牧业基地约 100 亩在广州市天河区柯木塱南路、农机化基地约 18 亩在白云区同沙路、淡水鱼养殖基地约 604 亩在南沙区东涌镇骏稳路，海水鱼养殖基地约 50 亩在惠州市大亚湾澳头镇衙前村边，中心总部位于广州天河区柯木塱南路 28-30 号。

中心主要负责拟订并组织实施全省农业技术推广体系建设规划；承担现代种业发展与技术服务工作；负责实施重大农业技术推广和生产技术攻关项目；承担农业新技术、新品种、新产品、新模式、新装备的引进、集成、试验、示范、推广应用；承担畜牧业调查统计、种畜禽生产性能测定与质量检测；承担渔业科研试验、成果转化及种质资源养护；承担农业机械产品、技术的试验鉴定和检测、示范推广；开展农业技术培训指导、咨询服务、合作交流等工作。

畜牧技术推广部（广东省种畜禽质量检测中心）：负责拟定并组织实施全省畜牧技术推广体系建设；承担畜牧业和畜禽良种关键技术的引进、试验、示范推广工作；承担畜禽统计工作；承担畜禽良种登记、种畜禽生产性能测定与质量检测等工作；承担畜禽遗传资源的调查、保护与利用等技术指导与服务工作。部门现在编在岗人员 13 人，其中正高资格 5 人，副高资格 3 人；具有博士学历 2 人，硕士学历 6 人。部门承建的农业农村部种猪质量检验检测中心（广州）通过了农业农村部组织的“二加一”复审认证，长期承担国家和省种畜禽质量安全抽检任务，承建的广东省奶牛生产性能测定实验室，通过全国畜牧总站评审，每年都承担检测任务。

项目参与单位（华南农业大学）

华南农业大学是国家“双一流”建设高校，隶属广东省教育厅，学校学科门类齐全，有101个本科专业，14个博士学位授权一级学科，1个博士专业学位类别，30个硕士学位授权一级学科，19个硕士专业学位类别。作物学入选国家“一流建设学科”，获批10个广东省高水平大学建设计划重点建设学科。植物学与动物学学科进入ESI全球排名前1‰。现有国家生猪种业工程技术研究中心、人兽共患病防控制剂国家地方联合工程实验室、畜禽育种国家地方联合工程研究中心（广东）、畜禽产品精准加工与安全控制技术国家地方联合工程研究中心（广东）、国家非洲猪瘟区域实验室（广州）等9个国家级科研平台，省部级科研平台104个，广东省高校特色新型智库3个。坚持“四个面向”，服务实现高水平科技自立自强。近五年来，获国家科学技术奖12项，实现国家科技进步、技术发明和自然科学三大奖全覆盖。

项目参与单位（广东省农业科学院动物科学研究所）

广东省农业科学院动物科学研究所，主要从事畜禽遗传育种、动物营养与饲料科学、水产科学、生态养殖与环境控制、草食动物与草业科学、系统微生物与合成生物学等方面的研究与技术开发。依托猪禽种业全国重点实验室等科研平台，收集保存10个地方特色鸡种，自主构建了省内唯一仍可持续构建的资源群体（18个世代），创制出专门化品系38个，其中3个为国家畜禽新品种，岭南黄鸡系列配套系在全国黄羽肉鸡市场占有率超过15%，对外供种能力国内领先。

项目参与单位（广东省养猪行业协会）

广东省养猪行业协会是由从事生猪生产、屠宰加工、贸易、疾病防治、教育科研与管理部门、设备、药械、饲料和饲料添加剂等相关行业代表组成的非盈利性社团组织。协会成立以来为省政府在养猪行业方面出谋献策，在政府与行业之间，发挥了纽带和桥梁作用，每年召开4-5次学术交流或专业技术报告会，先后和国内外进行了多次学术交流，协助会员单位解决生产中的技术难题，促进广东省养猪业的发展。

七、参与人员

序号	姓名	性别	工作单位	职务/职称	项目分工
1	刘建营	男	广东省农业技术推广中心	正高级	负责人
2	陈迎丰	男	广东省农业技术推广中心	正高级	项目骨干
3	曹长仁	男	广东省农业技术推广中心	副高级	项目骨干
4	郭建超	男	广东省农业技术推广中心	副高级	项目骨干
5	李品红	女	广东省农业技术推广中心	副高级	项目骨干
6	李亮	男	广东省农业技术推广中心	中级	项目骨干
7	许华钊	男	广东省农业技术推广中心	中级	项目骨干
8	樊福好	男	广东省农业技术推广中心	正高级	项目骨干
9	李莹	女	广东省农业科学院动物科学研究所	副高级	项目骨干
10	杜宗亮	男	广东省农业科学院动物科学研究所	正高级	项目骨干
11	罗威	男	广东省农业科学院动物科学研究所	中级	项目骨干
12	徐详	男	西藏德祥农牧科技有限公司	其他	项目骨干
13	罗成龙	男	广东省农业科学院动物科学研究所	正高级	项目骨干
14	张丽	女	广东省农业科学院动物科学研究所	其他	项目骨干
15	谢水华	男	广东省农业技术推广中心	正高级	项目骨干
16	李加琪	男	华南农业大学	正高级	项目骨干
17	刘小红	男	广东省养猪行业协会	正高级	项目骨干

八、审核意见

项目（课题）承担单位意见：

本单位对以上内容的真实性和准确性负责。



项目（课题）协作（参与）承担单位意见：



项目（课题）协作（参与）承担单位意见：



项目（课题）协作（参与）承担单位意见：



项目主管单位意见：



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 高亚辉 15 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	Deciphering genetic characteristics of South China and North China indigenous pigs through selection signatures	BMC GENOMICS 出版年: 2024 出版日期: DEC 18 卷期: 25 1 页码: - 文献号: 1191 文献类型: Article	第一作者	A 类	华南农业大学动物科学学院	SCI	IF2-year=3.7 IF5-year=4.2 (2024)	生物学 2 区 Top 期刊: 否 (2025)
2	Transcriptomic profiling of gastrointestinal tracts in dairy cattle during lactation reveals molecular adaptations for milk synthesis	JOURNAL OF ADVANCED RESEARCH 出版年: 2025 出版日期: MAY 卷期: 71 页码: 67-80 文献类型: Article	第一作者	T2 类	美国农业部、美国马里兰大学、华南农业大学动物科学学院	SCI	IF2-year=13.0 IF5-year=11.6 (2024)	综合性期刊 1 区 Top 期刊: 是 (2025)
3	PigBiobank: a valuable resource for understanding genetic and biological mechanisms of diverse complex traits in pigs	NUCLEIC ACIDS RESEARCH 出版年: 2024 出版日期: JAN 5 卷期: 52 D1 页码: D980-D989 文献类型: Article	并列第一作者	A 类	华南农业大学动物科学学院	SCI	IF2-year=13.1 IF5-year=16.8 (2024)	生物学 2 区 Top 期刊: 是 (2025)

4	A compendium of genetic regulatory effects across pig tissues	NATURE GENETICS 出版年: 2024 出版日期: JAN 卷期: 56 1 页码: - 文献类型: Article	并列第一作者	T2 类	华南农业大学 动物科学学院、美国农业部、美国马里兰大学	SCI	IF2-year=29.0 IF5-year=37.4 (2024)	生物学 1 区 Top 期刊: 是 (2025)
5	Genome-wide association analysis of heifer livability and early first calving in Holstein cattle	BMC GENOMICS 出版年: 2023 出版日期: OCT 21 卷期: 24 1 页码: - 文献号: 628 文献类型: Article	第一作者	A 类	美国马里兰大学、美国农业部	SCI	IF2-year=3.5 IF5-year=4.1 (2023)	生物学 2 区 Top 期刊: 是 (2023)
6	Comparative transcriptome in large-scale human and cattle populations	GENOME BIOLOGY 出版年: 2022 出版日期: AUG 22 卷期: 23 1 页码: - 文献号: 176 文献类型: Article	并列第一作者	T2 类	美国农业部、 美国马里兰大学	SCI	IF2-year=12.3 IF5-year=17.4 (2022)	生物学 1 区 Top 期刊: 是 (2022)
7	A multi-tissue atlas of regulatory variants in cattle	NATURE GENETICS 出版年: 2022 出版日期: SEP 卷期: 54 9 页码: 1438-+ 文献类型: Article	并列第一作者	T2 类	美国农业部、 美国马里兰大学	SCI	IF2-year=30.8 IF5-year=37.4 (2022)	生物学 1 区 Top 期刊: 是 (2022)

8	Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods	GENES 出版年: 2022 出版日期: MAY 卷期: 13 5 页码: - 文献号: 828 文献类型: Article	第一作者	B 类	美国农业部、 美国马里兰大学	SCI	IF2-year=3.5 IF5-year=3.9 (2022)	生物学 3 区 Top 期刊: 否 (2022)
9	Single-cell transcriptomic and chromatin accessibility analyses of dairy cattle peripheral blood mononuclear cells and their responses to lipopolysaccharide	BMC GENOMICS 出版年: 2022 出版日期: APR 30 卷期: 23 1 页码: - 文献号: 338 文献类型: Article	并列第一作者	A 类	山东省农科院、美国农业部	SCI	IF2-year=4.4 IF5-year=4.7 (2022)	生物学 2 区 Top 期刊: 是 (2022)
10	Towards the detection of copy number variation from single sperm sequencing in cattle	BMC GENOMICS 出版年: 2022 出版日期: MAR 17 卷期: 23 1 页码: - 文献号: 215 文献类型: Article	并列第一作者	A 类	美国农业部、 美国马里兰大学	SCI	IF2-year=4.4 IF5-year=4.7 (2022)	生物学 2 区 Top 期刊: 是 (2022)
11	Genome-wide recombination map construction from single sperm sequencing in cattle	BMC GENOMICS 出版年: 2022 出版日期: MAR 5 卷期: 23 1 页码: - 文献号: 181	并列第一作者	A 类	美国农业部、 美国马里兰大学	SCI	IF2-year=4.4 IF5-year=4.7 (2022)	生物学 2 区 Top 期刊: 是 (2022)

		文献类型: Article						
12	Functional annotation of regulatory elements in cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning	GENOMICS 出版年: 2022 出版日期: MAR 卷期: 114 2 页码: - 文献号: 110296 文献类型: Article	第一作者	B 类	美国农业部、 美国马里兰大学	SCI	IF2-year=4. 4 IF5-year=4. 2 (2022)	生物学 3 区 Top 期刊: 否 (2022)
13	Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning	GENOMICS 出版年: 2021 出版日期: JUL 卷期: 113 4 页码: 2045-2055 文献类型: Article	第一作者	B 类	美国农业部、 美国马里兰大学	SCI	IF2-year=4. 31 IF5-year=4. 38 (2021)	生物学 3 区 Top 期刊: 否 (2021)
14	Genome-wide association study of Mycobacterium avium subspecies Paratuberculosis infection in Chinese Holstein	BMC GENOMICS 出版年: 2018 出版日期: DEC 27 卷期: 19 页码: - 文献号: 972 文献类型: Article	第一作者	A 类	中国农业大学	SCI	IF2-year=3. 501 IF5-year=4. 142 (2018)	生物 2 区 Top 期刊: 否 (2018)

15	Short communication: Heritability estimates for susceptibility to Mycobacterium avium ssp paratuberculosis infection in Chinese Holstein cattle	JOURNAL OF DAIRY SCIENCE 出版年: 2018 出版日期: AUG 卷期: 101 8 页码: 7274-7279 文献类型: Article	第一作者	A 类	中国农业大学	SCI	IF2-year=3.082 IF5-year=3.208 (2018)	农林科学 2 区 Top 期刊: 是 (2018)
----	---	--	------	-----	--------	-----	--	---------------------------------

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效



华南农业大学图书馆SCAU LIB202519204

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 高亚辉 2 篇论文收录情况如下表。

序号	论文名称	发表刊物及发表的年月卷期/页码等	作者排名	论文等级	作者文中单位	收录情况	影响因子	中科院大类分区
1	Benchmarking 24 combinations of genotype pre-phasing and imputation software for SNP arrays in pigs	Journal of Integrative Agriculture 出版年：2024 卷期： 页码： - 文献号： 文献类型：期刊论文	通讯作者	T2 类	华南农业大学 动物科学学院	已发表， 暂未被 SCI 收录	IF2-year=4.4 IF5-year=4.8 (2024)	农林科学 1 区 Top 期刊：是 (2025)
2	基于全基因组重测序检测中国地方猪的体型选择信号	中国畜牧兽医 出版年：2024 出版日期：2024-07-30 卷期：51 8 页码：3438-3446 文献号： 文献类型：期刊论文	通讯作者	B 类	华南农业大学 动物科学学院	北大核心	无	无

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明:如未盖章,报告无效



RESEARCH

Open Access



Deciphering genetic characteristics of South China and North China indigenous pigs through selection signatures

Yahui Gao^{1†}, Xueyan Feng^{1†}, Shuqi Diao¹, Yuqiang Liu¹, Zhanming Zhong¹, Xiaotian Cai¹, Guangzhen Li¹, Jinyan Teng¹, Xiaohong Liu², Jiaqi Li¹ and Zhe Zhang^{1*}

Abstract

Background Indigenous pig breeds in China have accumulated significant genetic diversity due to regional selection pressures. Investigating the selection signatures of these populations helps to understand their adaptive evolution and contributes to genetic improvement programs.

Results We collected whole-genome sequencing data from 133 individuals, including South China and North China indigenous pigs and Asian wild boars. After data filtering, we retained 31,521,978 high-quality SNPs. Population structure analysis using PCA revealed distinct genetic clustering among these populations. Selection signature detection identified 5,227 loci under selection in South China indigenous pigs and 5,800 in North China indigenous pigs compared to Asian wild boars. Candidate genes were enriched in immune response pathways, reproductive traits, and pigmentation pathways. South China indigenous pigs exhibited selection signals for fat deposition and immune responses, while North China indigenous pigs showed stronger signals related to growth, blood physiology, and reproductive performance. Additionally, key genes such as *MC1R* and *KIT* were associated with coat color variation, and *IGF1R* and *IGF2R* were linked to growth regulation.

Conclusion Our results demonstrate that indigenous pigs in China have undergone selection for distinct traits aligned with their regional environments and farming systems. South China indigenous pigs have been selected for traits related to fat deposition and immunity, while North China indigenous pigs have been selected for growth and reproductive traits. The findings offer crucial insights into the genetic architecture of indigenous pig breeds, providing a valuable foundation for future genetic breeding programs.

Keywords China indigenous pigs, Selection signatures, Whole genome sequencing

[†]Yahui Gao and Xueyan Feng contributed equally to this work.

*Correspondence:

Zhe Zhang

zhezhang@scau.edu.cn

¹State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

²State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Introduction

Pigs serve as a consistent source of high-quality animal protein in human diets and play a vital role in human development and agricultural civilization. As one of the earliest domesticated animals in the shift from hunter-gatherer societies to agricultural civilizations, pigs probably underwent stages similar to other domesticated animals, including being hunted by humans, coexisting with humans, and eventually being domesticated [1–3]. Existing research has found that Eurasian wild boars started to diverge around one million years ago [4], with pig domestication beginning approximately 10,000 years ago [5]. It is now widely recognized that there are two independent centers of pig domestication globally, East Asia and the Near East [6], a conclusion supported by archaeological evidence [7–9]. China has a rich history of domestic pig breeding, with an abundance of indigenous pig resources. Over time, significant genetic variation has accumulated in traits and morphological characteristics, such as body size and skin color. This has led to the development of numerous indigenous pig breeds, each adapted to different environmental conditions, with varied body types and appearances, and distinct economic traits. Based on the current classification system, China indigenous pigs can be divided into six major types considering factors such as their origin distribution, body characteristics, and production performance: North China type, South China type, Jianghai type, Central China type, Southwest China type, and Plateau type [10]. North China indigenous pigs have medium-sized heads, black hair, large drooping ears, strong constitution, and adaptability to extensive farming, while the South type is smaller with sagging bellies and backs, thin skin, sparse hair, mostly black or black-and-white, and short broad bodies [10]. These two types differ in growth rate, feed efficiency, and lean meat rate. For instance, the weight and height of adult boars and sows of the Min pig breed, a North China type indigenous pig, were 227.10 ± 8.7 kg and 181.40 ± 10.27 kg, 89.10 ± 0.71 cm and 84.00 ± 0.42 cm, respectively [10]. For the breed of Luchuan pig, a South China type indigenous pig, the weight and height of adult boars and sows were 79.32 ± 2.94 kg and 78.52 ± 0.52 kg, 54.83 ± 0.81 cm, and 53.72 ± 0.18 cm, respectively [10]. Analyzing these differences aids in understanding the adaptive evolution history and breeding process, which provides a foundation and reference for genetic breeding work. At present, selection signature detection is a commonly used method that has been extensively applied in the genetic analysis of economic traits and the exploration of adaptive evolution in indigenous pigs and European commercial pig breeds.

In recent years, with the initiation and ongoing development of the Functional Annotation of Animal Genomes (FAANG) project [11], the FarmGTEx project

[12], and the PigGTEx project [13], researchers have gradually deepened their understanding and research from the genomic to the levels of gene expression, single-cell expression, and the regulation of functional elements within the genome. These projects have provided deeper biological insights into the tissue expression level and genetic regulatory mechanisms of important economic traits in pigs. By utilizing the advanced multi-omics findings from these livestock studies, the analysis of selection signature and the study of complex phenotypic regulation in livestock genomes are being advanced. In this study, we focused on the regions under selection and genetic variations affecting important economic traits in South China and North China indigenous pigs. Utilizing biological information from large-scale, multi-omics databases across different species, we performed signature mining analysis to thoroughly investigate the genetic basis under selection and the biological functions of key candidate genes in China indigenous pig populations.

Materials and methods

Sample collection and genotyping

We collected a total of 133 short-read whole-genome sequence (WGS) individual datasets from the PigGTEx project [13]. This dataset included 63 South China indigenous pigs of three breeds, 40 North China indigenous pigs of four breeds, and 30 Asian wild boars (Table 1). Among them, the Luchuan pig and Guangdongxiaoerhua pig were two subgroups of the breed Liangguangxiaoerhua pigs, and the Tunchang pig and Ding'an pig were two subgroups of the breed Hainan pigs.

The initial genotype file contained 42,523,218 single nucleotide polymorphisms (SNPs). We used PLINK v1.90 [14] to perform the following data quality control on the initial genotype: (1) retained SNPs with a minor allele frequency > 0.01 using the command “--maf 0.01”; (2) retained SNPs on autosomes using the command “--autosome”. Finally, 11,001,240 SNP loci were filtered out, leaving 31,521,978 SNPs for subsequent analysis after quality control.

Population genetic structure analysis

Principal component analysis

We applied PLINK v1.90 [14] to conduct principal component analysis (PCA) on the study population and calculate the eigenvalues and eigenvectors of the first ten principal components using the command “--pca 10”. The scatter plot of the first two principal components was generated using the R package ggplot2 v3.3.6 [15].

Phylogenetic tree construction

To investigate the phylogenetic relationships among the study populations, we constructed a phylogenetic tree based on genetic distances. Using PLINK v1.90 [14], we

Table 1 Sample information on South China indigenous pigs, North China indigenous pigs and Asian wild boars

Population	Breed		Sample size	Location/origin
North China indigenous pig breeds	Min		16	Suihua City, Heilongjiang Province
	Laiwu		6	Laiwu City, Shandong Province
	Hetaodaer		7	Bayannur City, Inner Mongolia Autonomous Region
	Bamei		11	Yulin City, Shaanxi Province
South China indigenous pig breeds	Liangguangxiaoerhua	Luchuan	17	Luchuan County, Yulin City, Guangxi Zhuang Autonomous Region
		Guangdongxiaoerhua	10	Suixi County, Zhanjiang City, Guangdong Province
	Hainan	Tunchan	10	Tunchang County, Hainan Province
		Dingan	10	Ding'an County, Hainan Province
	Bamaxiang		16	Bama County, Guangxi Zhuang Autonomous Region
Asian wild boars			30	/
Sum.			133	

calculated the identity by state (IBS) distances between individuals from 31,521,978 SNPs with the command “--genome”, representing genetic distances as 1-IBS. We used MEGA v7.0.14 [16] for genetic distance file format conversion, and constructed the phylogenetic tree using the Neighbor-joining (NJ) method. The tree was then visualized with the iTOL v6.7.2 tool [17].

Linkage disequilibrium decay analysis

We utilized PopLDdecay v3.40 [18] to perform LD decay analysis on three populations. Firstly, we extracted single-chromosome data for each population using BCFtools v1.12 [19] to generate genotype input files in *.vcf.gz format. We divided SNPs within 1 Mb into intervals as follows: “10 bp intervals for distances within 500 bp and 100 bp intervals for distances over 500 bp”, and calculated the average LD coefficient for all SNPs within these intervals. We then calculated LD per chromosome for a 1 Mb range with parameters “-MaxDist 1000 -bin1 10 -bin2 100 -break 500”. Finally, we merged the single-chromosome results for genome-wide LD calculations and visualized the results by plotting the LD decay curves for multiple populations.

Population genetic structure analysis

We performed population genetic structure analysis using admixture v1.3.0 [20]. We predefined ancestors (K) from 2 to 8 and conducted cross-validation analysis to compare the reliability of each K.

Genome-wide detection of selection signatures
Cross-population extended haplotype homozygosity (XP-EHH)

We carried out genome-wide selection signature detection on pairwise groups of South China indigenous pigs, North China indigenous pigs, and Asian wild boars. Using PLINK v1.90 [14], we extracted single-chromosome genotype files and conducted XP-EHH analysis with Selscan v1.3.0 [21], merging the results for all

chromosomes. Genetic distance files for genome-wide sites were obtained by converting physical positions, with 1 cM assumed to be equal to 1 Mb [22]. We applied two-tailed tests to the XP-EHH results, with SNPs in the top 1% considered significant. SNPs with scores below the 0.5th percentile indicated selection in population A, while scores above the 99.5th percentile indicated selection in population B. In the groupings of South/North China indigenous pigs with Asian wild boars, the wild boars were treated as population A, and the indigenous pigs as population B. In the grouping of South China vs. North China indigenous pigs, the North China indigenous pigs were designated as population A, and the South China indigenous pigs as population B.

Pairwise fixation index (F_{ST})

We employed VCFtools v0.1.13 [23] to calculate the per-site F_{ST} statistics for three groups. Subsequently, we ordered the F_{ST} statistics for all loci in each group from highest to lowest, and SNP loci exceeding the top 0.1% quantile were considered significant loci detected by this method.

Genome-wide association study with eigenvector decomposition (EigenGWAS)

We utilized GEAR v0.919 [24] with default parameters to perform EigenGWAS analysis on South China and North China indigenous pigs, aiming to screen for selection signals of population differentiation. First, we extracted the first principal component information from three groups: “South China indigenous pigs vs. Asian wild boars”, “North China indigenous pigs vs. Asian wild boars”, and “South vs. North China indigenous pigs” as the input phenotype. We then conducted calculations for each chromosome file and merged the results from individual chromosomes to achieve genome-wide site analysis results. To enhance the statistical power of this selection signature analysis, we utilized GC (Genetic Correction)-adjusted P values (P_{GC}) as the indicator for significant

loci. We applied the Bonferroni correction method, using “0.05 / total number of loci” as the significance threshold.

Biological annotation of significantly selected SNPs

Definition of significantly selected SNPs and intervals

We defined significantly selected SNPs as those detected as significant by the XP-EHH method and simultaneously detected as significant by either the EigenGWAS or F_{ST} methods, i.e., $(\text{Sig_XP-EHH} \cap \text{Sig_}F_{ST}) \cup (\text{Sig_XP-EHH} \cap \text{Sig_EigenGWAS})$. Additionally, we defined a potential selective region as the interval extending 50 Kb upstream and downstream from the significantly selected SNPs. Genes and associated QTLs located in candidate regions based on their chromosomal physical positions (*Sus scrofa* 11.1, Release 100) were regarded as candidate genes for those regions.

QTL region enrichment analysis

To determine if significant SNPs were significantly enriched in specific trait types, we conducted QTL region enrichment analysis on significant loci from each group using gff files downloaded from Animal QTLdb (Release 45) [25]. The steps were as follows: (1) To ensure the reliability of the results, we removed QTLs from the gff files that were insignificant, lacked a clear physical location, or had QTL intervals > 1 Mb; (2) Trait classification was based on the Trait Type from Animal QTLdb (Release 45) [25], and we excluded trait categories with fewer than 100 reports; (3) We used custom R scripts to extract and create input files: bed files with the physical location information of significant loci; (4) We performed permutation tests using the R package regioneR v1.26.1 [26], with 10,000 permutations for each set of significant SNPs. The significance threshold for permutation tests was set at $P \text{ value} < 0.05$.

Candidate gene pathway enrichment analysis

For a deeper understanding of the biological functions of candidate genes annotated by selection signatures in each group, we utilized the R package clusterProfiler v4.6.0 [27] and DAVID tool [28] for enrichment analysis, specifying the gene background species as pig (*Sus scrofa*). The significance criterion for enriched pathways was a $P \text{ value} < 0.05$. Specifically, we focused on the analysis of Biological Process terms from the GO analysis and the KEGG pathway categories.

Chromatin state enrichment analysis

To examine the roles of significant loci in different tissues and functional genomic layers, we performed chromatin state enrichment analysis on significant selected loci. The chromatin state information was obtained from the FANNG project [11]. These 15 chromatin states for 14 pig tissues (adipose, cecum, cerebellum, colon, cortex,

duodenum, hypothalamus, ileum, jejunum, liver, lung, muscle, spleen, and stomach) were divided into six categories: promoter-associated states (TssA, TssAHet, TssBiv), states related to proximal transcription regions near Transcription Start Sites (TSS) (TxFlnk, TxFlnkWk, TxFlnkHet), enhancer-associated states (EnhA, EnhAMe, EnhAWk, EnhAHet, EnhPois), ATAC island regions (ATAC_Is), repressive states (Repr, ReprWk), and quiescent states (Quiescent). For the chromatin state enrichment analysis, we excluded the Quiescent state due to its inactivity. We performed enrichment analysis using the R package LOLA v1.22.0 [29], and the significance threshold was set to $P_{\text{FDR}} < 0.05$ and enrichment fold > 1.

Enrichment analysis of complex traits in pigs

To further investigate the relationship between significant selected loci in South China and North China indigenous pigs and complex phenotypic traits in domestic pigs, we used 268 GWAS meta-analysis data from the PigGTEx project [13] to conduct enrichment analysis on the significantly selected SNPs for pig complex traits. We used the R package LOLA v1.22.0 [29] to conduct the analysis and perform Fisher's exact test, with the significance threshold set at $P_{\text{FDR}} < 0.05$ and enrichment fold > 1. Traits with odds ratio > 0 were plotted and converted to enrichment fold using the formula ‘log2(odds ratio + 1)’.

Multi-omics functional annotation analysis of candidate genes

We used large multi-omics databases such as HPA v22.0 [30], IMPC [31], GWASATLAS [32], and the PigGTEx project [13] to perform multi-level cross-species biological function annotation for key candidate genes, including transcriptomic expression levels in humans, pig, and mouse tissues, single-cell transcriptomic expression levels and protein expression levels in human, and associated phenotypes in human and mouse.

Results

Characteristics of the genome datasets

We collected 133 whole genome sequencing data of pigs from the pig genomics reference panel (PGRP) [13], including 30 Asian wild boars, 63 South and 40 North China indigenous pigs (Fig. 1a; Table 1). We aligned the clean reads to the *Sus scrofa* 11.1 reference genome [33] using BWA [34]. The sequence depth for each sample ranged from 5.24X to 69.40X, with an average of 19.94X (Table S1). Subsequently, we called SNPs on a population level and obtained 42,523,218 SNPs. After filtration of the raw variants, we kept 31,521,978 high-quality SNPs belonging to autosomes for subsequent analyses.

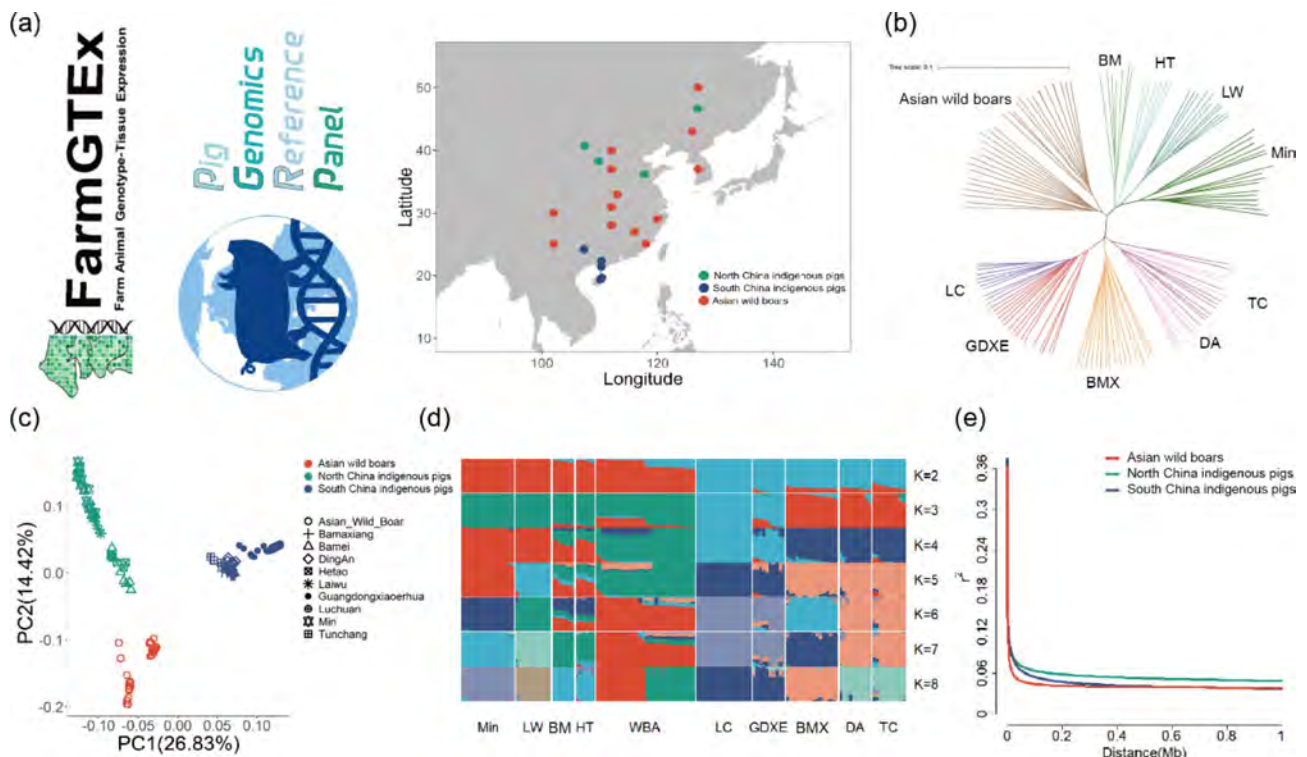


Fig. 1 Samples location and population genetic structures of South China and North China indigenous pigs and Asian wild boars. **a** Locations of the samples, which were collected from the PigGTEx project. **b** Neighbor-joining phylogenetic tree of 133 pigs. **c** Principal component analysis (PCA) result of 133 pigs on the first two PCs. **d** Genetic ancestry compositions with the assumed number of ancestries from $K=2$ to $K=8$. **e** Linkage disequilibrium decay in the distance of 1 Mb. LC, Luchuan pigs; GDXE, Guangdongxiaohua pigs; BMX, Bamaxiang pigs; DA, Ding'an pigs; TC, Tunchang pigs; LW, Laiwu pigs; HT, Hetao pigs; BM, Bamei pigs

Population structure analyses

The top two principal components (PCs) of the PCA contributed 26.83% and 14.42% genetic variance, respectively (Fig. 1c). PC1 and PC2 divided all the individuals into three groups defined by geographical distribution. Consistent with the relationships in the PCA, the Neighbor-joining tree showed that all individuals were clustered together according to their breeds (Fig. 1b). The assumed ancestral lineage compositions of all individuals were determined with a range of K values, where K represents the number of assumed ancestries (Fig. 1d). With an increasing K value, the populations within the three groups were gradually distinguished from each other. As depicted in Fig. 1e, the linkage disequilibrium (LD) decay among the three groups showed a similar trend, where r^2 decreased with increasing SNP marker distance, but the decay rates differed. Asian wild boars had lower LD levels and a faster decay rate than the other two populations. North China indigenous pigs exhibited a slower decay rate and higher LD levels compared to the South type. The maximum r^2 values were 0.3749 for South China indigenous pigs, 0.3737 for North type, and 0.3623 for Asian wild boars, with corresponding maximum LD distances of approximately 300 Kb, 200 Kb, and 50 Kb.

In addition, in preparation for using the EigenGWAS method in subsequent research, which requires population eigenvector files as input, we performed PCA on each pair of the three groups. As illustrated in Fig. S1, PC1 in each group successfully differentiated the corresponding two populations. Consequently, we extracted the PC1 values from these groups which will be used as input phenotype data for subsequent EigenGWAS.

Selection signatures in South China indigenous pig and Asian wild boar

Overview of selection signatures detection

In the XP-EHH test results for the grouping of South China indigenous pigs and Asian wild boars, the scores ranged from 1.2633 to 2.1607, with an average value of 0.1318. The 0.5% and 99.5% percentile XP-EHH values were -0.3793 and 1.1205 , respectively (Fig. S2a). Loci with XP-EHH values below the 0.5th percentile (XP-EHH score < -0.3793) were considered candidate loci under selection in Asian wild boars, while loci with XP-EHH values above the 99.5th percentile (XP-EHH score > 1.1205) were considered candidate loci under selection in South China indigenous pigs. There were 157,561 and 157,567 candidate loci detected in South China indigenous pigs and Asian wild boars, respectively,

with significant loci distributed across all chromosomes (Fig. 2a). The results of the F_{ST} test indicated that the F_{ST} statistics ranged from -0.0252 to 1 , with an average value of 0.0886 (Fig. S2a). In total, we detected 30,349 significant loci ($F_{ST} > 0.7314$) (Fig. 2b). From the EigenGWAS findings, we discovered 7,651 significant loci ($P < 1.6430 \times 10^{-9}$, Fig. S2), with peaks appearing on multiple chromosomes such as SSC1, SSC4, and SSC8 (Fig. 2c).

By using the XP-EHH method and confirming significance with either F_{ST} or EigenGWAS, we detected 5,227 loci as significantly selected (Fig. 2d). These loci were unevenly distributed across the autosomes, with the strongest signature peaks found on SSC1 and SSC3. The highest concentrations of significant loci were on SSC1, SSC15, and SSC2, with counts of 1,721, 745, and 533, respectively (Fig. 2d).

Biological annotation of significant selected loci

By extending 50 Kb both upstream and downstream from the significant selected loci to identify the selected regions, we conducted gene annotation for these regions. In South China indigenous pigs, we obtained 304 candidate genes, compared to 44 in Asian wild boars (Table S2). The pathway analysis indicated that the candidate genes in Asian wild boars were significantly enriched in a few pathways such as regulation of ventricular cardiac muscle cell membrane repolarization, cochlea development, integrin-mediated cell adhesion, prion diseases, and focal adhesion (Table S3). For South China indigenous pigs, candidate genes were significantly enriched in 21 GO biological processes and 17 KEGG pathways. Among these, the *DLX1* and *DLX2* genes were enriched in several brain nerve development pathways, including hippocampus development, fate commitment of GABAergic interneurons in the cerebral cortex, and subpallium development. Additionally, genes including *IKBKB*, *NFATC2*, *CD3E*, *CD3D*, *MAPK14*, and *NFATC2* were extensively enriched in pathways related to viral infection and immune response, such as Th1 and Th2 cell differentiation, C-type lectin receptor signaling pathway, and Chagas disease (Table S3).

QTL enrichment analysis for significant selected loci

Following the QTL enrichment analysis for significant selected loci in these two groups, it was observed that unlike Asian wild boars, which showed significant enrichment solely in blood parameters and meat texture traits ($P < 0.05$), South China indigenous pigs exhibited enrichment in five QTL trait categories: exterior, health, meat and carcass, production, and reproduction. Notably, in the health category, enrichment was noted in QTLs associated with disease resistance and immune capacity traits. In the reproduction category, the QTLs

related to reproductive traits, reproductive organs, and litter performance (Fig. 2f, Table S4).

Chromatin state enrichment analysis for significant selected loci

The enrichment revealed that differences between the two groups were concentrated in visceral organs, the digestive system, and cerebellum tissues. In Asian wild boars, selected loci were significantly enriched in enhancer regions of visceral organs (liver, spleen, lungs), with extreme enrichment in the weakly active enhancer chromatin state in the liver ($P_{FDR} < 0.001$, Table S5). For spleen, Asian wild boars showed significant enrichment in poised enhancer functional elements, while South China indigenous pigs were significantly enriched in weakly repressive Polycomb regions (ReprWk) ($P_{FDR} < 0.01$). Also, South China indigenous pigs' selected loci were enriched in the ATAC island state in cerebellar tissue and repressive states in the cecum and colon. Additionally, in the ileum, both were significantly enriched in weakly repressive Polycomb regions (ReprWk) and weak enhancer (EnhAWK) states, respectively (Fig. 2e).

Complex trait enrichment analysis for significant selected loci

In comparison to Asian wild boars, South China indigenous pigs showed significant enrichment in multiple complex traits within the growth, reproduction, and fat trait categories (Fig. 2g, Table S6). These included significant enrichment in growth traits such as days to reach 115 kg (DAYS_115) and average daily gain (ADG) ($P < 0.001$). In terms of reproductive traits, it mainly included the total number of piglets born (TNB) and total litter weight at weaning (TLWT_Weaning) ($P < 0.001$). There was also significant enrichment in backfat thickness (BFT) ($P < 0.001$).

Selection signatures in North China indigenous pig and Asian wild boar

Overview of selection signatures detection

The test results between this comparison group showed XP-EHH scores ranging from -1.0273 to 1.8237 , with an average of 0.0425 . The XP-EHH values at the 0.5% and 99.5% quantiles were -0.4107 and 0.7443 , respectively (Fig. 2). Loci with XP-EHH values less than the 0.5% quantile (XP-EHH score < -0.4107) were candidate selection loci in Asian wild boars, whereas those greater than the 99.5% quantile (XP-EHH score > 0.7443) were candidate in North China indigenous pigs. We found 157,530 and 157,528 significant loci in North China indigenous pigs and Asian wild boars. These loci were unevenly distributed across the chromosomes, with distinct selection signatures on SSC1, SSC5, SSC6, and SSC8 in the North China indigenous pigs (Fig. 3a). According to the F_{ST} test, the statistics ranged from -0.0300 to 0.9709 , with an

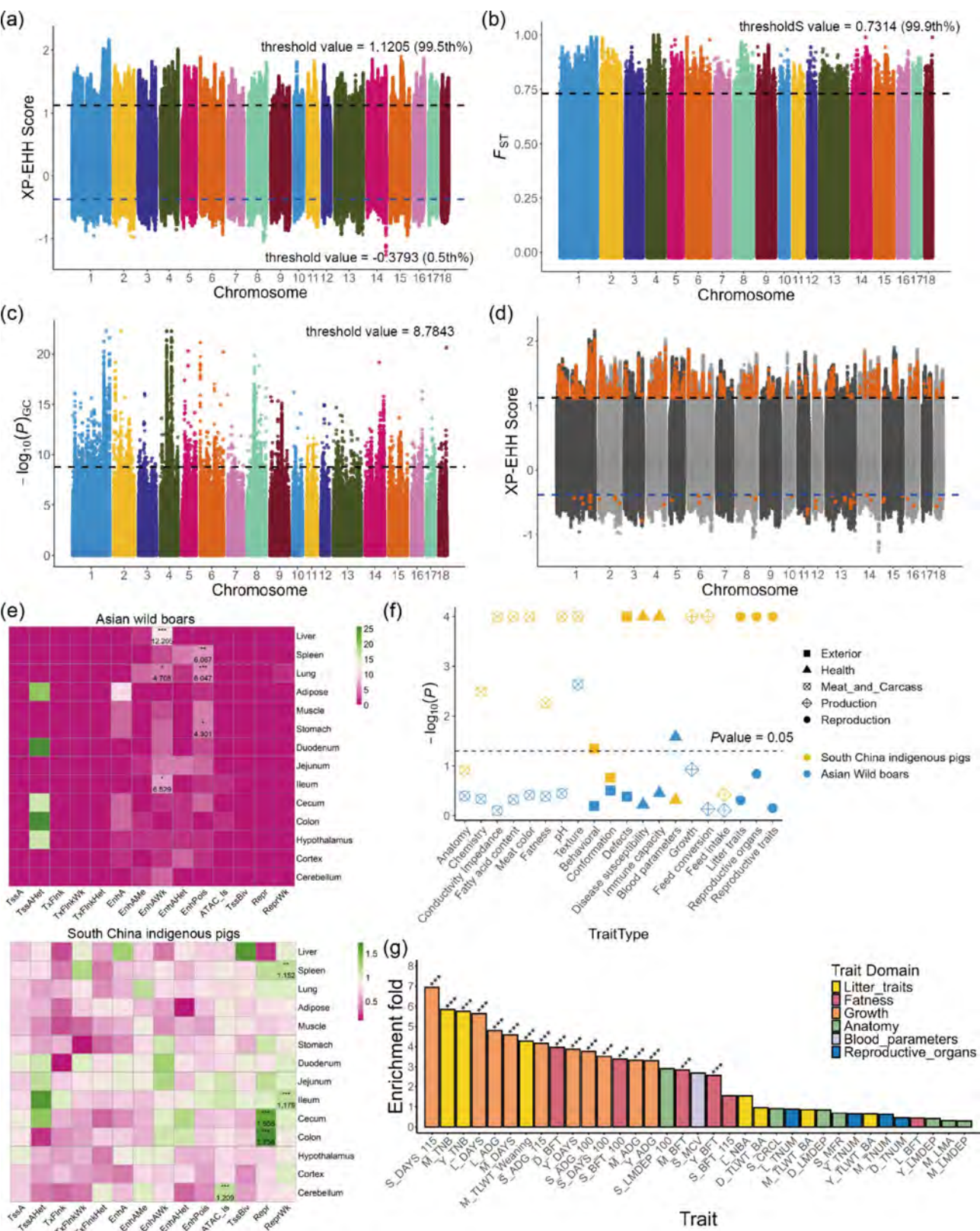


Fig. 2 Selection signatures and biological annotation of the paired South China indigenous pigs and Asian wild boars. **a–c** Manhattan plots of the selection signatures detected by XP-EHH, F_{ST} , and EigenGWAS. **d** Significant selected loci distribution in the whole genome. **e** Chromatin state enrichment analysis for significant selected loci in 14 major tissues of pigs. **f** QTL region enrichment analysis for the selected region. **g** Enrichment analysis of complex traits based on the selected SNP windows in pigs. The whole trait name showed in table S6

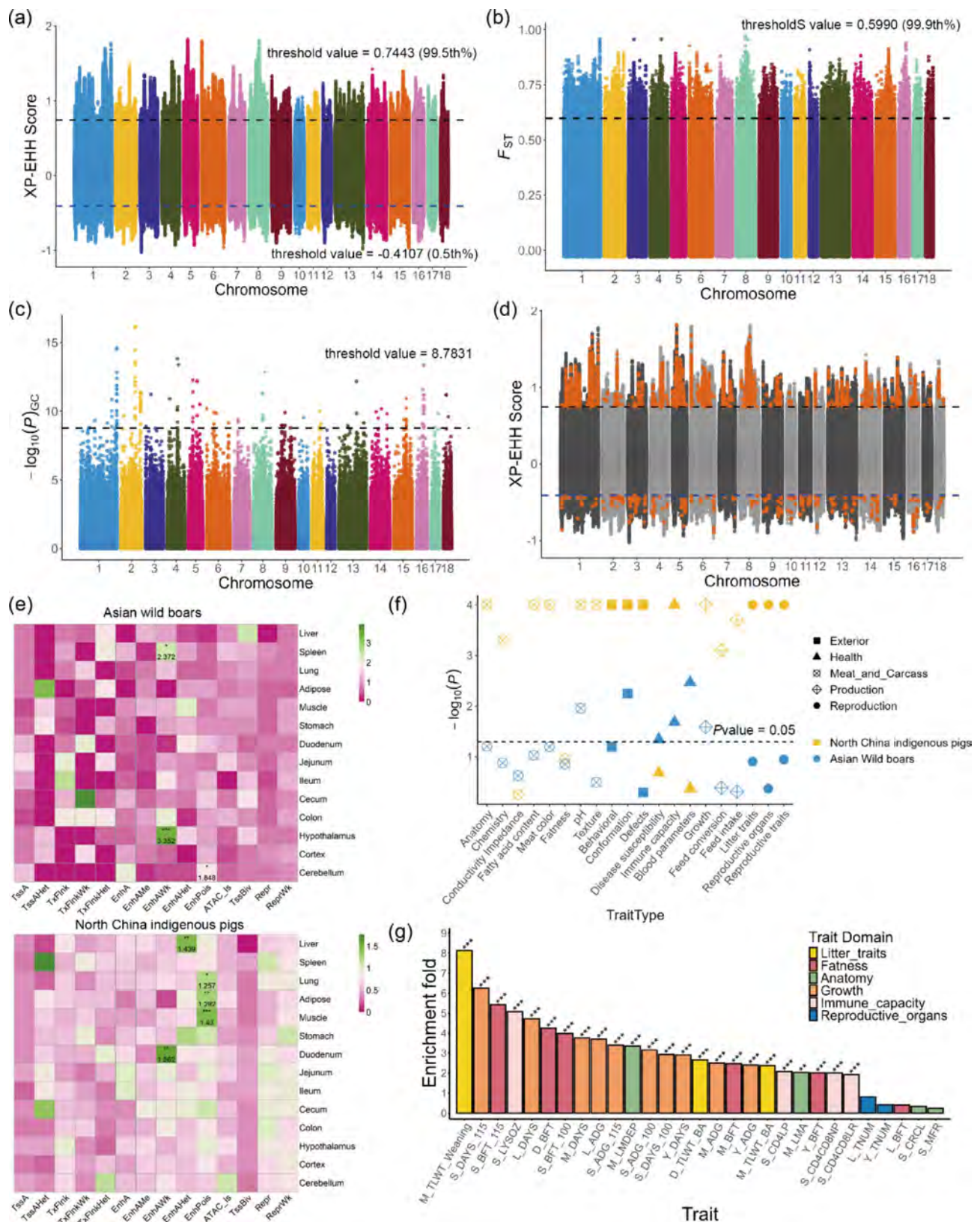


Fig. 3 Selection signatures and biological annotation of the paired North China indigenous pigs and Asian wild boars. **a–c** Manhattan plots of the selection signatures detected by XP-EHH, F_{ST} and EigenGWAS. **d** Significant selected loci distribution in the whole genome. **e** Chromatin state enrichment analysis for significant selected loci in 14 major tissues of pigs. **f** QTL region enrichment analysis for the selected region. **g** Enrichment analysis of complex traits based on the selected SNP windows in pigs. The whole trait name showed in table S6

average of 0.0670. The F_{ST} value at the 0.1% highest quantile was 0.5990 and we detected 30,346 significant loci using this method ($F_{ST} > 0.5990$) (Fig. 3b). The EigenG-WAS results showed that $-\log_{10}P_{GC}$ values ranged from 0 to 16.1805, with a mean value of 0.4223, and 191 significant loci were detected ($P_{GC} = 1.6476 \times 10^{-9}$). The most significant signature was found on SSC2 with peaks also present on SSC1, SSC2, and SSC16 (Fig. 3c).

Finally, we found 5,800 significant selection loci in the North China indigenous pig. These loci were distributed across all 18 chromosomes, showing an uneven distribution (Fig. 3d). The locus with the highest XP-EHH score was located on SSC5. SSC1 and SSC8 had the highest numbers of significant selection loci, with 1,305 and 1,526 loci, respectively. In the Asian wild boar group, 489 significant selection loci were detected.

Biological annotation of significant selected loci

In the regions showing selection signatures, we identified 363 and 157 candidate genes in North China indigenous pigs and Asian wild boars, respectively (Table S7). Pathway analysis of candidate genes indicated that those from the Asian wild boar group were significantly enriched in seven GO biological process terms and five KEGG pathways (Table S8). In North China indigenous pigs, candidate genes were significantly enriched in 26 GO biological process terms and 34 KEGG pathways, mainly related to immune response, inflammatory response, and viral infection. Among these, *RNF114*, *GAL3ST1*, *LIMK2*, *KIT*, *MAEL*, *SPATA2*, *AP3B1*, *CABS1*, *PATZ1*, and *ACVR2A* genes were significantly enriched in the spermatogenesis pathway, *GNAQ*, *KCNMA1*, *ITPR2*, *ITPR3*, *LYZ*, and *ADRA1A* genes were enriched in salivary secretion, and *MC1R*, *KIT*, and *SNAI2* genes were significantly enriched in the pigmentation pathway (Table S8).

QTL enrichment analysis for significant selected loci

The selection signatures in the North China indigenous pig population were significantly enriched in QTLs related to production trait class (growth, feed conversion ratio, and feed intake), reproduction traits class (reproductive traits, reproductive organs), and meat and carcass class (fatty acid content and meat color) (Fig. 3f). On the other hand, the significant signatures in Asian wild boars were predominantly enriched in QTLs associated with immunity and health.

Chromatin state enrichment analysis for significant selected loci

As depicted in Fig. 3e, the significant selection signatures in Asian wild boars were significantly enriched in the enhancer states of tissues such as the hypothalamus, cerebellum, and spleen. Meanwhile, the significant selection signatures in North China indigenous pigs were

prominently enriched in the enhancer states of visceral tissues (liver and lungs), muscle, fat, and duodenum (Fig. 3e, Table S5).

Complex trait enrichment analysis for significant selected loci

Selection signatures in North China indigenous pigs were enriched in various complex traits related to growth, reproduction, immunity, and fat characteristics (Fig. 3g, Table S6). Specifically, for growth traits, there was significant enrichment in traits like DAYS_115, ADG, and BFT ($P < 0.001$). For immune traits, significant enrichment was observed in lysozyme levels and the percentage of CD4-positive leukocytes ($P < 0.001$). TLWT_Weaning showed the highest level of enrichment. In contrast, the selection signatures in Asian wild boars were significantly enriched in the phenotype of the number of teats (TNUM) ($P < 0.001$, Table S6).

Selection signatures in South China and North China indigenous pig

Overview of selection signatures detection

In this comparison group, XP-EHH scores ranged from -1.1870 to 2.0498 , with an average of 0.0778 . The XP-EHH values at the 0.5% and 99.5% quantiles were -0.4868 and 0.8738 , respectively. Loci with XP-EHH values below the 0.5% quantile (XP-EHH score < -0.4868) were candidate selection loci in North China indigenous pigs, while those above the 99.5% quantile (XP-EHH score > 0.8738) were candidate in South types. We identified 157,569 and 157,567 significant loci in South China and North China indigenous pigs, respectively. These loci were distributed across every chromosome, with significant ones on SSC1, SSC4, and SSC12 in South China indigenous pigs, and on SSC8 and SSC11 in North types (Fig. 4a). The F_{ST} statistics ranged from -0.0209 to 0.9701 , with an average value of 0.1017 . The F_{ST} statistic at the 0.1% highest quantile was 0.7265 . This method identified 30,612 significant loci with F_{ST} values exceeding 0.7265 (Fig. 4b). The EigenG-WAS results showed $-\log_{10}P_{GC}$ values ranging from 0 to 23.9567 , with a mean value of 0.4456 . There were 2,165 significant loci with P_{GC} values greater than the threshold ($P_{GC} = 1.6331 \times 10^{-9}$), with significant peaks appearing on SSC1, SSC4, SSC14, and SSC16 (Fig. 4c).

We detected 3,527 significant selection loci in South China indigenous pigs, distributed unevenly across the genome. Significant selection signatures were particularly evident on SSC1, SSC3, and SSC12, with SSC1 having the most significant one, totaling 1,285 (Fig. 4d). In North types, 958 significant selection loci were detected, with distinct selection signals on SSC1, SSC2, and SSC6. The highest numbers of significant loci were on SSC6 and SSC2, with 173 and 109 loci, respectively. The loci with the highest XP-EHH values were located on SSC8 (Fig. 4d).

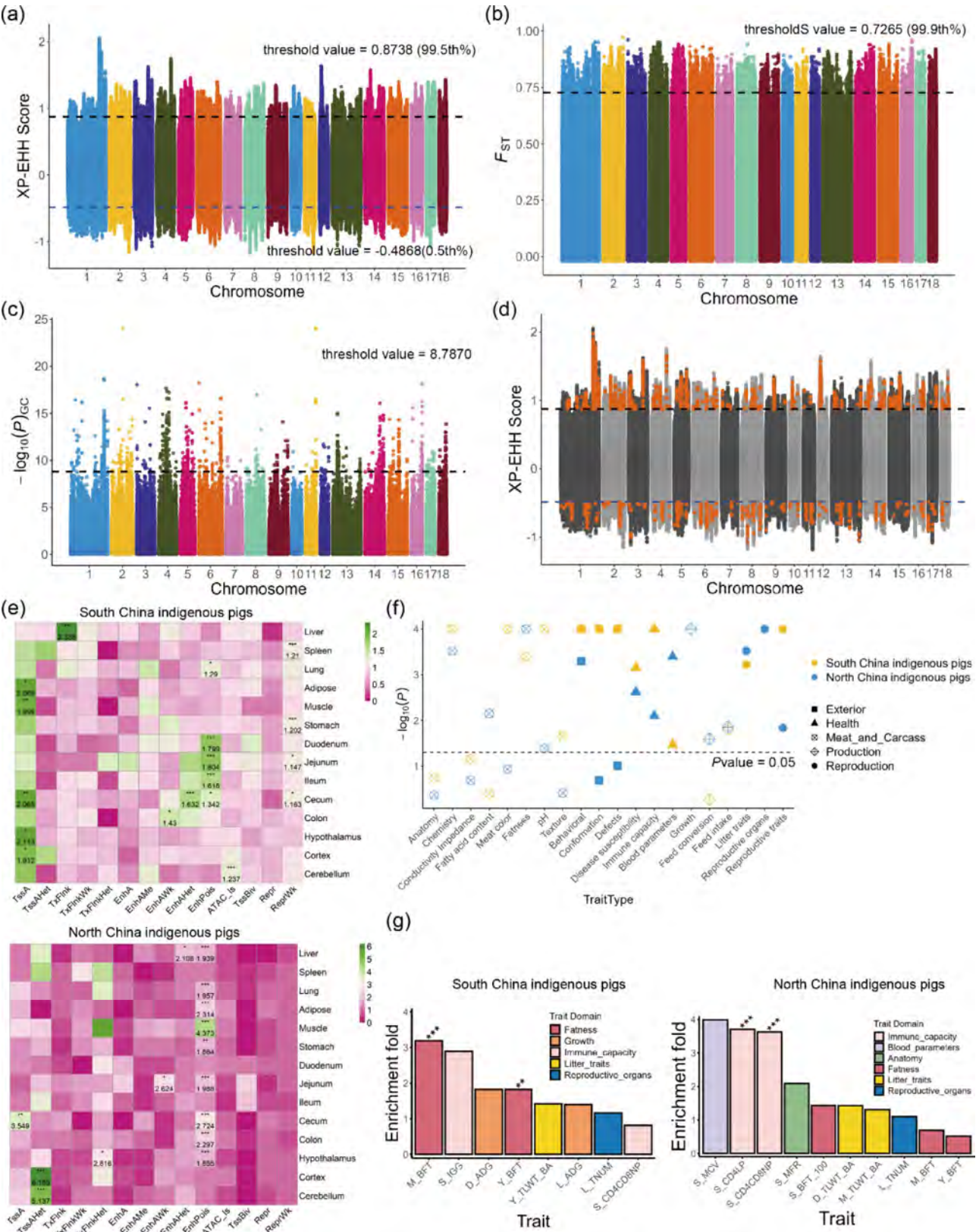


Fig. 4 Selection signatures and biological annotation of the paired South China indigenous pigs and North China indigenous pigs. **a-c** Manhattan plots of the selection signatures detected by XP-EHH, F_{ST} , and EigenGWAS. **d** Significant selected loci distribution in the whole genome. **e** Chromatin state enrichment analysis for significant selected loci in 14 major tissues of pigs. **f** QTL region enrichment analysis for the selected region. **g** Enrichment analysis of complex traits based on the selected SNP windows in pigs. The whole trait name showed in table S6

Biological annotation of significant selected loci

We annotated a total of 243 genes in South China indigenous pigs (Table S9), which were significantly enriched in 15 biological process pathways and one KEGG pathway, predominantly involved in complex developmental differentiation processes (Table S10). In North China indigenous pigs, 175 genes were annotated. These genes were enriched in 21 significant biological process pathways and 25 significant KEGG pathways, primarily linked to metabolic regulation (Table S10).

QTL enrichment analysis for significant selected loci

The enrichment analysis revealed that South China indigenous pigs' significant loci were enriched in meat quality, carcass, and appearance categories (Fig. 4f). In comparison, North China indigenous pigs' significant loci were more enriched in production and reproductive traits. For production traits, both populations showed enrichment in QTLs related to feed intake. Moreover, North China indigenous pigs' selection loci were enriched in QTLs related to growth traits and feed conversion efficiency (Fig. 4f).

Chromatin state enrichment analysis for significant selected loci

The results showed selection signatures enriched in the enhancer regions of the digestive system (stomach, small and large intestines) (Fig. 4e, Table S5). In South China indigenous pigs, significant selection signals were enriched in the EnhPois state of the small intestine (duodenum, jejunum, ileum) and cecum, and also enriched in the repressive functional regions of the stomach, jejunum, and spleen. The regulatory states in cecum tissue were diverse, with enrichment in TssA, EnhAHet, and ReprWk states. Additionally, there were differences between South China and North China indigenous pigs in the chromatin region enrichment of brain, muscle fat, and liver. Significant loci in South China indigenous pigs were enriched in the TssA of the cerebral cortex and hypothalamus and in the ATAC_Is of the cerebellum. In North China indigenous pigs, significant enrichment was found in the EnhPois and TxFlnkHet of the hypothalamus and in the TssAHet of the cerebellum and cerebral cortex. Both populations showed specific enrichment in the TssA and EnhPois states of adipose and muscle tissues, respectively. In the liver, the selected loci of South China indigenous pigs were enriched in TxFlnk, while those of North types were enriched in EnhAHet and EnhPois (Table S5). These differences in chromatin state enrichment in various tissues may be related to the different selection objectives and mechanisms for fat deposition, growth, and immune performance between these two populations.

Complex trait enrichment analysis for significant selected loci

The enrichment analysis for complex traits in pigs indicated that the most significant differences in selection direction between South China and North China indigenous pigs were observed in fat traits and immune traits (Fig. 4g). The candidate loci in South China indigenous pigs were enriched in the BFT trait, while those in North types were enriched in the CD4LP and CD4CD8NP traits (Fig. 4g).

Functional annotation analysis of genes under selection in indigenous pigs

For a more detailed analysis of the biological functions of selection signature in South China and North China indigenous pigs, all genes identified in the same type of pig population across three groups, South vs. North China indigenous pigs, South/North China indigenous pigs vs. Asian wild boars, were considered potential selected genes for that population. Repeatedly detected genes were classified as key selected genes.

Pathway analysis of selected genes in South China indigenous pigs

There were 439 potentially selected genes in the South China indigenous pig, of which 108 were key selected genes (Fig. 5a). These pathways, in which these genes were significantly enriched, formed clusters of biological function pathway networks, which are mainly associated with the activation, proliferation, and differentiation of immune cells (T cells, lymphocytes, leukocytes) as well as with the regulation of neural cell development and cell differentiation (Fig. 5b, Table S11). The enriched pathways for key selected genes were involved in functions such as cell development and metabolism, immune response, and others. *ABCA1*, in particular, was enriched in pathways that stabilize vascular endothelial cells and were associated with anti-atherosclerosis (Table S12).

Pathway analysis of selected genes in North China indigenous pigs

There were 503 potentially selected genes in the North China indigenous pig, with 35 identified as key selected genes (Fig. 5a). Clustering of functional pathway networks in which genes potentially under selection in North China indigenous pigs were significantly enriched showed that they were mainly associated with biological pathways related to blood circulation process, regulation of organism growth and development, and immune response regulation (Fig. 5c, Table S13), as well as reproductive physiological developmental pathways such as maternal processes during pregnancy, follicular development, and spermatogenesis (Table S11). The pathway enrichment result for key selected genes was shown in Table S11, with five genes (*LYN*, *KCNMA1*, *ITPR2*,

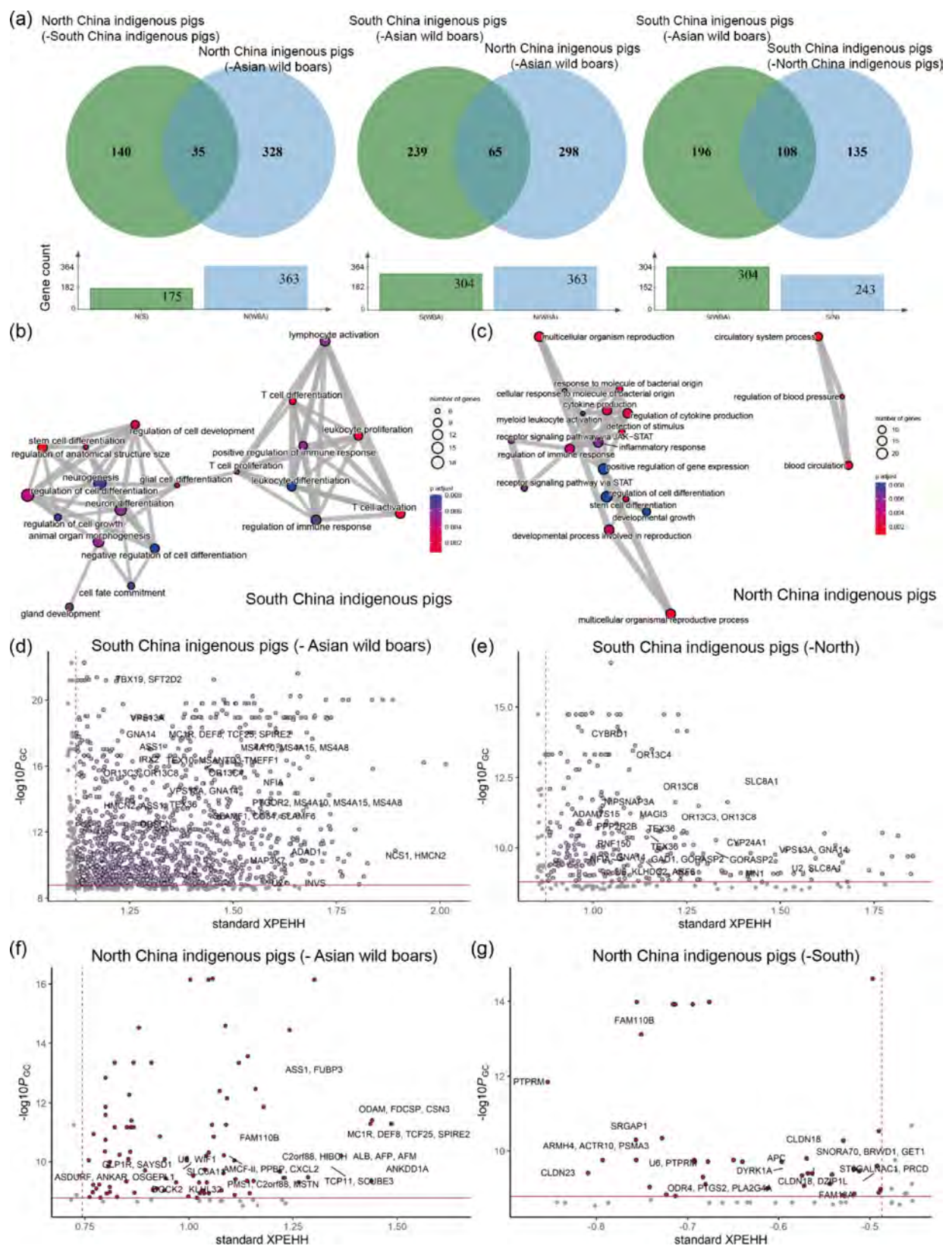


Fig. 5 The selected genes and their biological functions of South China and North China indigenous pigs. **a** Venn diagrams of the genes detected in different tests. **b-c** Network of the biological function terms where the selected genes in South China and North China indigenous pigs were enriched. **d-g** Key selected genes distribution in different tests

PLA2G4A, *PTGS2*) significantly enriched in six KEGG pathways, associated with neural cell signaling, immune response, and cardiovascular functions (Table S12).

Pathway analysis of shared selected genes in China indigenous pigs

In the selection signature detection analysis with Asian wild boars as the control group and indigenous pigs as the observation group, we identified 65 shared selected genes in both South China and North China indigenous pigs (Fig. 5a). These shared candidate genes were enriched in six GO biological processes and two KEGG pathways (Table S12), with several interferon (IFN) and interleukin (IL) gene family members enriched in pathways related to immune regulation processes and cytokine signaling.

Screening of differentially selected genes in China indigenous pigs

To delve deeper into the differential selection signals between South China and North China indigenous pigs, we kept the differential loci detected by XP-EHH, F_{ST} , and EigenGWAS, and defined candidate genes annotated with more than 10 of these loci as strongly selected genes. We found 33 and 15 strongly selected candidate genes in South China indigenous pigs (Fig. 5d-e, Fig. S3) and North China indigenous pigs (Fig. 5f-g, Fig. S3), respectively, of which 11 genes (*ASS1*, *FUBP3*, *MC1R*, *DEF8*, *TCF25*, *ODAM*, *C2orf88*, *FDCSP*, *CSN3*, *DOCK2*, *SPIRE2*) were commonly annotated as strongly selected (Table S14, Fig. 5d-g). These distinct genes were connected to different physiological functions, such as reproductive physiology, brain neurodevelopment, coat color, and immunity.

The transcriptome expression profiles from the HPA database indicated that genes such as *VPS13A* and *TEX36* were associated with male infertility in humans, with both showing specific high expression levels in the testicular tissue of pigs and humans. *EDRF1* was also highly expressed in the testicular tissue of pigs and humans, with single-cell transcriptome data showing significant enrichment in early and late-stage sperm cell clusters, suggesting a role in spermatogenesis. *ESR1* demonstrated high tissue-specific expression in reproductive organs such as the cervix and fallopian tubes (Fig. S4a-d). *CSN3* was involved in mammalian lactation, being expressed only in the salivary glands and mammary tissue, with particularly high expression in mammary tissue (Fig. S5a-c). Similarly, the pig transcriptome atlas showed that the *CSN3* gene had specific high expression in the lactating tissue of pigs (Fig. S5b).

NFIA, *BRWD1*, and *ST18* were associated with brain neurodevelopment. For example, *NFIA* was linked to brain development, brain malformations, and lethality in

mice. *BRWD1* was specifically enriched in human brain tissue and highly expressed in porcine fetal development tissues such as oocytes and blastomeres. *ST18* showed transcriptome-level expression only in human brain tissue but was highly expressed in various pig brain tissues.

The *MC1R* gene, associated with coat color, had an average expression level of TPM > 1 in pig brain tissues (frontal cortex, cerebrum, hypothalamus) (Fig. S5d-h). In humans, it showed tissue-specific high expression in the pituitary gland and testicular tissues (Fig. S5e). Additionally, in phenotype association tests, *MC1R* was linked to abnormal hair and hair pigmentation phenotypes in mice (Fig. S5f) and various skin and hair color phenotypes in humans (Fig. S5g).

Additionally, we identified several genes associated with mammalian immune function, such as *FDCSP*, *PIK3API*, and *DOCK2*. *FDCSP* and *ODAM* exhibited tissue-specific expression in lymphoid tissues and salivary glands and were linked to metabolic traits in humans (Fig. S4e-h). *PIK3API* showed high expression in lymphoid tissues, liver, and salivary glands in both human and pig transcriptome profiles and was associated with immune and metabolic complex traits in humans, and with increased neutrophil and monocyte counts and decreased lymphocyte counts in mice. *DOCK2* was expressed in various pig tissues, with higher expression levels in fetal thymus, lymph nodes, spleen, macrophages, and blood (Fig. 6). In humans, *DOCK2* had specific high expression in bone marrow, lungs, and lymphoid tissues (Fig. 6b), was specifically enriched in immune response clusters in lymphoid tissues (Fig. 6d), and was associated with decreased bone density and increased spleen weight phenotypes in mice (Fig. 6e).

Image credit: a, PigGTEx-Portal [13], <http://piggtex.far.mgtxe.org/>. b-d, Human Protein Atlas [30], www.protein-atlas.org. e, International Mouse Phenotyping Consortium [31], www.mousephenotype.org.

Discussion

Based on the WGS data, we employed three methods, i.e., F_{ST} , XP-EHH, and EigenGWAS, to detect genome-wide selection signatures in South China and North China indigenous pig breeds. We annotated the significant selected loci and functional genes using enrichment of chromatin state, QTL region, complex trait, and pathway enrichment. The results indicated that China indigenous pigs have been positively selected for traits related to feeding habits and feed conversion efficiency. Moreover, different types of indigenous pigs had distinct breeding directions: South China indigenous pigs showed selection signatures concentrated in traits related to fat deposition, meat quality, body shape, and immune function, while North China pigs exhibited signals related to

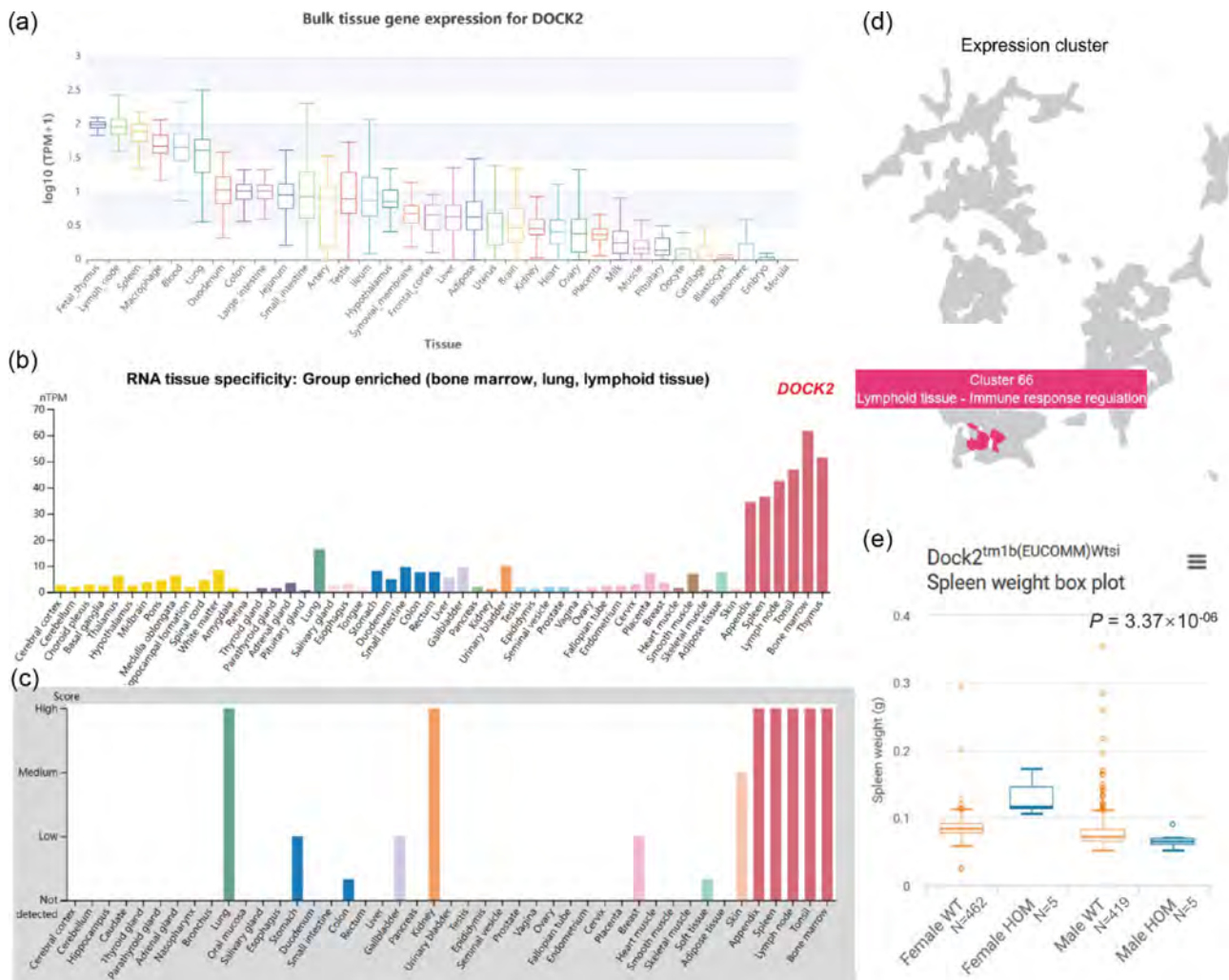


Fig. 6 Gene expression and biological phenotypes regulated by immune-associated gene *DOCK2* in mammals. **a** RNA expression overview across tissues on pigs. **b-c** RNA expression and protein level overview across tissues in humans. nTPM, normalized expression levels. Color coding is based on tissue groups, each consisting of tissues with functional features in common. **d** *DOCK2* is expressed specifically in the cluster Lymphoid tissue - Immue response regulation. **e** Related phenotype: increased spleen weight on *Dock2*^{-/-} female mice, compared with WT female mice

growth and development, blood physiology, and reproductive performance.

Firstly, by integrating the results from PCA, phylogenetic trees, and ancestry detection, we observed that North China indigenous pigs were genetically closer to Asian wild boars. This was hypothesized to be related to the geographical proximity of the Asian wild boar populations and the occurrence of gene flow events between them.

Secondly, through signature detection analysis of pairwise combinations among three groups, we identified candidate genes shared among China indigenous pigs, primarily associated with coat color, such as *MC1R*, *EDNRB*, and *KIT*. The *MC1R* gene was annotated in the SSC6: 0.120-0.28 Mb region in both South China and North China indigenous pigs. This gene was shown in humans to have multiple allelic mutations and was widely

reported to be associated with pigmentation, skin color [35], and susceptibility to melanoma [36]. Additionally, polymorphic mutations in this gene were related to changes in melanin synthesis in cattle coat color [37], horse coat color [38], chicken plumage color [39], and mice [40]. In pigs, Kijas et al. [41] were the first to reveal the role of *MC1R* variations in pig coat color diversity. These findings highlighted that *MC1R* played a central role in regulating the synthesis of eumelanin (black/brown) and pheomelanin (red/yellow) in mammalian melanocytes. Recent researchers used genetic engineering to create *MC1R* gene-edited pigs [42], manipulating pig coat color to cater to future consumer demands in the meat market, and evaluating the breeding of new pig breeds. We annotated the *KIT* gene in the SSC8: 41.46–41.56 Mb region in North China indigenous pigs. This gene was reported to have a clear association with the

white coat color phenotype in indigenous pigs [43] and was sensitive to melanocytes involved in the pigmentation of the epidermis and hair follicles [44]. Additionally, this gene was found to influence *MC1R* expression and was associated with the increased white spotting phenotype in horses [45]. Furthermore, we annotated the *EDNRB* gene in the SSC11: 50.04-50.14 Mb region in South China indigenous pig populations. Ai et al. [46] revealed through analysis of the “two-end-black” coat color in indigenous pig populations that *EDNRB* might be associated with the appearance of white coat color in indigenous pigs. This gene was also enriched in pigmentation and melanocyte differentiation pathways in Pudong White pigs [47]. In other mammals, the deletion of *EDNRB* in mice [48] and horses [49] results in a white banded coat color phenotype similar to the “two-end-black” color in China indigenous pigs. Additionally, an SNP site in this gene was reported by Yan et al. [50] as the causative mutation for albinism in canines, specifically in the Chinese raccoon dog.

Thirdly, the results of the study on the performance selection directions of South China and North China indigenous pigs demonstrated that North China indigenous pigs exhibited better growth characteristics, while South China indigenous pigs have a higher capacity for fat deposition. We annotated the *IGF1R* and *IGF2R* genes in South China and North China indigenous pigs, respectively. Zhan et al. [51] compared the expression patterns of *IGF1R* and *IGF2R* in different tissues and myocytes in Nanjiang Yellow goats, confirming that these two genes synergistically promoted muscle tissue development and myocyte proliferation and differentiation. The *IGF1R* gene was first discovered and confirmed to play a key role in the regulation of neuroendocrine functions and growth and development in animals [52, 53]. It was involved in regulating cell proliferation, migration, and organ formation during the developmental process of animals and played an important role in the insulin signaling pathway that regulates animal body size. Previous studies on selection gene loci in large-sized pigs and small-sized China indigenous pig breeds identified a mis-sense mutation in the *IGF1R* gene that occurred at a low frequency only in large-sized pigs, suggesting that *IGF1R* may be a key candidate gene for regulating body size and organ development in domestic pigs [54]. Similar results were validated in mice [55]. Additionally, the *IGF1R* and *IGF2R* genes were reported by Wang et al. [56] to play an important role in pig growth and development.

Conclusion

In this study, we performed population genetic analysis and selection signature detection on Asian wild boars, and South China and North China indigenous pigs using whole-genome resequencing data. The population

genetic structure analysis showed that North China indigenous pigs are genetically closer to Asian wild boars than South China indigenous pigs. Both South China and North China indigenous pigs have been selected for growth, meat quality, and reproductive traits, but there are differences in the selection directions. South China indigenous pigs were more selected for fat deposition ability, neural development, small body size, and coat color, whereas selection signatures in North China indigenous pigs were more related to blood physiology, immunity, and reproductive performance. Furthermore, we identified multiple selected genes associated with reproductive physiology, pigmentation, and immune function in both South China and North China indigenous pigs.

Abbreviations

SNP	Single nucleotide polymorphism
WGS	Whole genome sequencing
QTL	Quantitative Trait Locus
Mb	Mega base pair
Kb	Kilo base pair
WBA	Asian wild boars
PCA	Principal component analysis
LD	Linkage disequilibrium
NJ	Neighbor-joining
EigenGWAS	Eigenvector genome-wide association study
XP-EHH	Cross population extended haplotype homozygosity
F _{ST}	Fixation index
GO	Gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
GWAS	Genome-wide association study
PigGTEx	Pig Genotype-Tissue Expression
IMPC	International Mouse Phenotyping Consortium
HPA	Human Protein Atlas
PGRP	Pig genomics reference panel
SSC	Sus Scrofa Chromosome
ADG	Average daily gain
DAYS_115	Days to reach 115 kg from birth
TNB	Total number of piglets born
TLWT_Weaning	Total litter weight at weaning
BFT	Backfat thickness
TNUM	Teat number
CD4LP	CD4 positive leukocyte percentage
CD4CD8NP	CD4 positive, CD8 negative leukocyte percentage=

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-024-11119-y>.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5
Supplementary Material 6
Supplementary Material 7
Supplementary Material 8
Supplementary Material 9
Supplementary Material 10

Supplementary Material 11
 Supplementary Material 12
 Supplementary Material 13
 Supplementary Material 14
 Supplementary Material 15
 Supplementary Material 16
 Supplementary Material 17
 Supplementary Material 18
 Supplementary Material 19

Acknowledgements

We also acknowledge technical support from the National Supercomputer Center in Guangzhou. We thank anonymous reviewers and editors for their constructive comments and suggestions.

Author contributions

ZZ, YG and XF conceived and designed the experiments. SD, YL, ZZ, XC, and GL provided technical assistance and revised the manuscript. ZZ, XC and JT helped to draw the figures. ZZ, JL and XL designed the analysis scheme and revised the manuscript. YG, XF and ZZ drafted the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by the China Agriculture Research System (CARS-35), Guangzhou Science and Technology Planning Project (2024A04J3806), Specific university discipline construction project (2023B10564001, 2023B10564003), Guangdong Province Rural Revitalization Strategy Special Fund Seed Industry Revitalization Project (2022-440000-43010101-9501), the Young Scientists Fund of the National Natural Science Foundation of China (32402714), the National Key R & D Program of China (2023YFD1300400), and Guangxi Science and Technology Program Project (GuikJ23023003).

Data availability

All raw data analyzed in this study are publicly available from CNGB GSA (<https://ngdc.cnbc.ac.cn/>) and NCBI SRA (<https://www.ncbi.nlm.nih.gov/sra/>) databases. Details of WGS dataset can be found in Supplementary Table S1.

Declarations

Ethics approval and consent to participate

Not applicable. No animals or animal materials have been used in this study, and ethical approval for the use of animals was not necessary.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 26 September 2024 / Accepted: 3 December 2024

Published online: 18 December 2024

References

- Duarte CM, Marbá N, Holmer M, Ecology. Rapid domestication of marine species. *Science*. 2007;316:382–3.
- Marom N, Bar-Oz G. The prey pathway: a regional history of cattle (*Bos taurus*) and pig (*Sus scrofa*) domestication in the northern Jordan Valley, Israel. *PLoS ONE*. 2013;8:e55958.
- Zohary D, Tchernov E, Horwitz LK. The role of unconscious selection in the domestication of sheep and goats. *J Zool*. 1998;245:129–35.
- Kijas JM, Andersson L. A phylogenetic study of the origin of the domestic pig estimated from the near-complete mtDNA genome. *J Mol Evol*. 2001;52:302–8.
- Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 2012;491:393–8.
- Giuffra E, Kijas JM, Amarger V, Carlborg O, Jeon JT, Andersson L. The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics*. 2000;154:1785–91.
- Ervynck A, Dobney K, Hongo H, Meadow R. Born free ? New evidence for the Status of *Sus scrofa* at Neolithic Çayönü Tepesi (Southeastern Anatolia, Turkey). *Paléorient*. 2001;27:47–73.
- Jing Y, Flad RK. Pig domestication in ancient China. *Antiquity*. 2002;76:724–32.
- Zeder MA, Emshwiller E, Smith BD, Bradley DG. Documenting domestication: the intersection of genetics and archaeology. *Trends Genet*. 2006;22:139–55.
- Wang LY, Wang AG, Wang LX, Li K, Yang GS, He RG, et al. Animal genetic resources in China: pigs. Beijing, China: China Agriculture; 2011. (in Chinese).
- Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project. *Genome Biol*. 2015;16:57.
- Liu SL, Gao YH, Canela-Xandri O, Wang S, Yu Y, Cai WT, et al. A multi-tissue atlas of regulatory variants in cattle. *Nat Genet*. 2022;54:1438–47.
- Teng JY, Gao YH, Yin HW, Bai ZH, Liu SL, Zeng H, et al. A compendium of genetic regulatory effects across pig tissues. *Nat Genet*. 2024;56:112–23.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–75.
- Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2016.
- Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. *Mol Biol Evol*. 2016;33:1870–4.
- Letunic I, Bork P. Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49:W293–6.
- Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*. 2019;35:1786–8.
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10:giab008.
- Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19:1655–64.
- Szpiech ZA, Hernandez RD. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol*. 2014;31.
- Ma YL, Wei JL, Zhang Q, Chen L, Wang JY, Liu JF, et al. A genome scan for selection signatures in pigs. *PLoS ONE*. 2015;10:e0116850.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–8.
- Chen GB, Lee SH, Zhu ZX, Benyamin B, Robinson MR. EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity (Edinb)*. 2016;117:51–61.
- Hu ZL, Park CA, Reecy JM. Bringing the animal QTLdb and CorDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res*. 2022;50:D956–61.
- Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*. 2016;32:289–91.
- Wu TZ, Hu EQ, Xu SB, Chen MJ, Guo PF, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innov (Camb)*. 2021;2:100141.
- Sherman BT, Hao M, Qiu J, Jiao XL, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50:W216–21.
- Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics*. 2016;32:587–9.
- Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347:1260419.
- Koscielny G, Yaikhom G, Iyer V, Meehan TF, Morgan H, Atienza-Herrero J, et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res*. 2014;42:802–9. Database issue:D.

32. Watanabe K, Stringer S, Frei O, Umičević Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet.* 2019;51:1339–48.
33. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. *Nucleic Acids Res.* 2023;51:D933–41.
34. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–60.
35. Akey JM, Wang H, Xiong M, Wu H, Shriver WL. MD, Interaction between the melanocortin-1 receptor and P genes contributes to inter-individual variation in skin pigmentation phenotypes in a tibetan population. *Hum Genet.* 2001;108.
36. Shi H, Cheng Z. MC1R and melanin-based molecular probes for theranostic of melanoma and beyond. *Acta Pharmacol Sin.* 2022;43:3034–44.
37. Rouzaud F, Martin J, Gallet PF, Delourme D, Goulemot-Leger V, Amigues Y, et al. A first genotyping assay of French cattle breeds based on a new allele of the extension gene encoding the melanocortin-1 receptor (Mc1r). *Genet Sel Evol.* 2000;32:511–20.
38. Marklund L, Moller MJ, Sandberg K, Andersson L. A missense mutation in the gene for melanocyte-stimulating hormone receptor (MC1R) is associated with the chestnut coat color in horses. *Mamm Genome.* 1996;7:895–9.
39. Kerje S, Lind J, Schütz K, Jensen P, Andersson L. Melanocortin 1-receptor (MC1R) mutations are associated with plumage colour in chicken. *Anim Genet.* 2003;34:241–8.
40. April CS, Barsh GS. Skin layer-specific transcriptional profiles in normal and recessive yellow (Mc1re/Mc1re) mice. *Pigment Cell Res.* 2006;19:194–205.
41. Kijas JM, Wales R, Törnsten A, Chardon P, Moller M, Andersson L. Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics.* 1998;150:1177–85.
42. Zhong HW, Zhang J, Tan C, Shi JS, Yang J, Cai GY, et al. Pig Coat Color Manipulation by MC1R Gene Editing. *Int J Mol Sci.* 2022;23:10356.
43. Johansson Moller M, Chaudhary R, Hellmén E, Höyheim B, Chowdhary B, Andersson L. Pigs with the dominant white coat color phenotype carry a duplication of the KIT gene encoding the mast/stem cell growth factor receptor. *Mamm Genome.* 1996;7:822–30.
44. Aoki H, Yamada Y, Hara A, Kunisada T. Two distinct types of mouse melanocyte: differential signaling requirement for the maintenance of non-cutaneous and dermal versus epidermal melanocytes. *Development.* 2009;136:2511–21.
45. Patterson Rosa L, Martin K, Vierra M, Lundquist E, Foster G, Brooks SA, et al. A KIT variant Associated with increased White spotting epistatic to MC1R genotype in horses (*Equus caballus*). *Anim (Basel).* 2022;12:1958.
46. Ai HS, Huang LS, Ren J. Genetic diversity, linkage disequilibrium and selection signatures in Chinese and western pigs revealed by genome-wide SNP markers. *PLoS ONE.* 2013;8:e56001.
47. Zhang Z, Xiao Q, Zhang QQ, Sun H, Chen JC, Li Z-C, et al. Genomic analysis reveals genes affecting distinct phenotypes among different Chinese and western pig breeds. *Sci Rep.* 2018;8:13352.
48. Ceccherini I, Zhang AL, Matera I, Yang G, Devoto M, Romeo G, et al. Interstitial deletion of the endothelin-B receptor gene in the spotting lethal (sl) rat. *Hum Mol Genet.* 1995;4:2089–96.
49. Metallinos DL, Bowling AT, Rine J. A missense mutation in the endothelin-B receptor gene is associated with Lethal White Foal Syndrome: an equine version of Hirschsprung disease. *Mamm Genome.* 1998;9:426–31.
50. Yan SQ, Bai CY, Qi SM, Li ML, Si S, Li YM, et al. Cloning and association analysis of KIT and EDNRB polymorphisms with dominant white coat color in the Chinese raccoon dog (*Nyctereutes procyonoides procyonoides*). *Genet Mol Res.* 2015;14:6549–54.
51. Zhan SY, Ding X, Tao Zhong, Wang LJ, Li L, Zhang HP. Comparison of IGF1R and IGF2R expression patterns in different tissues and muscle cells of Nanjiang Brown Goats. *Acta Vet Et Zootechnica Sinica.* 2019;50:701–11. (in Chinese).
52. Baker J, Liu JP, Robertson EJ, Efstratiadis A. Role of insulin-like growth factors in embryonic and postnatal growth. *Cell.* 1993;75:73–82.
53. Zanol N, Gailly P. Skeletal muscle hypertrophy and regeneration: interplay between the myogenic regulatory factors (MRFs) and insulin-like growth factors (IGFs) pathways. *Cell Mol Life Sci.* 2013;70:4117–30.
54. Li WB, Zhu YL, Ai HS, Guo TF. Identifying signatures of selection related to small body size in pigs. *Acta Vet Et Zootechnica Sinica.* 2016;47:1977–85. (in Chinese).
55. Lee Y, Wang Y, James M, Jeong JH, You M. Inhibition of IGF1R signaling abrogates resistance to afatinib (BIBW2992) in EGFR T790M mutant lung cancer cells. *Mol Carcinog.* 2016;55:991–1001.
56. Wang K, Wu PX, Yang Q, Chen DJ, Zhou J, Jiang A, et al. Detection of selection signatures in Chinese landrace and Yorkshire pigs based on genotyping-by-sequencing data. *Front Genet.* 2018;9:119.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Original Manuscript

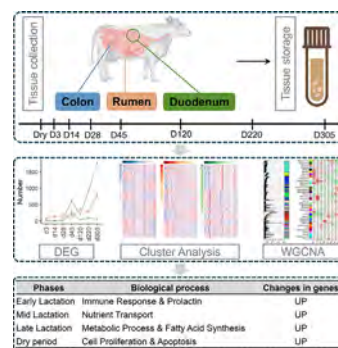
Transcriptomic profiling of gastrointestinal tracts in dairy cattle during lactation reveals molecular adaptations for milk synthesis

Yahui Gao^{a,b,c}, George E. Liu^a, Li Ma^b, Lingzhao Fang^d, Cong-jun Li^a, Ransom L. Baldwin VI^{a,*}^a Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705, USA^b Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA^c State Key Laboratory of Livestock and Poultry Breeding, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China^d Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

HIGHLIGHTS

- Lactation in dairy cattle is a critical period that demands significant adjustments in the rumen and digestive system to meet the increased nutrient requirements for milk production.
- We assembled the transcriptome and compared gene expression patterns in the epithelial tissue of the colon, duodenum, and rumen from dairy cattle in dry and lactating cows.
- The serial sampling approach using biopsied tissues enabled direct comparison of gene expression patterns within and among tissues during different phases of lactation.
- With in-depth computational analyses, this resource provided comprehensive insight into adaptations required in service tissues during lactation in dairy cows and revealed the specific characteristics of gastrointestinal tract tissues and genomic mechanisms controlling the process.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 9 February 2024

Revised 11 June 2024

ABSTRACT

During lactation, dairy cattle's digestive tract requires significant adaptations to meet the increased nutrient demands for milk production. As we attempt to improve milk-related traits through selective pressure, it is crucial to understand the biological functions of the epithelia of the rumen, small intestine, and colonic tissues in response to changes in physiological state driven by changes in nutrient demands

Abbreviations: BARC, Beltsville Agricultural Research Center; cg-like SC, channel-gap-like spinous cells; GO, Gene Ontology; PCA, principal components analysis; PCG, protein-coding genes; QTL, quantitative trait locus; RFI, residual feed intake; SEP, similarly expressed patterns; TPM, Transcripts Per Million; WGCNA, weighted gene co-expression network analysis.

* Corresponding author.

E-mail address: ransom.baldwin@usda.gov (R.L. Baldwin VI).<https://doi.org/10.1016/j.jare.2024.06.020>

2090-1232/© 2024 The Authors. Published by Elsevier B.V. on behalf of Cairo University.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Accepted 21 June 2024
Available online 24 June 2024

for milk synthesis. In this study, we obtained a total of 108 transcriptome profiles from three tissues (epithelia of the colon, duodenum, and rumen) of five Holstein cows, spanning eight time points from the early, mid, late lactation periods to the dry period. On average 97.06% of reads were successfully mapped to the reference genome assembly ARS-UCD1.2. We analyzed 27,607 gene expression patterns at multiple periods, enabling direct comparisons within and among tissues during different lactation stages, including early and peak lactation. We identified 1645, 813, and 2187 stage-specific genes in the colon, duodenum, and rumen, respectively, which were enriched for common or specific biological functions among different tissues. Time series analysis categorized the expressed genes within each tissue into four clusters. Furthermore, when the three tissues were analyzed collectively, 36 clusters of similarly expressed genes were identified. By integrating other comprehensive approaches such as gene co-expression analyses, functional enrichment, and cell type deconvolution, we gained profound insights into cattle lactation, revealing tissue-specific characteristics of the gastrointestinal tract and shedding light on the intricate molecular adaptations involved in nutrient absorption, immune regulation, and cellular processes for milk synthesis during lactation.

© 2024 The Authors. Published by Elsevier B.V. on behalf of Cairo University. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

Dairy cattle are a critically important ruminant species as they supply high quality milk and milk based dairy products for human consumption and agribusiness [1]. Moreover, as the most economically significant trait, improving milk production-related traits remains the most profitable breeding objective [2]. Advanced technology facilitates the genetic improvement of dairy cattle productivity by effectively using genomic information [3]. Lactation is the process by which mammals secrete milk from mammary glands following parturition and in the modern high producing dairy cow this can greatly increase the nutrient and energy demands of the cow. In fact, modern cows can exceed maintenance energy intake by 4 to 5 times during lactation in order to support the synthesis of milk. Thus, during late gestation and lactation, the gastrointestinal tract must undergo significant adaptations to meet increased nutritional demands. These changes, observed in both ruminants and non-ruminants, include (1) Increased size and morphology: expansion of the stomach and intestines to enhance nutrient absorption. (2) Cell proliferation: enhanced turnover of enterocytes, increasing the absorptive surface area. (3) Motility adjustments: Optimized gut motility for better nutrient extraction. (4) Microbiome shifts: Changes in gut microbiome composition to improve digestion and nutrient synthesis. (5) Hormonal influences: Hormonal changes affect gut function and nutrient uptake [4,5].

The lactation cycle in dairy cattle includes four phases: early lactation (d0-d120), mid-lactation (d120-d240), and late lactation (d240-d305), each lasting approximately 120 days, and a non-milking period, commonly referred to as the “dry period,” generally lasting up to 65 days. (<http://www.holsteinfoundation.org/>). This cycle of production is coupled with increased nutrient requirements to support milk synthesis, which in modern dairy cattle increases her energetic needs from 1-times maintenance to 4- or 5-times maintenance [6]. Therefore, dairy cow physiology must adapt to the new nutritional demand caused by lactation, and this adaptation includes major functional adjustments to the rumen and digestive system. For dairy cattle, the common practice is to feed a diet with increased fermentability immediately before parturition. After parturition, a more fermentable diet is further boosted to meet the energy demands arising from the onset of lactation [7]. Quantity and composition of milk produced follow a characteristic pattern, increasing from early postpartum to a peak, followed by a slow decline until the next dry period (typically about 305 d). Therefore, in response to these changes in milk production nutritional demand varies substantially with the lactation phase [4].

Despite extensive research on the liver [8–11] and mammary glands [12–16] of ruminants, there have been few studies on the

gastrointestinal tract [17–19]. Examining the gastrointestinal tract throughout lactation is crucial due to the fluctuating nutrient demands associated with milk production. Gastrointestinal tissues play a pivotal role in nutrient absorption, processing, and assimilation, significantly contributing to the cow's energy requirements by utilizing and supplying nutrients [20]. Understanding the biological functions of the epithelia of the rumen and digestive tract during lactation is indispensable for improving milk-related traits [21]. Transcriptomics, which reflects tissue- and situation-specific genomic activities, can reveal lactation-specific adaptations in the gastrointestinal tract. While several studies have documented gene expression during development in humans [22], mice [23], zebrafish [24], sheep [25], pigs [26], and cattle [27], research on the gastrointestinal tissues of dairy cattle remains limited. This gap highlights the need for more comprehensive investigations into the transcriptomic dynamics across the entire lactation period to better understand these essential adaptations.

To explore the molecular basis for adaption by the gastrointestinal tract to provide the increased nutrient requirements for milk synthesis in dairy cattle, we used RNA-seq to profile gene expression patterns in three tissues, the colon, duodenum, and rumen at eight-time points across the lactation cycle including the dry period and various days in milk (DIM) during lactation (dry period and 3, 14, 28, 45, 120, 220, and 305 DIM). Multiple sampling allow direct comparison of expression patterns within and among tissues during different lactation phases. In addition, we split the early lactation stage into early and peak lactation stages (Fig. 1A). Overall, this resource provided comprehensive insight into cattle gastrointestinal responses to lactation and revealed the specific characteristics of gastrointestinal tract tissues to understand the complexity of genomic activities during lactation.

Results

Global profiling of transcriptomes in gastrointestinal tracts at different lactation stages

To elucidate the potential molecular networks regulating gastrointestinal response to lactation, we used RNA-seq to quantify the transcriptomes of 108 samples, including three tissue types (colon, duodenum, and rumen) at eight discrete days in milk (DIM) which represent lactational stages (dry and 3, 14, 28, 45, 120, 220 and 305 DIM; Fig. 1A). Cows were fed a corn silage-based ration (46.7 ± 0.2 %DM as fed; 17.92 ± 0.08 % crude protein) with a Net Energy for lactation value of 0.76 ± 0.002 in order to support milk production over the 305 d of $10,075 \pm 855$ kg with 3.74 ± 0.16 % milk fat and 3.14 ± 0.07 % protein. The average number of mapped reads we obtained was over 20.42 million with an aver-

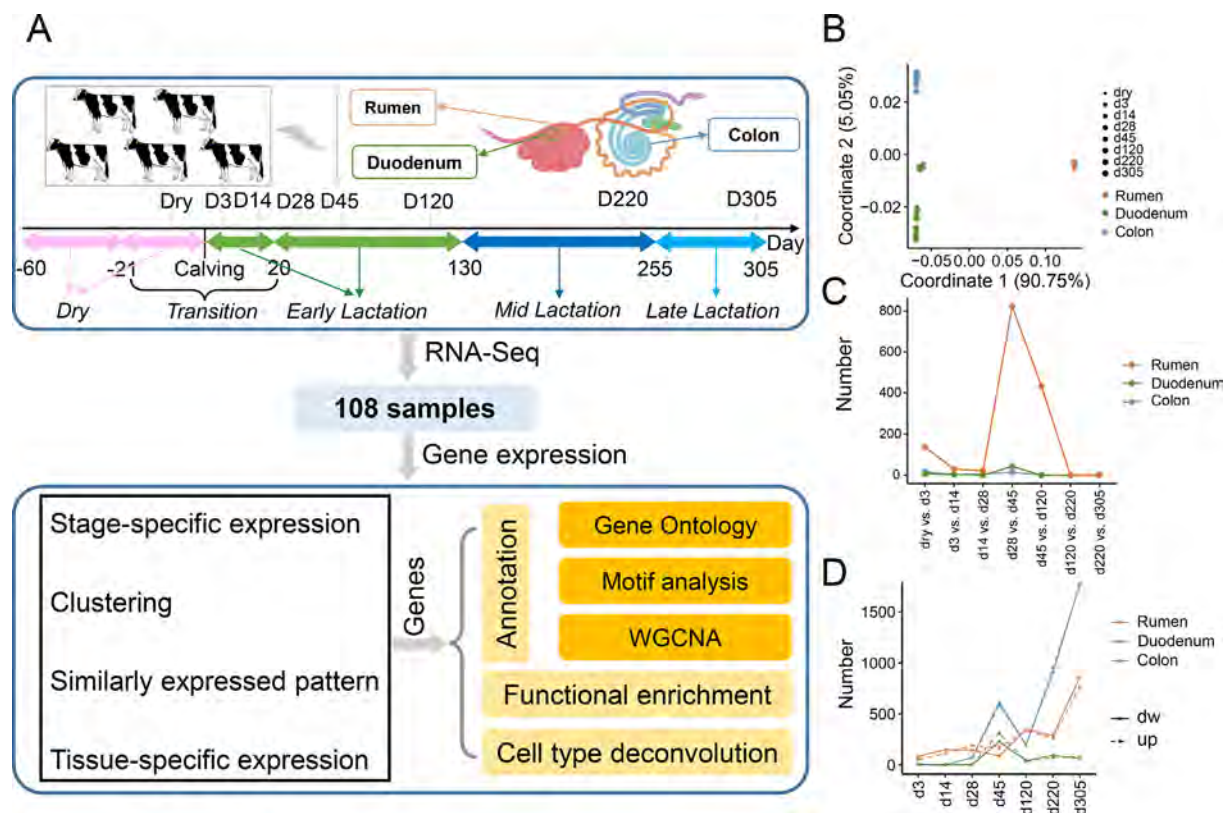


Fig. 1. (A) Tissue collection and global analysis of the cattle transcriptomes. We explored expression specificity patterns by multiple analyses (differentially expressed gene analysis, stage-/tissue-specific expression, co-expression network analysis, functional enrichment, deconvolution analysis, etc.) using 108 samples from three gastrointestinal tissues (rumen, duodenum and colon) from eight lactation periods of dairy cattle. (B) The Principal Component Analysis (PCA) of three tissues across eight stages. Tissues are represented by different colors, whereas the size of the dots indicates different lactation stages. (C) The number of genes differentially expressed between adjacent stages for three tissues. (D) The number of genes differentially expressed between dry and other stages for three tissues.

age mapping rate of 97.06 % (Table S1). Within each tissue, the distribution of mapped reads and mapping rate were similar for each lactation stage, indicating good quality generally (Fig. S1A). We subjected each gene to the principal components analysis (PCA) by its normalized expression (Transcripts Per Million, TPM). Altogether, with more than 90 % of the variance explained by the first three PCs, PCA for these three tissues depicted the predominant clustering according to tissue types (Fig. 1B). The PCA exhibited a striking separation between the three tissues, indicating the activation of a potential different transcript profile in each tissue. The various stages of the colon and rumen replicates were closely clustered. At the same time, the duodenum replicates exhibited a clear transition from one stage to another (Fig. 1B). We specified genes with TPM > 0.1 as those expressed. Based on the expressed genes, we identified the differentially expressed genes (DEGs) in adjacent stages of the three tissues (Fig. 1C). The trend curves were similar in the colon and duodenum but showed an inverse V-shape in the rumen. When the dry stage was used as a Control to compare the other stages, the trend curves were broadly similar for the three tissues in the first five time points. However, they differed in the last two time points (Fig. 1D). When comparing all the possible pairwise stages, the rumen and colon showed similar patterns, but the duodenum trend curve was distinct and flat (Fig. S1B). The similar patterns observed in the rumen and colon could be attributed to their shared roles in fermentation and similar microbial environments, which might influence similar gene expression pathways involved in immune response and metabolic processes. In contrast, the duodenum, primarily engaged in nutrient absorption, exhibits a more stable expression pattern, possibly due to its consistent nutrient exposure and lower microbial variability.

Identification of stage-specific genes for different lactation stages

Gene expression shows different fluctuations during lactation stages. The functions of stage-specific genes can reflect the known biological attributes of their respective stage. We identified 1645, 813, and 2187 stage-specific genes in the colon, duodenum, and rumen, respectively (Table S2), showing distinct patterns at different lactation stages in each tissue (Fig. 2). Gene Ontology (GO) enrichment analysis unveiled that in the dry period, upregulated genes in the colon were mainly involved in cell proliferation, especially in the epithelium (Fig. 2A, Table S3). In contrast, upregulated genes in the duodenum and rumen were primarily associated with immune functions (Fig. 2B–C, Table S3). In the early lactation stage (d3), upregulated genes in the colon and duodenum were mainly involved in immune response (Fig. 2A–B, Table S3). In contrast, those in the rumen were implicated in cell proliferation and fatty acid metabolic process (Fig. 2C, Table S3). In the peak (d14, d28, d45, d120) and mid (d220) lactation stages, upregulated genes of three tissues were significantly enriched in material digestion and absorption, immune functions, and cell cycle activities (Fig. 2, Table S3). In late lactation, upregulated genes in the colon still mainly participated in lipid metabolism, while those in the duodenum and rumen were primarily involved in immune functions (Fig. 2, Table S3). Similarly, downregulated genes of three tissues in different stages were mainly engaged in the same pathways (Fig. 2, Table S3).

To probe whether specific TFs regulated stage-specific genes, we performed a motif enrichment analysis of stage-specific genes either upregulated or downregulated for each tissue separately (Table S4–S6). In summary, in the three tissues, the motifs mainly

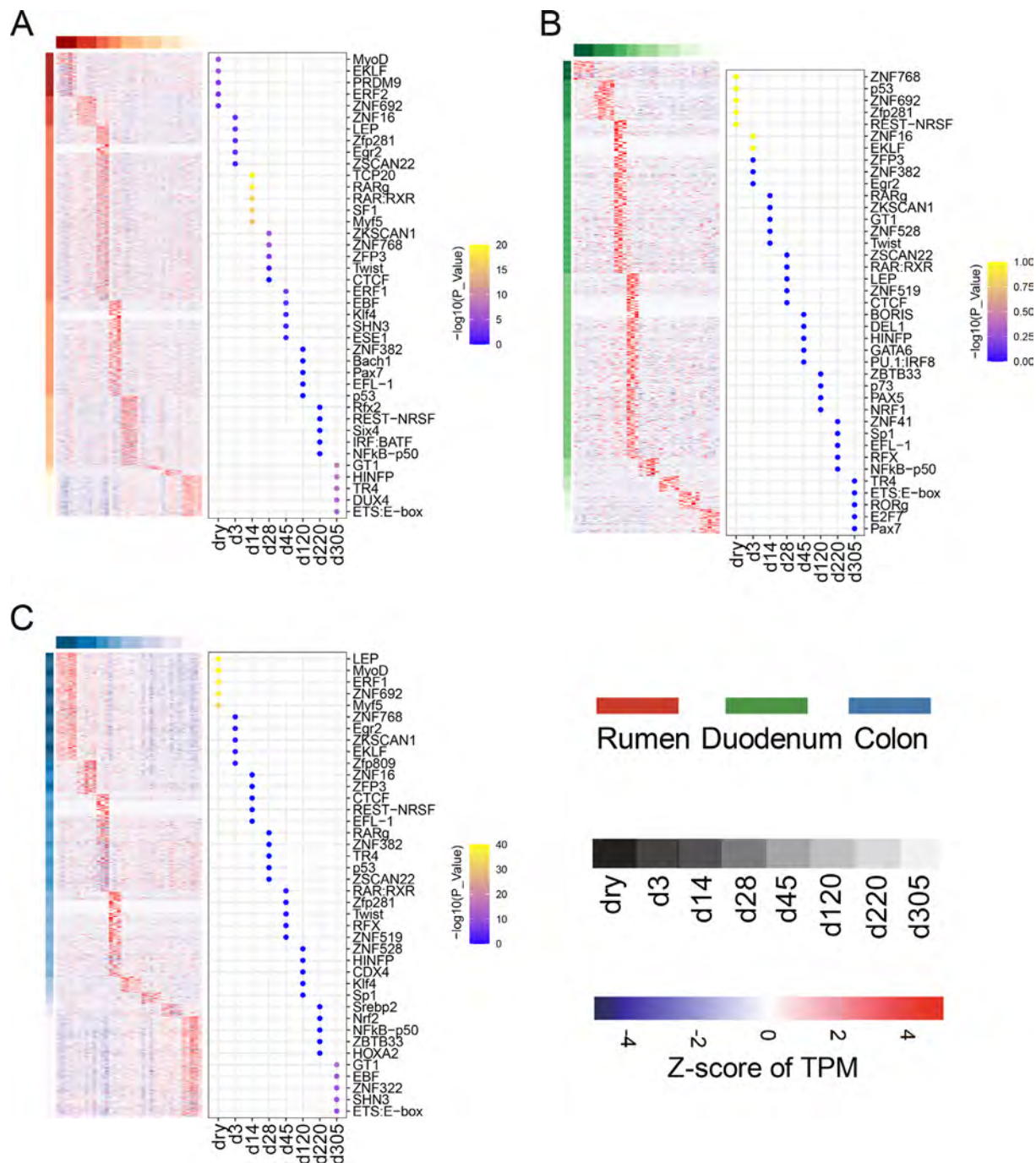


Fig. 2. Dynamic expression patterns of stage-specific genes across stages. (A) For the rumen, the heatmap shows the expression level (Z-score of TPM) of upregulated stage-specific genes across stages; the dot plot shows motifs of transcriptional factors (TFs) are significantly enriched in promoters of upregulated stage-specific genes. The same patterns are for the duodenum (B) and colon (C).

included cell differentiation-related motifs (MyoD, ZNF16, and ZNF382, etc.), adipocytes related-motifs (LEP, Twist, etc.), and immune-related motifs (Zfp809, RFX, and NFkB-p50, etc.) (Fig. 2 and Fig. S2). Furthermore, we detected different zinc finger proteins at each stage. Some motifs were shared among three tissues, such as ZNF692 and ETS: E-box. Therefore, the transition from the dry period to the lactation stages was accompanied by marked fluctuations in the expression of genes related to cell differentiation, nutrient process, transport, and immune functions. Through these comparisons, we gained specific insights into the molecular and physiological adaptations that occur in different tissues to meet the demands of lactation. For instance, there is a shift from

cell proliferation in the dry period to immune response in early lactation in the rumen and an enrichment of nutrient transport genes in the colon during peak lactation (d28).

Clustering of RNA-seq data for each tissue

To identify different gene expression profiles across lactation in each tissue, we next performed clustering using the maSigPro algorithm [28]. We identified four clusters in each tissue, and the number of genes covered by these clusters ranged from 19 to 654 (Fig. 3A, Table S7), which exhibit distinct expression patterns (Fig. 3B). In the colon, genes in clusters 1 and 2 showed similar

expression patterns, being upregulated at dry, d3 and d14, but downregulated at other stages. In general, the expression of cluster 2 genes was slightly higher than that of cluster 1. In contrast, genes of cluster 3 were downregulated in dry, d3, and d14 but upregulated in d45, d120, and d305, while genes of cluster 4 were only upregulated in d28 (Fig. 3B). GO enrichment revealed that genes from clusters 1 and 2 were mainly concerned with pathways of cell cycle activities and immune functions (Fig. 3C, Table S8). At the same time, they shared some motifs (Fig. 3D, Table S9). Genes of cluster 3, upregulated in d45, d120, d220, and d305, were mainly involved in the metabolism process, such as protein secretion and lipid biosynthetic process (Fig. 3C, Table S8). The enriched motifs were associated with basic necessary biological processes (Fig. 3D, Table S9). Genes of cluster 4 were upregulated in d28 and were mainly involved in keratinization and immune function (Fig. 3C, Table S8), with the corresponding motifs primarily associated with different zinc finger proteins (Table S9).

In the duodenum, genes from cluster 1 were upregulated at dry, d3 and d28, genes from cluster 2 on days 45, 120, and 220, genes from cluster 3 on day 14, and genes from cluster 4 on d305 (Fig. 3B). Different clusters of genes were enriched in different pathways and motifs. Cluster 1 showed enrichment mainly in metabolic processes and immune functions. In addition to immune functions, cluster 2 was involved in cell cycle activities. While both clusters 3 and 4 were related to fundamental biological processes

(Fig. 3C, Table S8). Each cluster's enriched motifs were similar to the corresponding GO enrichments (Fig. 3D, Table S9).

Similarly, in the rumen, genes of cluster 1, upregulated at d3, d14, and d28, were involved in various pathways such as energy metabolism, cell cycle activities, fatty degradation, and immune function. Genes of cluster 2, upregulated in d45, were in the immune function pathways. Genes of cluster 3, upregulated in d120, d220, and d305, participated in cell cycle activities. Genes of cluster 4, upregulated in d305, were mainly engaged in the immune function (Fig. 3C, Table S8). The motifs of each cluster exhibited comparable to the corresponding GO enrichments (Fig. 3D, Table S9).

Co expression gene network analysis

We modeled networks using the unsigned weighted gene co-expression network analysis (WGCNA) [29] between genes to explore the associations among modules and three tissues across different stages. WGCNA revealed distinct gene co-expression modules in the colon, duodenum, and rumen, each containing genes with similar expression patterns. We detected 36, 36, and 14 co-expression modules for the rumen, colon, and duodenum, respectively (Fig. 4). For rumen, d3, d14, d28, d45, and d305 were correlated with brown, orange, tan/dark red, black, and turquoise/yellow/royal blue/blue modules, respectively (Fig. 4A). These mod-



Fig. 3. Clustering of genes for each tissue during lactation time-course. (A) For the rumen, all genes were clustered into four non-redundant groups using maSigPro. Temporal expression profiles of the four clusters. The red lines represent median gene expression levels, and the gray lines represent gene expression levels for each gene in the relative cluster during stages; Heatmap of the expression levels (Z-scores of TPM) of genes in each cluster; GO terms for all genes in each cluster; Motifs of transcriptional factors (TFs) are significantly enriched in promoters of cluster genes. The same patterns are for the duodenum (B) and colon (C). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ules were also involved in multiple biological processes such as metabolic pathways (brown, black), basic biological process (orange), immune functions (tan/dark red, black, royal blue), cell cycle (turquoise), nutrition metabolism (yellow, blue) (Fig. 4A and Table S10). For the colon, dry period, d3, d28, and d305 were correlated with purple/turquoise, dark red, steel blue, and saddle brown modules, respectively, as labeled by stars (Fig. 4B). GO enrichment revealed that genes coexpressed in these modules were associated with diverse pathways such as muscle activities (purple and turquoise modules), immune functions (purple, turquoise, dark red module, saddle brown modules), the absorption and digestion of materials (steel blue, saddle brown modules) (Fig. 4B and Table S10). For the duodenum, d45 was correlated with the pink module, which was mainly involved in the cell cycle activities (Fig. 4C). D120 was correlated with the purple module, displaying enrichment in immune functions (Fig. 4C and Table S10). These findings provide a detailed understanding of the molecular mechanisms and biological processes active at different stages of lactation in various gastrointestinal tissues. For example, the identification of immune response genes in the rumen during early lactation (d3) underscores the importance of immune function in adapting to the physiological changes at the onset of lactation. Similarly, the enrichment of nutrient transport genes in the colon during peak lactation (d28) highlights its critical role in nutrient absorption to support high milk production.

Clustering of RNA-seq data for all tissues

In addition, we clustered all genes expressed during the lactation stages across the three tissues. We classified the expressed genes into 14 clusters based on each tissue's Z-score of expression levels (Fig. 5A, Table S11). Cluster 4 had the highest number of genes (1,853) relatively highly expressed in the colon. Similarly, clusters 3 and 5, the other two colon-specific clusters, had a steady expression of genes throughout lactation. Cluster 2, the second largest cluster with 1,430 genes, was relatively highly expressed in both the colon and duodenum. Clusters 7 and 10 were composed of genes primarily expressed in the duodenum and rumen, respectively. GO enrichment analysis of genes in each tissue-specific cluster revealed that the colon (C3) was mainly involved in basic necessary biological processes (Fig. 5B, Table S12); the duodenum (C10) was involved primarily in metabolism processes such as fat digestion and absorption (Fig. 3F, Table S12); the rumen (C7) was mainly involved in diverse pathways, such as cell cycle activities, fatty degradation, and immune function (Fig. 5B, Table S12).

Moreover, we were interested in shared clusters between three tissues to explore the common biological processes during the transition from dry to lactation stages. To this end, we utilized TimesVector [30] to cluster similarly expressed patterns (SEP). We detected 36 SEP clusters (Fig. 5C), indicating that the genes of the three tissues were in a similar pattern overall. Cluster 223 had the highest number of genes (814 genes), followed by cluster 439 (407 genes), 49 (145 genes), and 78 (127 genes) (Table S13). We subsequently performed GO enrichment analysis to determine the functional roles of the genes in the SEP clusters (Table S14). GO analyses of the genes in cluster 223 suggested significant enrichment in various biological processes, including cell cycle activities, methylation, and immune function (Fig. 5D). Genes in cluster 439

were mainly involved in cell differentiation and metabolism of fat and protein. In contrast, those in clusters 49 and 78 were associated with basic necessary biological processes (Fig. 5D). These results suggested that the three tissues probably shared similar patterns of cell cycle activities and immune function during the transition from dry to lactation stages.

Tissue specific expression patterns in gastrointestinal tracts

To illuminate the biological functions of tissue-specific gene expression patterns across three tissues, we employed TAU values to estimate the specificity of each tissue. Gene expression was either tissue-specific or ubiquitous (Fig. 6A). Nonetheless, the expression of PCG is less tissue-specific than that of non-coding genes (Fig. 6B). We detected between 453 (rumen) and 811 (duodenum) tissue-specific expressed genes across three tissues (Table S15). These tissue-specific genes showed distinct expression patterns (Fig. 6C and Fig. S3–S5), functions of which reflect the known biology of their respective tissues. GO enrichment revealed that colon-specific genes mainly participate in membrane-related activities; duodenum-specific genes associated with the metabolic process, and rumen-specific genes were involved in diverse processes, including keratinization, amino acid metabolism, and immune functions (Fig. 6D, Table S16). Lastly, we calculated the stage-specificity of each gene across lactation stages. Stage- and tissue-specificity are highly correlated: tissue-specific genes are more potentially expressed in a slimmer time window, and vice versa.

Functional enrichment and cell type deconvolution analysis

Many studies on the bovine rumen use different omics data to date. Hence, we conduct a subsequent analysis using the rumen as an example. By examining the 15 previously predicted chromatin states [31], we found that specific genes for the dry, d3, and d14 showed a higher enrichment of active regulatory elements than other stages. In comparison, dry- and d3-specific genes also had a higher enrichment of enhancers than others (Fig. 7A). All the stage-specific genes except d220 showed a higher depletion of repressed regions. Furthermore, we downloaded cattle QTLs from Animal QTLdb (release 48, Aug. 24, 2022) [32]. We classified them into six categories, i.e., health, production_meat, production_milk, reproduction, residual feed intake (RFI), and type (Fig. 7A). The functional enrichment exhibited that dry-specific genes were highly correlated with health QTLs; d28-specific genes were correlated with production QTLs; d120- and d220-specific genes were correlated with reproduction QTLs; and only d45-specific genes were lightly correlated with RFI QTLs (Fig. 7A). In the meantime, we tested the enrichment of 15 chromatin states and QTLs with rumen-specific genes. We found that rumen-specific genes highly enriched for active regulatory elements and milk-production, health, and reproduction QTLs (Fig. 7B). To exploit the cell type components in the rumen tissue, we conducted a cell type deconvolution analysis in the rumen tissue, demonstrating the variation of cell type composition across bulk tissue samples. Deconvolution analysis indicated that channel-gap-like spinous cells (cg-like SC) constituted the most significant percentage of rumen tissue samples (Fig. 7C). This finding suggests

Fig. 4. The weighted gene co-expression network analysis (WGCNA). (A) For the rumen, the left plot shows that functional modules are represented in different colors. Each major branch represents a color-coded module that contains a group of highly connected genes; the heatmap shows correlations between gene modules and lactation stages. The statistical significance of the module-lactation stage relationship was corrected by multiple testing using the FDR method. Each cell contains the correlation and the corresponding FDR value in the bracket; the right plot shows GO terms for all genes in significant modules. The same patterns are for the duodenum (B) and colon (C).



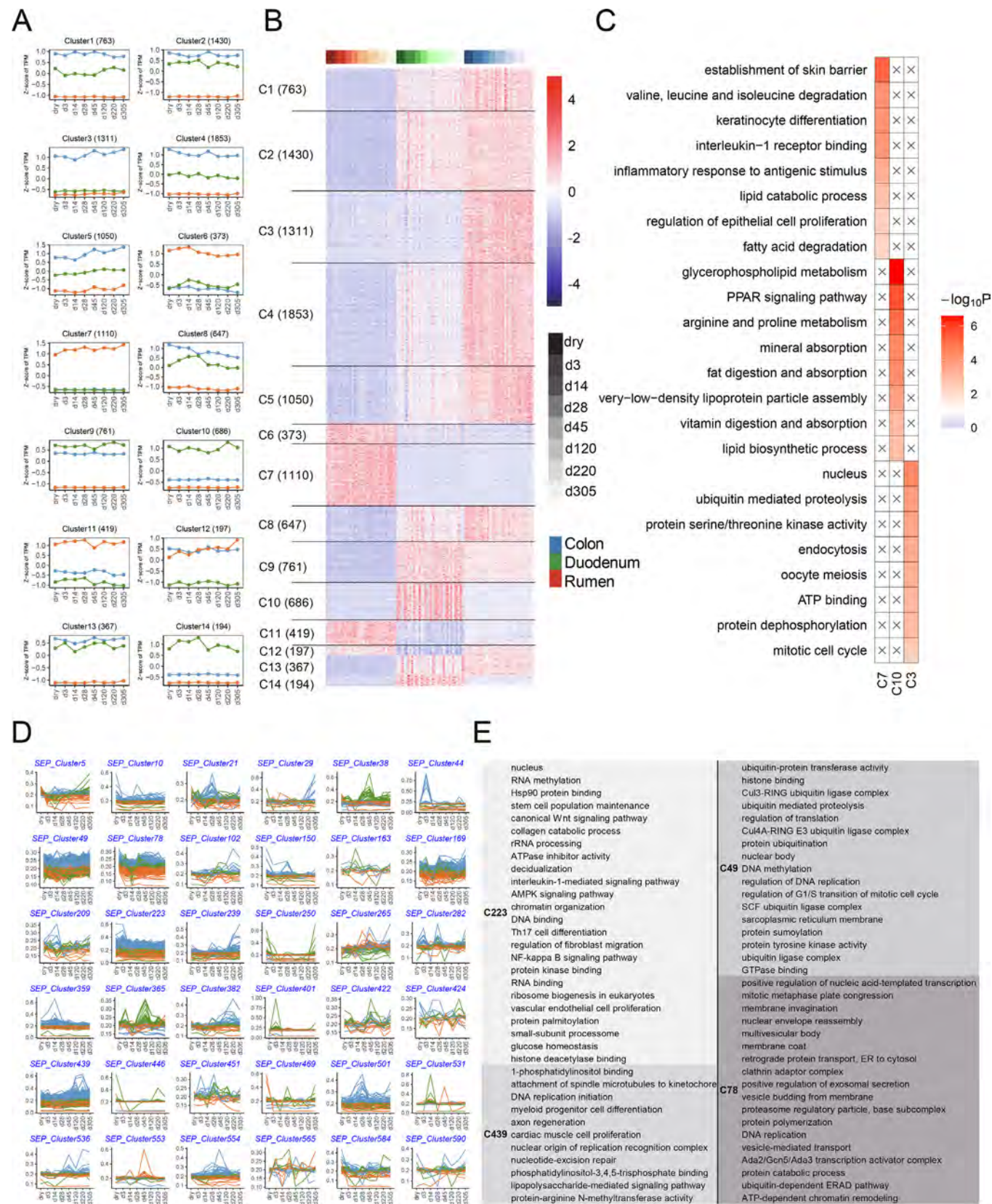


Fig. 5. Clustering of RNA-seq data for all tissues. (A) All genes from three tissues were clustered into 14 non-redundant groups using MaSigPro, based on the median expression level of the genes in each cluster. (B) Heatmap of the normalized expression levels (Z-scores of TPM) of genes in each cluster. (C) GO terms for genes in clusters 3, 7, and 10. (D) 36 SEPs were obtained. The expression patterns of genes in each cluster per tissue. Each line color represents a tissue color corresponding to Fig. 1B. The x and y axis represent the time points and normalized gene expression levels, respectively. (E) GO terms for genes in clusters 223, 439, 49, and 78.

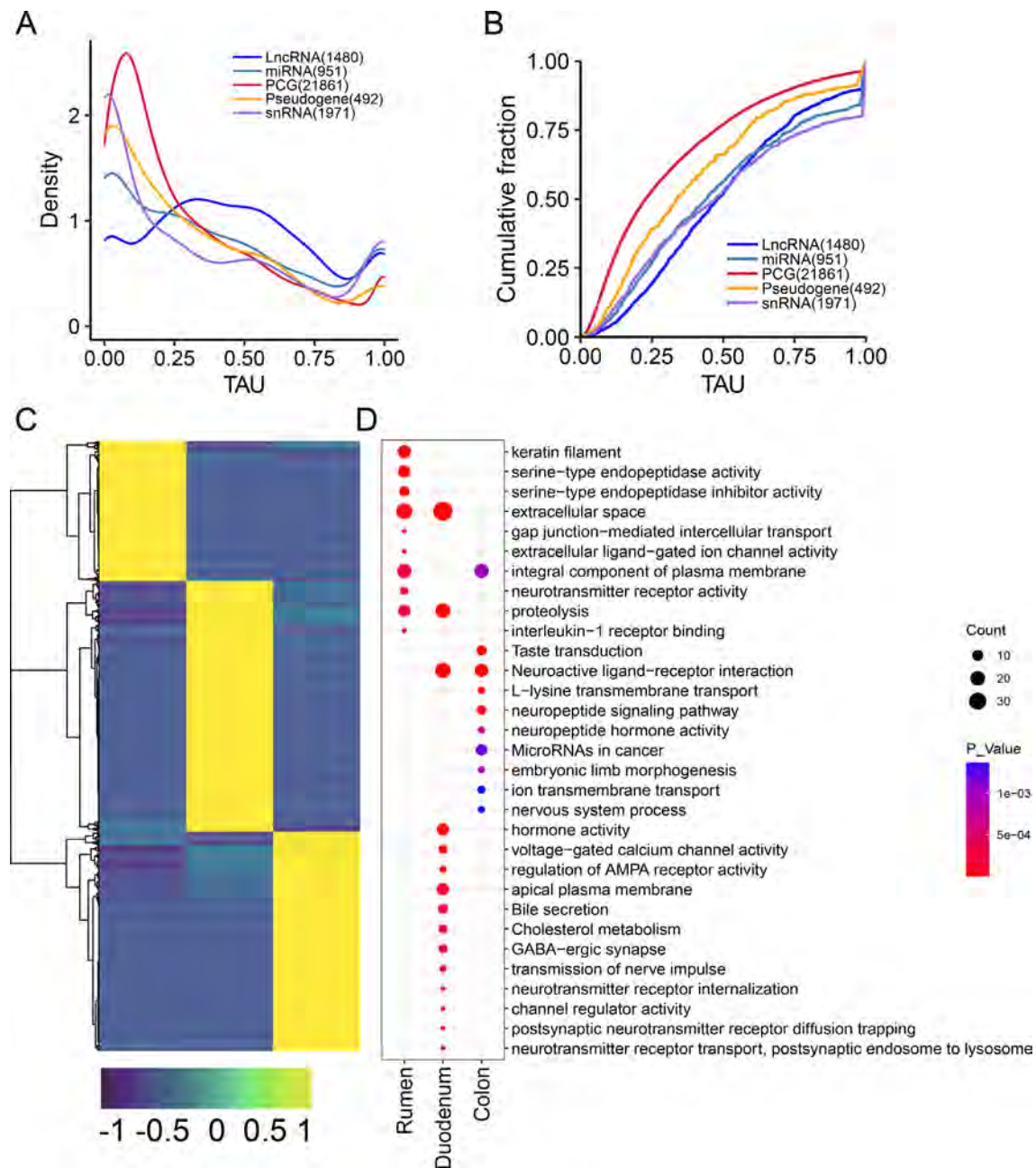


Fig. 6. Tissue specificity. (A) Tissue-specificity distributions represented by TAU values in different genotypes. PCG: protein-coding genes. (B) Tissue-specific expression of five transcript types reflected by the TAU score. (C) The expression patterns of tissue-specific genes in three tissues. (D) The significantly enriched GO terms.

these cells are likely involved in maintaining the structural integrity of the rumen epithelium and facilitating nutrient absorption, as previously reported by Wu et al. [33].

Discussion

Tissue-specific or physiological phase-specific transcriptome represents the functional outcomes of the genome. It can be interpreted as controlling the tissue response to the phase or a reaction to the animal's physiological needs. Transcriptomic analysis has driven many of the advances in functional genomic research. This report used transcriptomic profiling to explore the molecular basis of adaption to provide for the increased nutrient requirements for milk synthesis during lactation in gastrointestinal tracts (epithelia of the colon, duodenum, and rumen) of dairy cattle. We assembled

the transcriptome and compared gene expression patterns in the colon, duodenum, and rumen epithelial tissue from dairy cattle across the lactation cycle.

The gastrointestinal epithelium is a highly metabolically active tissue that performs significant functions such as absorption, transport, and protection. Total gastrointestinal tissues consume a disproportionate amount of the energy that the animal uses (about 25 % of total oxygen consumption), considering its relative size (about 6 % of body weight) [20]. Elucidating the molecular mechanisms behind the transition from dry to lactating by the gastrointestinal tract broadens our insights into milk-producing characteristics in dairy cows. The present study presented a comprehensive profile of gene expression dynamics in three tissues at eight lactation stages (dry, d3, d14, d28, d45, d120, d220, and d305). Our catalog of tissue-specific and stage-specific genes

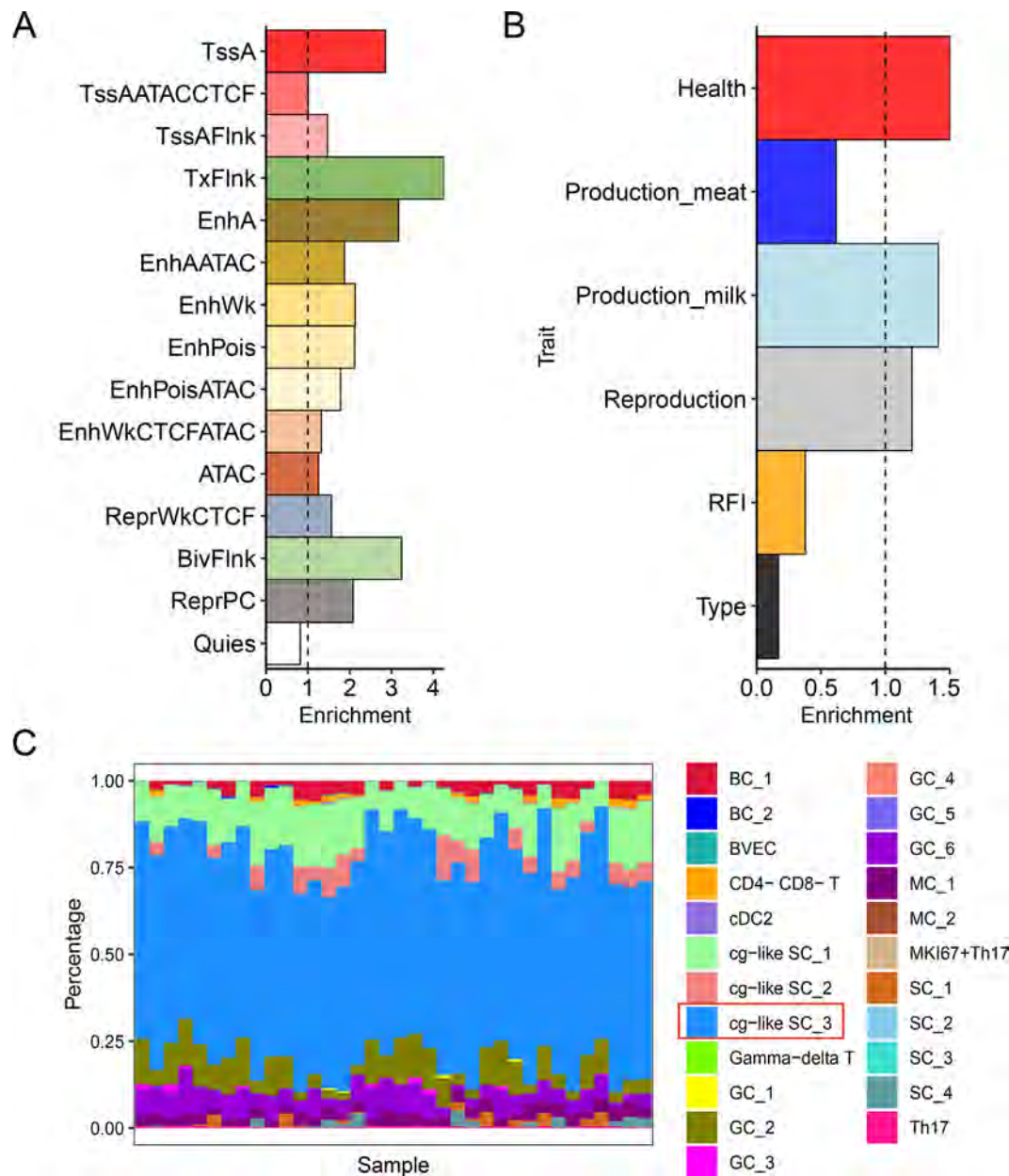


Fig. 7. Functional enrichment and cell type deconvolution analysis. (A) Enrichments of stage-specific genes for chromatin states and QTLs in rumen tissue. (B) Enrichments of rumen-specific genes for chromatin states and QTLs. (C) Distribution of relative cell fractions across 36 rumen samples. BC: basal cell; BVEC: blood vascular endothelial cell; CD4 T: helper T cells; CD8 T: cytotoxic T cells; cDC2: conventional type 2 dendritic cell; cg-like SC: channel-gap-like spinous cell; Gamma-delta T: $\gamma\delta$ T cell; GC: granule cell; MC: mitotic cell; MKI67 + Th17: MKI67 + T helper type 17 cell; SC: spinous cell; Th17: T helper type 17 cell.

delivered an extensive resource for elucidating the transcriptional diversity in the gastrointestinal tracts of cattle during the transition from the dry period to the lactation stages.

We detected DEGs correlated with changes in milk production between lactation stages through pairwise or time series analyses (Table S3 and Table S8). Examples include (1) Early Lactation (d0-120): Immune Response Genes: Active to protect against infections; Prolactin: Stimulates milk production and milk protein synthesis. (2) Mid Lactation (d120-d240): Nutrient Transport Genes: Essential for meeting high nutrient demands. (3) Late Lactation (d240-d305): Metabolic Process Genes: Upregulated to manage declining milk production; Fatty Acid Synthesis Genes: Maintain milk fat content. (5) Dry Period (Non-milking period): Cell Proliferation Genes: Promote tissue regeneration; Apoptosis-related Genes: Remove old or damaged cells. These genes generally reflect

the physiological adaptations needed for milk production at different lactation stages. Interestingly, casein and whey protein genes, crucial for maintaining milk yield and quality, were not detected in these comparisons. This might be due to the gastrointestinal tissues not being the primary sites for their expression.

Specifically, at parturition, the mammary gland is triggered to begin milk secretion. In modern dairy, producers target a complete cycle of lactation from a dry period of 60 days and lactation that lasts approximately 305 days. Several physiological changes occur as cows progress through lactation. The period between three weeks before and three weeks post-parturition is considered the transition period, a crucial physiological stage as cows increase mammary cell numbers and metabolic machinery to begin producing milk. As milk production begins, cows may experience metabolic and infectious diseases within this phase due to the

dramatic physiological and nutritional shifts that must occur [33]. During the transition in our results by integrating stage-specific analysis (Fig. 2 and Table S3), clustering (Fig. 3 and Fig. 5), and WGCNA (Fig. 4), we observed, in all three tissues, changes in barrier structure, nutrient transport, cell cycle activities, and immune functions. The results obtained aligned with the hypothesis put forward by Bath et al [17]. During the dry period of the colon, changes in barrier structure were the most prominent change, which could be explained as preparation for increased ration intake requiring enhanced nutrient flux and immune function changes in support of milk production. For both the duodenum and rumen, changes in immune functions were apparent. On d3, striking changes in immune function gene expression begin to occur in the rumen (Table S3), as is also the case for the colon (Table S3), while the duodenal gene expression changes are associated with both immune functions as well as nutrient transport (Table S3), which is consistent with the rapid increase in nutrient demands of the mammary gland which results in a rapid increase in dry matter intake. In the mid-transition period, d14, all three tissues exhibit nutrient transport-related gene expression changes. During peak lactation (d28–d120), expression of genes related to immune function and nutrient transport is apparent for all three tissues. In mid-lactation (d220), the tissue changes in the expression of nutrient transport function-related genes were altered. In late lactation (d305), the colon tissue expression of nutrient transport genes differed, while the duodenal and ruminal epithelia had more apparent changes in immune function-related gene expression. In general, changes in the duodenum more resemble those of the rumen, possibly because these tissues are fundamentally responsible for absorbing nutrients and/or initiating the assimilation of nutrients for metabolism in support of lactation. Colonic tissue has essential immune and solute transport functions but less of a nutrient absorption role.

Increased immune-related gene expression during lactation is consistent with the fact that the intestines serve as a barrier to luminal contents as a vital immune organ. This function requires a complex cellular network, secreted peptides and proteins, and interactions with other host defenses. Innate immunity plays a critical role in intestinal immune defense against invading pathogens. It also functions as a conduit for activating the adaptive immune system. Pattern recognition molecules of microorganisms are a fundamental component for pinpointing invading pathogens. Initiating the innate immune response and accelerating a unique pattern recognition in the extracellular matrix for microbial pathogens, such as lipopolysaccharides, is crucial. Thus, it is possible that immune cells within the epithelium layer (like CD4 T cells or $\gamma\delta$ T cells) can bind directly to bacteria and their components and serve as an opsonin for macrophage phagocytosis of bacteria [30]. In contrast, it is interesting that T helper type 17 (Th17) cells (highly expressing CD4 and IL17A) were enriched in the rumen and other forestomach tissues. They can regulate the capacities of epithelial cells to uptake short-chain fatty acids through IL-17 signaling [34].

From the perspective of tissue-specific clustering of gene expression patterns, the colon mainly exhibits changes in essential biological processes, the duodenum exhibits changes in nutrient transport. In contrast, the rumen exhibits changes in a variety of processes, including barrier structure, nutrient transport, cell cycle activities, and immune function, highlighting the unique functional roles of each tissue. Integrating these tissue-specific findings across lactation, we observe similar patterns to those observed by tissue-specific clusters, confirming that genes such as *ST6GAL2* (colon), *TMPRSS15* (duodenum), and *SERPINB10* (rumen) were expressed with high tissue specificity (Fig. S3–5 and Table S15). At the same time, however, similar expression patterns indicate shared common biological functions. While *ST6GAL2* encodes a sialyltransferase, *TMPRSS15* encodes an enzyme that converts the

pancreatic proenzyme trypsinogen to trypsin, which activates other proenzymes including chymotrypsinogen and procarboxypeptidases. Diseases associated with *TMPRSS15* include enterokinase deficiency and diarrhea. Additionally, one of the related pathways for *SERPINB10* is the innate immune system.

When we assessed chromatin status using the QTL database to perform enrichment analyses, the rumen was found to be highly enriched for active regulatory elements and health traits, indicating that the rumen tissue expression is dynamic during lactation. Channeled slit-like spine cells (cg-like SCs) were found to be most prevalent within rumen tissue samples. Spinous cells, or prickly cells, are keratin-producing epithelial cells that are spiny in shape due to characteristically high numbers of intracellular connections. They constitute the stratum spinosum (prickly layer) of the tissue important for a continuous reticular protective layer to exclude large molecules and microbes from being absorbed. Thus, to prepare for, or in response to, the dramatic changes in nutritional demand of the lactating cow, restructuring by the ruminal epithelium during the transition period is necessary to ensure tissue integrity while meeting increased nutrient absorptive capacity.

Conclusion and future directions.

We assembled the transcriptome and compared gene expression patterns in the epithelial tissue of the colon, duodenum, and rumen from dairy cattle in dry and lactating cows. The serial sampling approach using biopsied tissues enabled direct comparison of gene expression patterns within and among tissues during different phases of lactation. With in-depth computational analyses, this resource provided comprehensive insight into adaptations required in service tissues during lactation in dairy cows and revealed the specific characteristics of gastrointestinal tract tissues and genomic mechanisms controlling the process. This study identified many DEGs from RNA-seq data using three or five replicates. However, due to the complex nature of gene regulatory networks, key signatures such as main hubs and master regulators are not yet evident. This may limit our ability to further pinpoint these genes' specific temporal and spatial expression patterns. The next steps involve generating epigenomics data, including DNA methylation, ATAC-seq, histone modifications, and transcription factor binding sites. Future research will focus on constructing gene regulatory networks using multi-omics approaches.

Materials and methods

Ethics statement

All animal procedures were conducted under the approval of the Beltsville Agricultural Research Center (BARC) Institutional Animal Care Protocol Number 18–005.

Animal collection and tissue preparation

The research dairy herd at Beltsville Agricultural Research Center, Agricultural Research Service, US Department of Agriculture (USDA, ARS, BARC) is representative of the U.S. Holstein population and serves as a great model for this work. Briefly, second (n = 2) and third (n = 3) lactation cows were surgically fitted with a duodenal sampling cannula (general anesthesia) during the dry period prior to initiation of sampling. Following at least a two-week recovery, the rumen fistula (local anesthesia) was placed at least 28 days prior to the expected calving date. Following parturition, a total of 108 sample biopsies of three gastrointestinal tissues (colon, duodenum, and rumen) across eight lactation stages (D3, 14, 28, 45, 120, 220, 305, and Dry), with five replicates of each stage (two cows were not sampled at days 14 and 28; Table S1). Rumen epithelial tissue (papillae) was obtained using grab biopsies

without total rumen evacuation. Duodenal fiber optic endoscopic biopsies will be obtained using sterile biopsy forceps aided by a Pentax EC-383IL camera inserted through the duodenal cannula. Colonic tissue biopsies were obtained using sterile biopsy forceps aided by a Pentax EC-383IL camera inserted through the anus. After isolating the three gastrointestinal tissues (colon, duodenum, and rumen) as described, the tissue samples were serially rinsed in saline solutions before being snap-frozen in liquid nitrogen and stored at -80°C for future use.

Library construction and RNA profiling

All tissue samples were processed by a commercial service provider, Admera Health LLC (South Plainfield, NJ), for RNA isolation, Quality control, library construction, and sequencing. Briefly, RNA was isolated using Qiagen RNeasy Plus Mini Kit (Cat No./ID: 74134, Qiagen). The quality of RNA was checked using TapeStation RNA HS Assay (Agilent Technologies, CA, USA). To ensure the RNA quality, RNA integrity numbers (RINs) are set at a minimum of 7.5. and RNA samples are quantified by Qubit RNA HS assay (ThermoFisher). Ribosomal RNA depletion was performed with a Ribo-zero Magnetic Gold Kit (Catalog number MRZG12324, Illumina Inc., San Diego, CA).

RNA-SEQ library construction and sequencing

RNA samples are randomly primed and fragmented based on the manufacturer's recommendation (NEBNext[®] Ultra[™] RNA Library Prep Kit for Illumina[®]). The first strand is synthesized with the Protoscript II Reverse Transcriptase with a longer extension period (40 min for 42°C). All remaining steps for library construction were used according to the NEBNext[®] Ultra[™] RNA Library Prep Kit for Illumina[®]. Illumina 8-nt dual indices were used.

Sequencing was performed in paired-end mode (2×150 bp reads) on the Illumina HiSeq 2500 sequencing platform (Illumina, San Diego, CA, USA). The raw sequencing data generated in this study have been submitted to the NCBI SRA database [<https://identifiers.org/ncbi/insdc.sra:SRP441033>].

Gene expression analysis

We removed adaptors and discarded poor-quality reads using Trimmomatic (v0.39) [35] with parameters: TruSeq3-PE.fa:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:36. After building the ARS-UCD1.2 [36] reference genome index with HISAT2-build software, we mapped the clean reads to the reference genome using HISAT2 (v2.2.1) [37]. We utilized the samtools (v1.12) [38] to categorize and convert the SAM files to BAM format and index BAM files. We assembled transcripts and genes for each sample and merged them into a unique transcriptome by Stringtie (v2.2.1) [39]. We extracted read counts of genes by featureCounts (v2.0.3) [40] and obtained their normalized expression (i.e., TPM) using StringTie (v2.2.1) [39].

We assessed the quality of the sequenced libraries by visualizing the variance in gene expression among samples using principal components analysis (PCA) with distance = $(1-r)$, where r is Spearman's correlation coefficient of gene expression. We used DESeq2 (v1.30.0) [41] for the identification of DEGs between all pairwise comparisons and required the adjusted P value to be ≤ 0.05 and the absolute log2 fold change value to be ≥ 0.1 .

Stage and Tissue specific gene detection

To identify genes with specifically high expression in a stage, we conducted differential gene expression (Z-score of TPM) analysis by comparing the target stage with the other stage using limma

(v3.46.0) R package [42]. We considered genes with P value ≤ 0.05 and absolute LogFC > 1.5 as differentially expressed between stages.

To investigate the specificity of gene expression in the different tissues, we employed tspx [43] to calculate the TAU metric. TAU ranges from 0 to 1, where 0 indicates broadly expressed, and 1 is specific [44]. To calculate the tissue specificity, we adopted the maximum expression observed among different stages in each tissue.

Clustering of RNA-seq data

To explore genes with dynamic temporal expression profiles, we performed the time-series analysis using maSigPro (v1.62.0) [28], a package designed for transcriptomic time course analysis in R (v4.0.3) (R Core Team 2018). We analyzed each tissue using a degree of seven and maSigPro functions. We tested different numbers of clusters (k) by using the "see.genes (get\$genes, $k = \dots$)" command to compare the clusters for each step of k with the previous ones to obtain a robust cluster. The best cluster for each tissue was $k = 4$. We plotted the median values of genes using ggplot2 in R.

To investigate the shared expression clusters among three tissues, we utilized TimesVector (v1.5) [30] to anchor similarly expressed patterns (SEP). The value of K , the number of clusters targeted for detection, is evaluated using the following equation: $K = -85.71 + 28.57x$, where x is the product of the number of tissues and time points. We adjusted the K -value according to the data characteristics as suggested by the developers. TimesVector was applied to tissue groups exhibiting similar expression patterns at $K = 600$.

Weighted gene co-expression network analysis (WGCNA)

We constructed co-expression networks for each of the three tissues based on TPM data using the R package WGCNA (v1.71) [29]. We involved 22,636 (colon), 22,600 (duodenum), and 20,797 (rumen) genes with TPM > 0.1 for further analysis. We calculated the Pearson correlation matrix between all gene pairs, which was transformed into an adjacency matrix. We performed the cluster analysis with the hclust function using the average agglomeration method and determined the soft threshold (β) with a scale-free distribution. We then constructed the gene network and detected modules using the one-step and dynamic hybrid cutting methods with options minModuleSize 30 and mergeCutHeight 0.25. We defined the eigengenes of these modules as the first principal component of the corresponding expression matrix, which was then associated with all eight lactation stages. We considered modules with correlation coefficients greater than 0.5 as the candidate modules.

Gene Ontology (GO) and cis-motif enrichment analysis

We performed GO term enrichment in stage/tissue-specific genes, cluster genes, and module genes with KOBAS (v3.0) [45] and considered GO terms with a P value lower than 0.05 significantly enriched. For cis-motif identification and enrichment analysis, we first obtained Fasta sequences of target genes based on their coordinates via the bedtools (v2.30.0). We carried out the motif enrichment analysis using the "findMotif.pl" program in the HOMER suite (<https://homer.ucsd.edu/homer/motif/>), considering the whole genome as background. We used the FDR methods as adjusted P-values for multiple testing.

Functional enrichment

To explore the stage/tissue-specificity, we downloaded 15 chromatin states predicted in cattle rumen tissue [31] and cattle QTLs from Animal QTLdb (release 48, Aug. 24, 2022) [32]. We conducted the chromatin state/QTL enrichment analysis of tissue-specific genes as described in ChromHMM (v1.22) [46]: $(C/A)/(B/D)$, where A is set as the number of bases in the chromatin state/QTL, B is set as the number of bases in stage/tissue-specific genes, C is set as the number of bases in both the chromatin state/QTL and stage/tissue-specific genes, and D is set as the number of bases in the entire genome. We calculated the statistical significance of enrichment using the Chi-squared test.

Cell type deconvolution analysis

We obtained 23 cell types in the rumen tissue from the single-cell RNA-Seq data [34] and then applied CIBERSORTx [47] to estimate the fraction of these cell types in bulk RNA-Seq samples from rumen tissues. We extracted 150 cells from each cell cluster using the subset function, implemented in Seurat (v 3.0.2) [48], to create a signature matrix using the CIBERSORTx [47] online tool by the custom option with default parameters. We then uploaded the gene expression (TPM) matrix of bulk RNA-Seq samples as the mixture file. We imputed cell fractions based on the signature and mixture files by running the Impute Cell Fractions analysis with the custom mode. We used the permutation test (100 times) to determine the significance level.

Declarations

Ethics approval and consent to participate

All animal procedures were conducted under the approval of the Beltsville Agricultural Research Center (BARC) Institutional Animal Care Protocol Number 18-005.

Consent for publication

Not applicable.

Availability of data and materials

All RNA sequencing data were submitted to NCBI, SRA database (SUB3040669, BioProject ID: PRJNA658627). All other newly generated sequencing data were submitted to NCBI, SRA database (SUB8420017, BioProject ID: PRJNA672996). The reference genome and gene annotation files (including all the sequence ontology, orthologues genes among mammals, and evolutionarily conserved regions) of ARS-UCD1.2 were downloaded from Ensembl v105 [49]. The Cattle QTLdb (release 48, Aug. 24, 2022) was obtained from [50]. The selection signatures in cattle were obtained from [51].

Authors' contributions

RLB, CJL, and GEL conceived and designed the experiments. RLB and CJL collected samples and/or generated data. YG, LF, and LM performed computational and statistical analyses. YG, CJL, GEL, and RLB wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported in part by AFRI grant numbers 2013-67015-20951, 2016-67015-24886, 2019-67015-29321, 2020-67015-02848, and 2021-67015-33409 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome and Reproduction Programs and BARD grant number US-4997-17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund. GEL was supported by appropriated project 8042-31000-112-00-D, "Accelerating Genetic Improvement of Rumi-

nants Through Enhanced Genome Assembly, Annotation, and Selection" of the Agricultural Research Service of the United States Department of Agriculture. RLB and CJL were supported by appropriated project 8042-31310-114-00-D, "Improving Dairy Cow Feed Efficiency and Environmental Sustainability Using Genomics and Novel Technologies to Identify Physiological Contributions and Adaptations."

Acknowledgements

We thank Reuben Anderson, Mary Bowman, Donald Carbaugh, Christina Clover, Cecelia Niland, and Sara McQueeney for technical assistance and sample collection. We thank the Council on Dairy Cattle Breeding for genotype, phenotype, and pedigree data, Inter-bull for global trait evaluations, and the anonymous reviewers for many helpful comments. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture (USDA). The USDA is an equal opportunity provider and employer. Cartoons in Graphical Abstract and Fig. 1A were created with BioRender.com.

Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jare.2024.06.020>.

References

- [1] Gilbert M, Nicolas G, Cinardi G, Van Boeckel TP, Vanwambeke SO, Wint GRW, et al. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Sci Data* 2018;5:180227.
- [2] Caroli AM, Chessa S, Erhardt GJ. Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. *J Dairy Sci* 2009;92(11):5335–52.
- [3] Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic Selection in Dairy Cattle: The USDA Experience. *Annu Rev Anim Biosci* 2017;5:309–27.
- [4] Strucken EM, Laurenson YC, Brockmann GA. Go with the flow-biology and genetics of the lactation cycle. *Front Genet* 2015;6:118.
- [5] Looor J, Bionaz M, Invernizzi G. Systems biology and animal nutrition: insights from the dairy cow during growth and the lactation cycle. In 2011::215–45.
- [6] Council NR. Nutrient Requirements of Dairy Cattle: Seventh Revised Edition, 2001. Washington, DC: The National Academies Press; 2001.
- [7] NRC: **Nutrient Requirements of Dairy Cattle**, 7th rev. ed. edn. Washington DC: Natl. Acad. Sci.; 2001.
- [8] Grunberg W, Staufenbiel R, Constable PD, Dann HM, Morin DE, Drackley JK. Liver phosphorus content in Holstein-Friesian cows during the transition period. *J Dairy Sci* 2009;92(5):2106–17.
- [9] Moran B, Cummins SB, Creevey CJ, Butler ST. Transcriptomics of liver and muscle in Holstein cows genetically divergent for fertility highlight differences in nutrient partitioning and inflammation processes. *BMC Genomics* 2016;17(1):603.
- [10] Pascottini OB, De Koster J, Van Nieuwerburgh F, Van Poucke M, Peelman L, Fievez V, et al. Effect of overconditioning on the hepatic global gene expression pattern of dairy cows at the end of pregnancy. *J Dairy Sci* 2021;104(7):8152–63.
- [11] Veshkini A, H MH, Vogel L, Delosiore M, Viala D, Dejean S, Troscher A, Ceciliani F, Sauerwein H, Bonnet M: Liver proteome profiling in dairy cows during the transition from gestation to lactation: effects of supplementation with essential fatty acids and conjugated linoleic acids as explored by PLS-DA. *J Proteomics* 2022;252:104436.
- [12] Accorsi PA, Pacioni B, Pezzi C, Forni M, Flint DJ, Seren E. Role of prolactin, growth hormone and insulin-like growth factor 1 in mammary gland involution in the dairy cow. *J Dairy Sci* 2002;85(3):507–13.
- [13] Annen EL, Fitzgerald AC, Gentry PC, McGuire MA, Capuco AV, Baumgard LH, et al. Effect of continuous milking and bovine somatotropin supplementation on mammary epithelial cell turnover. *J Dairy Sci* 2007;90(1):165–83.
- [14] Bernier-Dodier P, Girard CL, Talbot BG, Lacasse P. Effect of dry period management on mammary gland function and its endocrine regulation in dairy cows. *J Dairy Sci* 2011;94(10):4922–36.
- [15] Watanabe A, Hata E, Sláma P, Kimura K, Hirai T. Characteristics of mammary secretions from Holstein cows at approximately 10 days before parturition: with or without intramammary infection. *J Appl Anim Res* 2017;46(1):604–8.
- [16] Zhao X, Ponchon B, Lanctot S, Lacasse P. Invited review: accelerating mammary gland involution after drying-off in dairy cattle. *J Dairy Sci* 2019;102(8):6701–17.

- [17] Bach A, Guasch I, Elcoco G, Chaucheyras-Durand F, Castex M, Fabregas F, et al. Changes in gene expression in the rumen and colon epithelia during the dry period through lactation of dairy cows and effects of live yeast supplementation. *J Dairy Sci* 2018;101(3):2631–40.
- [18] Aschenbach JR, Zebeli Q, Patra AK, Greco G, Amasheh S, Penner GB. Symposium review: the importance of the ruminal epithelial barrier for a healthy and productive cow. *J Dairy Sci* 2019;102(2):1866–82.
- [19] Li CJ, Lin S, Ranilla-Garcia MJ, Baldwin RL. Transcriptomic profiling of duodenal epithelium reveals temporally dynamic impacts of direct duodenal starch-infusion during dry period of dairy cattle. *Front Vet Sci* 2019;6:214.
- [20] Johnson DE, Johnson KA, Baldwin RL. Changes in liver and gastrointestinal tract energy demands in response to physiological workload in ruminants. *J Nutr* 1990;120(6):649–55.
- [21] Gross JJ. Limiting factors for milk production in dairy cows: perspectives from physiology and nutrition. *J Anim Sci* 2022;100(3).
- [22] Cardoso-Moreira M, Halbert J, Vallotton D, Velten B, Chen C, Shao Y, et al. Gene expression across mammalian organ development. *Nature* 2019;571(7766):505–9.
- [23] Rahmanian S, Murad R, Breschi A, Zeng W, Mackiewicz M, Williams B, et al. Dynamics of microRNA expression during mouse prenatal development. *Genome Res* 2019;29(11):1900–9.
- [24] White RJ, Collins JE, Sealy IM, Wali N, Dooley CM, Digby Z, et al. A high-resolution mRNA expression time course of embryonic development in zebrafish. *Elife* 2017;6.
- [25] Zhao B, Luo H, He J, Huang X, Chen S, Fu X, et al. Comprehensive transcriptome and methylome analysis delineates the biological basis of hair follicle development and wool-related traits in Merino sheep. *BMC Biol* 2021;19(1):197.
- [26] Shi L, Li H, Huang X, Shu Z, Li J, Wang L, Yan H, Wang L. Integrated analysis of transcriptome and metabolome revealed biological basis of sows from estrus to lactation. *iScience* 2023, 26(1).
- [27] Zhang T, Wang T, Niu Q, Xu L, Chen Y, Gao X, et al. Transcriptional atlas analysis from multiple tissues reveals the expression specificity patterns in beef cattle. *BMC Biol* 2022;20(1):79.
- [28] Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* 2014;30(18):2598–602.
- [29] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* 2008;9:559.
- [30] Jung I, Jo K, Kang H, Ahn H, Yu Y, Kim S. TimesVector: a vectorized clustering approach to the analysis of time series transcriptome data from multiple phenotypes. *Bioinformatics* 2017;33(23):3827–35.
- [31] Gao Y, Liu S, Baldwin VI RL, Connor EE, Cole JB, Ma L, et al. Functional annotation of regulatory elements in cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning. *Genomics* 2022;114(2):110296.
- [32] Hu ZL, Park CA, Reecy JM. Bringing the animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res* 2022;50(D1):D956–61.
- [33] Drackley JK. ADSA foundation scholar award. Biology of dairy cows during the transition period: the final frontier? *J Dairy Sci* 1999;82(11):2259–73.
- [34] Wu JJ, Zhu S, Gu F, Valencak TG, Liu JX, Sun HZ. Cross-tissue single-cell transcriptomic landscape reveals the key cell subtypes and their potential roles in the nutrient absorption and metabolism in dairy cattle. *J Adv Res* 2022;37:1–18.
- [35] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30(15):2114–20.
- [36] Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience* 2020;9(3).
- [37] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37(8):907–15.
- [38] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. Genome project data processing S: the sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25(16):2078–9.
- [39] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 2015;33(3):290–5.
- [40] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;30(7):923–30.
- [41] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15(12):550.
- [42] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43(7):e47.
- [43] Camargo AP, Vasconcelos AA, Fiamenghi MB, Pereira GAG, Carazzolle MF. Tspex : a Tissue-Specificity Calculator for Gene Expression Data. 2020:1–7.
- [44] Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* 2017;18(2):205–14.
- [45] Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res* 2021;49(W1):W317–25.
- [46] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 2012;9(3):215–6.
- [47] Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;37(7):773–82.
- [48] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck 3rd WM, et al. Comprehensive integration of single-cell data. *Cell* 2019;177(7):1888–1902 e1821.
- [49] Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P et al. Ensembl variation resources. *Database (Oxford)* 2018; 2018.
- [50] Hu ZL, Park CA, Wu XL, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Res* 2013;41.
- [51] Chen N, Fu W, Zhao J, Shen J, Chen Q, Zheng Z, et al. BGVD: an integrated database for bovine sequencing variations and selective signatures. *Genom Proteom Bioinform* 2020.

PigBiobank: a valuable resource for understanding genetic and biological mechanisms of diverse complex traits in pigs

Haonan Zeng^{1,†}, Wenjing Zhang^{1,†}, Qing Lin^{1,†}, Yahui Gao^{1,†}, Jinyan Teng¹, Zhiting Xu¹, Xiaodian Cai¹, Zhanming Zhong¹, Jun Wu¹, Yuqiang Liu¹, Shuqi Diao¹, Chen Wei¹, Wentao Gong¹, Xiangchun Pan¹, Zedong Li¹, Xiaoyu Huang¹, Xifan Chen¹, Jinshi Du¹, The PigGTEx Consortium, Fuping Zhao², Yunxiang Zhao³, Maria Ballester⁴, Daniel Crespo-Piazuelo⁴, Marcel Amills^{5,6}, Alex Clop^{5,7}, Peter Karlskov-Mortensen⁸, Merete Fredholm⁸, Pinghua Li^{9,10}, Ruihua Huang^{9,10}, Guoqing Tang¹¹, Mingzhou Li¹¹, Xiaohong Liu¹², Yaosheng Chen¹², Qin Zhang¹³, Jiaqi Li¹, Xiaolong Yuan¹, Xiangdong Ding^{14,*}, Lingzhao Fang^{15,*} and Zhe Zhang^{1,*}

¹State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

²Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing 100193, China

³College of Animal Science and Technology, Guangxi University, Nanning 530004, China

⁴Animal Breeding and Genetics Programme, Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Torre Marimon, Caldes de Montbui, Spain

⁵Department of Animal Genetics, Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus de la Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

⁶Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain

⁷Consejo Superior de Investigaciones Científicas, Barcelona, Catalonia, Spain

⁸Animal Genetics, Bioinformatics and Breeding, Department of Veterinary and Animal Sciences, University of Copenhagen, Frederiksberg C1870, Denmark

⁹Institute of Swine Science, Nanjing Agricultural University, Nanjing 210095, China

¹⁰Key Laboratory in Nanjing for Evaluation and Utilization of Livestock and Poultry (Pigs) Resources, Ministry of Agriculture and Rural Areas, China, Nanjing 210095, China

¹¹State Key Laboratory of Swine and Poultry Breeding Industry, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China

¹²State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China

¹³College of Animal Science and Technology, Shandong Agricultural University, Tai'an 271018, China

¹⁴College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

¹⁵Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark

*To whom correspondence should be addressed. Tel +86 20 85282019; Email: zhezhang@scau.edu.cn

Correspondence may be also addressed to Lingzhao Fang. Tel +45 89991301; Email: lingzhao.fang@qgg.au.dk

Correspondence may be also addressed to Xiangdong Ding. Tel +86 10 62734277; Email: xding@cau.edu.cn

[†]The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

Abstract

To fully unlock the potential of pigs as both agricultural species for animal-based protein food and biomedical models for human biology and disease, a comprehensive understanding of molecular and cellular mechanisms underlying various complex phenotypes in pigs and how the findings can be translated to other species, especially humans, are urgently needed. Here, within the Farm animal Genotype-Tissue Expression (FarmGTEx) project, we build the PigBiobank (<http://pigbiobank.farmgtex.org>) to systematically investigate the relationships among genomic variants, regulatory elements, genes, molecular networks, tissues and complex traits in pigs. This first version of the PigBiobank curates 71 885 pigs with both genotypes and phenotypes from over 100 pig breeds worldwide, covering 264 distinct complex traits. The PigBiobank has the following functions: (i) imputed sequence-based genotype-phenotype associations via a standardized and uniform pipeline, (ii) molecular and cellular mechanisms underlying trait-associations via integrating multi-omics data, (iii) cross-species gene mapping of complex traits via transcriptome-wide association studies, and (iv) high-quality results display and visualization. The PigBiobank will be updated timely with the development of the FarmGTEx-PigGTEx project, serving as an open-access and easy-to-use resource for genetically and biologically dissecting complex traits in pigs and translating the findings to other species.

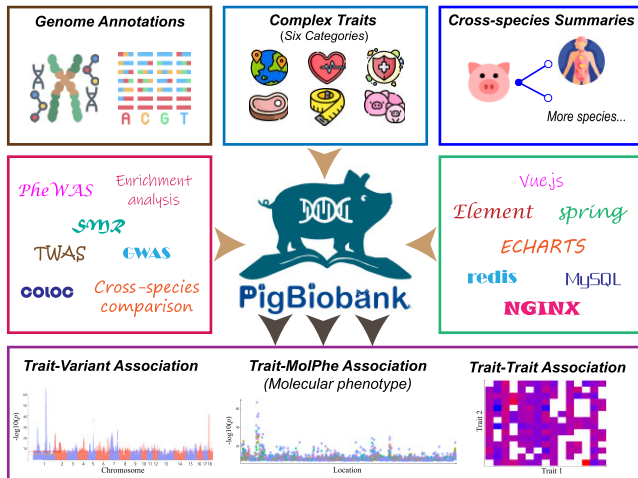
Received: August 14, 2023. Revised: October 13, 2023. Editorial Decision: October 24, 2023. Accepted: October 27, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License

(<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Graphical abstract



Introduction

The immense value of pigs in the realms of agriculture and human biomedicine has been widely acknowledged. Pork is the second-largest meat source worldwide and plays an important role in ending the global hunger and malnutrition crisis, as reported by the Food and Agriculture Organization of the United Nations (FAO) (1). Additionally, pigs are considered to be more similar to humans in terms of anatomical characteristics, physiology, immunology, and genome than other model organisms such as rodents (2). Lunney *et al.* (3) comprehensively reviewed the recent applications of pigs as human biomedical models in reproduction and fetal development, brain and neurodegenerative disease, and xenotransplantation, shedding light on the myriad advantages that pigs can offer in biomedical research (4–6). Hence, it becomes crucial to thoroughly explore and elucidate the underlying molecular mechanisms responsible for diverse complex traits in pigs, which will substantially contribute to both expediting the genetic gain via selective breeding and establishing the pig as an invaluable biomedical model.

Over the past decades, genome-wide association study (GWAS) has proven to be a valuable tool for understanding the genetic basis of complex traits and diseases. However, compared with large-scale public GWAS summary statistics of various complex traits in humans like GWAS Catalog (7), GWASdb (8), GWAS ATLAS (9), Brain Catalog (10) and Open Target (11), GWASs of complex traits in pigs are typically performed using SNP arrays in small populations with highly related individuals, where the linkage disequilibrium (LD) among SNPs is high. What is even worse: the access to full GWAS summary statistics in the public domain is limited, posing a hindrance to the aggregation of multiple independent GWAS datasets for performing meta-analyses. Moreover, the majority of these GWAS loci are primarily located in non-coding regions of the genome (12,13), implying that they exert their effects on complex traits via altering gene regulation and expression (14–16). Consequently, the underlying biological mechanisms that explain the effects of these non-coding variants on complex traits and diseases are still not well understood. Although the AnimalQTLdb (17) curated tens of thousands of QTLs of 279 traits from 800 publications, it lacks full

summary statistics of each individual GWAS. Recently, Teng *et al.* (14) released the Pig Genotype-Tissue Expression (Pig-GTEx) resource, a highly valuable catalog of regulatory variants across multiple pig tissues. It provides an extensive collection of millions of molecular QTLs associated with five distinct types of molecular phenotypes (i.e. protein coding gene expression, exon expression, lncRNA expression, enhancer expression and alternative splicing) derived from 34 different pig tissues. The advent and future development of Pig-GTEx will greatly enhance our understanding of the regulatory mechanisms underlying complex traits in pigs and it will also facilitate cross-species gene mapping via transcriptome-wide association studies (TWAS) such as pigs vs. humans (14), chickens vs. humans (18), and rats vs. humans (19).

Here, to fully unleash the potential of pigs as an agricultural species and for human biomedical applications, we constructed the PigBiobank database (<http://pigbiobank.farmgtex.org>) by integrating large-scale imputed sequence-based GWAS of 264 distinct complex traits with the PigGTEx resource. To ensure the reliability and credibility of our data and findings, we implemented rigorous quality control measures and a standardized analytical pipeline throughout the entire process, from raw data collection to data analysis within the PigBiobank. It will serve as the most extensive resource for discerning candidate causal variants, genes, molecular networks, and tissues underlying complex traits in pigs. In addition, the PigBiobank not only allows the users to explore the genetic relationships between pig traits, but also establish connections with traits in other species (e.g. humans) via TWAS to uncover the shared genetic basis of homologous traits (e.g. human body weight vs. pig body weight) across species.

Materials and methods

Data collection and pre-processing

GWAS resources and pre-processing

The original data of PigBiobank is defined as a set of GWAS summary statistics generated from a large-scale meta-analysis of GWAS (metaGWAS, dataset 1) (20), and GWAS

on the eigenvector composition (eigenGWAS) and environmental phenotypes (envGWAS) using PGRP from PigGTEx project (14) (dataset 2). A total of 71 885 pigs with both genotypes and phenotypes from over 100 pig breeds, containing up to 264 complex traits that are classified into six main trait categories (17) (Supplementary Table S1, S2). To obtain accurate and valuable information, a standardized protocol was set up to deal with such complex datasets.

Dataset 1 consists of 70328 pigs from 59 populations in 14 pig breeds (20). SNPs genotyped with six types of SNP array from previous pig reference genome version were lifted over to the current Sscrofa11.1 (Ensembl v100) using R (v.3.6). Only the successfully mapped autosomal bi-allelic SNPs were retained. Afterwards, the SNP array data of each population was imputed to the sequence level via Beagle (v5.1) (21) using the Pig Genomics Reference Panel (PGRP) version 1 in PigGTEx project (14), which consists of 1602 WGS samples and 42 523 218 SNPs. For primary phenotypes, we manually removed phenotypic outliers and ensured consistency of the same trait among populations by the phenotypic descriptive statistics using R (v3.6) (<https://www.R-project.org>). Finally, a total of 232 complex traits were included in the subsequent analyses after quality control (QC), which were classified into five main trait categories (i.e. Reproduction, Meat and Carcass, Production, Health, Exterior).

Dataset 2 consisting of 1557 pigs from over 100 pig breeds was extracted from PGRP v1 (14). To elucidate the evolutionary patterns of environmental adaptability, a total of 30 environment-related phenotypes for envGWAS were obtained from the WorldSIM 2 (22), the High-resolution gridded datasets (23), and the GLOBMAP Leaf Area Index (LAI) (24) using the information on latitude and longitude for each indigenous pig breed. Moreover, to interpret the evolution and adaptation of the characteristics, two genetic differentiation phenotypes derived from the genotypic data are used to perform GWAS with eigenvector composition (eigenGWAS) within Asian-European pig populations and within Asian south-north domestic pig populations. In total, dataset 2 encompasses 32 traits that are further classified under the main trait category of 'Adaptation'.

Cis-molQTLs, chromatin states, and phastCons score resources. We downloaded 15 chromatin states from 14 pig tissues (25) and conservation scores (phastCons) from UCSC (<http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.100way.phastCons/>) (phastCons100way). Moreover, we generated five different types of cis-molecular quantitative trait locus (cis-molQTL, <1 Mb to the TSS of genes) data involved in 34 tissues (i.e. tissues, and organ systems) from the PigGTEx project (14). To be specific, it included cis-eQTL for protein coding gene expression, cis-eeQTL for exon expression, cis-lncQTL for lncRNA expression, cis-enQTL for enhancer expression and cis-sQTL for alternative splicing.

Summary-based analysis

SNP-level and gene-level association analysis

For each of the populations in dataset 1, we performed GWAS on binary and continuous traits with fastGWA-GLMM and fastGWA models implemented in GCTA (v1.94.0beta) (26–28), respectively. For reproductive traits (with repeated recordings), association analysis was conducted using MMAP

(released on 2021–08-19) (<https://mmap.github.io/>) with de-regressed estimated breeding value (dEBV) as the phenotype. The metaGWAS was performed using the inverse variance-weighted fixed effects model in METAL (released on 2011-03-25) (29) after QC for GWAS on the basis of sample size, effect size and standard error. Ultimately, the 232 complex traits comprised of 268 studies were further classified into five main trait categories (i.e. Reproduction, Meat/Carcass, Production, Health, and Exterior). For more details of processing raw individual data, we referred to Xu *et al.* (20). For dataset 2, GWAS analysis of environment-related phenotypes was conducted with GEMMA (v0.98.5) (30), including five genotypic principal components as covariates. In addition, we conducted eigenGWAS for two genetic differentiation phenotypes using default parameters in GEAR (v0.919) (31). The analysis yielded 32 summary statistics falling under the category of 'Adaptation'. Lastly, leveraging SNP-level GWAS, we conducted the gene-level association analysis using MAGMA (v1.0) (32).

Estimation of heritability and genetic correlations of complex traits

We conducted the SNP heritability estimation for all the GWAS summary statistics by LDSC (v1.0.1) (33) with default parameters. In order to build the landscape of genetic correlations of pig complex traits, we used LDSC (v1.0.1) (33) to calculate the global genetic correlation for each pair of the aforementioned GWAS studies. By default, linkage disequilibrium (LD) pattern was calculated from biallelic SNPs in the reference panel PGRP v1. We also tested whether genetic correlations significantly deviated from 0 with the chi-square test ($df = 1$) using the Wald statistic.

Integrative analysis

TWAS, colocalization and SMR analysis

We performed gene prioritization through integrating the GWAS with five cis-molQTLs from 34 tissues using the following three complementary strategies, colocalization (COLOC), summary-based mendelian randomization (SMR), and transcriptome-wide association study (TWAS). To explore whether a genetic variant affects both the intermediate molecular phenotypes and the complex trait of interest, we used fastENLOC (v1.0) (34) to quantify the regional colocalization probability (RCP) for each independent molQTL signal clusters and GWAS hits, and considered a gene to be significant if its RCP ≥ 0.9 in the COLOC analysis. TWAS was to test associations between a complex trait of interest and genetically predicted gene expression levels. We applied S-PrediXcan (35) in single tissues and S-MultiXcan (36) in multiple tissues to detect transcriptionally regulated genes underlying complex traits with a stringent Bonferroni multiple-testing correction. Furthermore, SMR was performed to identify molecular phenotypes that are associated with a complex trait because of a shared candidate causal variant (i.e. pleiotropy or causality) (37). A Benjamini–Hochberg method correction ($FDR < 0.05$) was used in each SMR analysis and the heterogeneity in dependent instruments (HEIDI) test was applied to distinguish pleiotropy from linkage with a threshold of 0.05 ($HEIDI > 0.05$) (37).

Enrichment analysis

Identification of trait-relevant functional elements and molQTLs

We utilized SnpEff (v.4.3) (38) to annotate the biallelic SNPs in the PGRP v1 VCF file, resulting in 20 functional categories. For each functional category, the enrichment fold of molQTL was calculated by R package fmsb (v0.7.5). In addition, we utilized the stratified LD score regression (S-LDSC) (39) to comprehensively explore the heritability enrichment of GWAS with five molQTL annotations across all 34 tissues in dataset 1.

Inference for trait-relevant tissues

We applied four computational strategies to detect the tissues associated with complex traits. The first was QTLEnrich (v2) (40), as a rank- & permutation-based method, which aims to test for enrichment of trait-associations in molQTLs (e.g. eQTL, eeQTL, IncQTL, enQTL, sQTL) specific in each tissue. For each type of *cis*-molQTLs, QTLEnrich was used to test whether the molQTLs in a given tissue were significantly enriched for given traits ($P < 0.05$). Second, we extracted genomic regions by expanding 100 kb windows around the top 1000 genes that were highly expressed in each of the 34 tissues. Subsequently, we utilized BEDTools (v2.25.0) (41) to calculate the enrichment fold of these regions with trait-associations. Permutation tests with 10000 replicates were conducted to determine the P values using the R package regioneR (v1.24.0) (42). Third, we applied stratified LD score regression (S-LDSC) (43) to the above-mentioned genomic regions from 34 tissues to evaluate whether the heritability of each of the 232 traits in dataset 1 was significantly enriched in tissue-specific expressed gene regions. Lastly, we utilized BEDTools (v2.25.0) (41) to calculate the enrichment fold of the genomic region of each of 15 chromatin states in 14 tissues, and the permutation test with 10000 replicates was conducted to obtain the P values using the R package regioneR (v1.24.0) (42).

Cross-species comparison analysis

To explore the sharing patterns of the genetic architecture of complex traits between species, we further investigated the trait similarity between pigs and humans on the level of GWAS and TWAS summary statistics. On the one hand, we calculated the Pearson's correlation of the z-scores of GWAS for homologous variants between pigs (*Sus scrofa*11.1) and humans (GRCh38/hg38). On another hand, we obtained the TWAS summaries statistics from PigGTEx project (14) and calculated the Pearson's correlation of the absolute standardized effect size of TWAS for orthologous genes between pigs and humans.

Database design

The PigBiobank was designed with a decomposing framework using Vue (<https://github.com/vuejs/core>) as the front-end and Spring Boot in Java as the back-end. NGINX was used as the reverse proxy server for balancing the network load. To develop the user-friendly interface, we used Element (<https://github.com/ElementFE/element>) for beautifying the page layout, ECharts (<https://github.com/apache/echarts>) and IGV (44) for data visualization. To fit and invoke the multi-omics data, we used MySQL as an engine for both data storage and data querying.

Results

Overview of PigBiobank

Herein, the current version of PigBiobank is of six perspectives: Trait, Resource, Biology, Analysis, Search and Download (Figure 1). With each perspective, a general exploration of the database is presented. (i) Trait. PigBiobank encompasses a comprehensive collection of 264 complex pig traits, which are meticulously classified into six main trait categories, including adaptation, exterior, health, meat and carcass, production, and reproduction (17). (ii) Resource. PigBiobank systematically collects, processes and consolidates the data resource from multiple databases, i.e. the PigGTEx-portal, Functional Annotation of Animal Genomes (FAANG) project, UK Biobank, Human GTEx project, Ensembl, UCSC and The National Center for Biotechnology Information (NCBI). (iii) Biology. PigBiobank integrates the aforementioned large-scale multi-omics data to facilitate users to effortlessly explore the regulatory mechanisms underlying various complex traits in pigs. The platform also strives to establish connections with traits in other species (currently only humans) to unveil the shared genetic basis of homologous traits. (iv) Analysis. PigBiobank provides comprehensive features collected from multi-layer analyses, including trait-variant association, trait-molecular QTL association, and trait-trait association. (v) Search. Users can utilize the user-friendly quick search function to explore specific traits, genomic variants, genes, or genomic regions of interest by entering relevant keywords. They can also jump seamlessly from one page to another to explore connections and interactions among traits, genes and genomic variants. (vi) Download. Users can intuitively visualize and freely download all the results of association analysis or query data from the PigBiobank.

Web interface and usage

We developed a user-friendly interface allowing users to access all the information from any device and location by means of searching, browsing, visualizing and downloading. The current version of the PigBiobank mainly contains six menus, namely Home, Trait Browser, Module, Download, Contact and Help (Figure 2A). The homepage provides the basic summary of the database and a search box on a very prominent position. Users can utilize the quick search function in search of traits, genomic variants, genes, or genomic regions of interest via typing relevant keywords (Figure 2A). Searching by trait is always a central point of database construction and web design, so we have made a separate overview interface for these traits in the menu of 'Trait browser'. The page displays concise summaries of 264 traits comprised of 300 studies, such as trait names, trait types, synonyms, main and sub trait categories, breeds, total sample size, total SNPs, lead SNPs, and single GWAS population. Users can click the items of interesting traits and then go into the specific web presentation for each trait that is similar to use the search function on the homepage. The menu of 'Module' provides detailed functions for point-to-point searching and analyzing such as GWAS and phenome-wide association study (PheWAS). The menu of 'Download' provides the download entry for the list of available files. The remaining menus (i.e. Contact, Help) allow users to get detailed information and documentation about the PigBiobank and convenient communications with us.

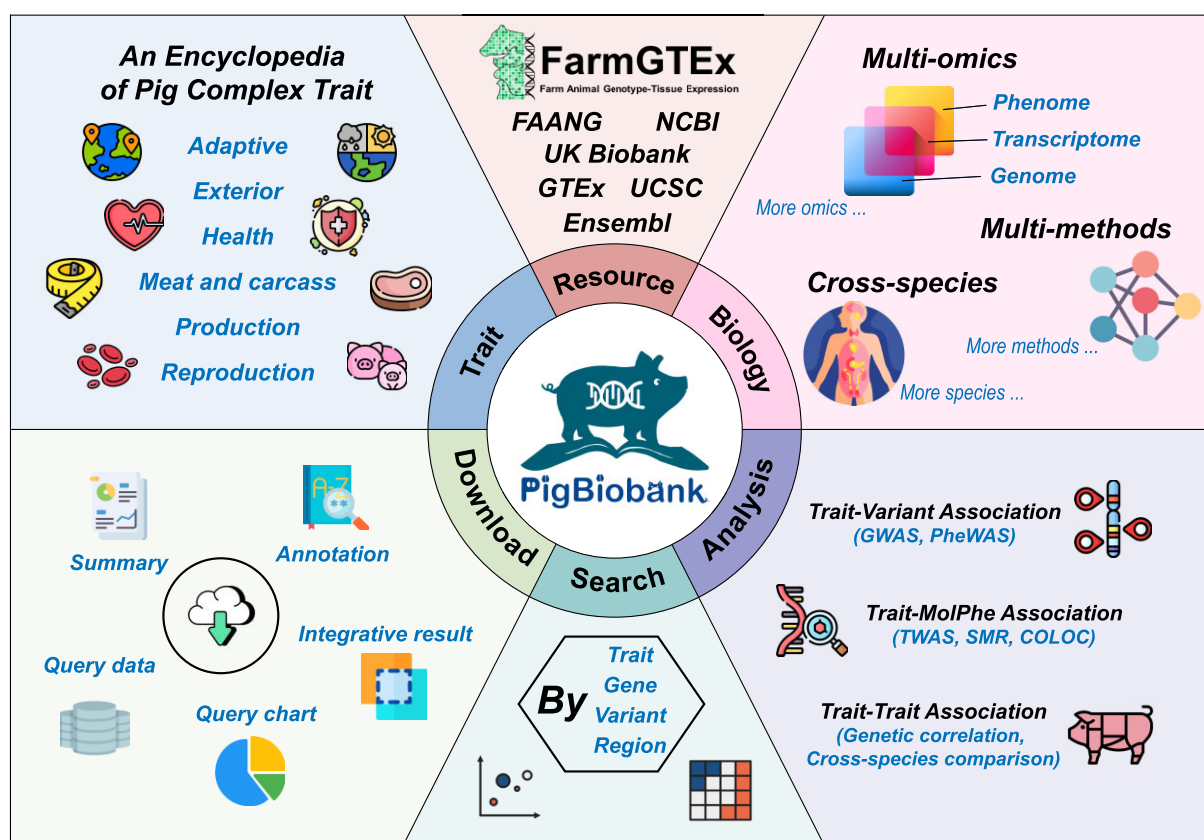


Figure 1. Schematic overview of the PigBiobank. It consists of six components. (i) Trait: 264 complex pig traits derived from six main trait categories. (ii) Resource: data collected from multiple relevant international projects. (iii) Biology: biological elucidation using multi-omics data, multi-methods analysis, and cross-species comparison. (iv) Analysis: seven online analysis modules/tools are available in the PigBiobank web portal. (v) Search: users can query the database in four different ways (by trait, gene, variant, or region) in a user-friendly manner. (vi) Download: data files or results are available for free download.

To help users elucidating the potential regulatory mechanism of complex traits in pigs, we categorize the research contents into six sections (Figure 2B), including (i) ‘Summary-based Analysis’ containing SNP-based GWAS, gene-based GWAS, genetic parameter estimation and PheWAS. (ii) ‘Content of Region’ by genomics viewer. (iii) ‘Integrative Analysis’ presented by TWAS, SMR, and COLOC. (iv) List of ‘associated Variants/Genes’ for target region or variant based on various methods. (v) ‘Enrichment Analysis’ for detecting trait-relevant functional elements and tissues. (vi) ‘Cross-species Analysis’ based on GWAS and TWAS. Moreover, we use interactive or static visualizations for the results from most analyses. For instance, as shown in Figure 2B, PigBiobank provides a variety of diagrams such as the scatter plot to show trait-associated SNPs and genes for the GWAS, PheWAS, TWAS, SMR, COLOC, and the heatmap plot is mainly utilized to visually depict the results of enrichment analysis and cross-species comparisons (i.e. human). For the genetic correlation analysis, the histogram is available for single trait search and heatmap for multiple traits search in the database. It is worth noting that PigBiobank provide all the accession of the images presented in the website.

Case study for complex trait search module

To demonstrate the utility of the PigBiobank in deciphering the molecular mechanisms behind complex traits in pigs and to showcase the functionality of the trait search mod-

ule within PigBiobank, we present the study named ‘MetaGWAS_M_BFT’ focusing on the trait of ‘Average backfat thickness’ as a case study. Upon querying the database using the keyword ‘MetaGWAS_M_BFT’, following five sections are presented: (i) Summary of Trait, (ii) Summary-based Analysis, (iii) Integrative Analysis, (iv) Enrichment Analysis and (v) Cross-species analysis.

The section titled ‘Summary of Trait’ presents a table that provides extra information about the study named ‘MetaGWAS_M_BFT’ (Figure 3A). The table includes the following information: (i) trait details, including the trait name, trait type, trait category and breeds involved; (ii) association analysis details, which consist of the number of GWASs used for meta-analysis, phenotype records, tested SNPs, genome-wide significant SNPs, lead SNPs and the analysis software used and (iii) heritability estimated along with the standard error.

The section of ‘Summary-based Analysis’ presents the result of metaGWAS, gene-based GWAS and genetic correlation (Figure 3B). As an example, the ‘SNP-based metaGWAS’ panel displays significant signals of metaGWAS for ‘MetaGWAS_M_BFT’. Among these metaGWAS signals, the top QTL (chr1: 160114440–161114440) with the lead SNP of 1_160614440_C_A ($P = 4.1620 \times 10^{-65}$) was identified. Notably, the putative gene *MC4R*, which has been supported by laboratory experiments and multiple association studies (45,46), is located within the QTL region. In the gene-based GWAS, we also find *MC4R* significantly associ-



Figure 2. The web interface of the PigBiobank. (A) Query entries for the PigBiobank database. The menu options 'Home', 'Trait Browser', and 'Module' provide access to the quick search, detailed trait browses, and online analysis modules, respectively. (B) Query visualization. A total of six sections namely 'Summary-based Analysis', 'Content of Region', 'Integrative Analysis', 'Enrichment Analysis', 'Cross-species Analysis', and 'Associated Variants/Genes' are presented in the database.

ated ($P = 1.2741 \times 10^{-11}$) with BFT. Additionally, the 'genetic correlation' panel reveal genetic correlations between the searched trait and other traits. For instance, BFT exhibit a significant genetic correlation with reproductive trait such as the number born alive (NBA), the total number of born (TNB), and the number born of healthy pigs (NBH). This observation supports that fatness has the potential to impact farrowing performance (47–49).

The 'Integrative Analysis' section presents the result of integrating regulatory variants from PigGTEx with GWAS of complex traits to identify the candidate causal genes for the searched trait (Figure 3C). The in-line panels display the plot and table showcasing significant results from TWAS, SMR and colocalization analyses between GWAS of complex traits and five types of molecular phenotypes/QTLs. For 'MetaGWAS_M_BFT', we detect that *MC4R* was a significant gene in TWAS ($P = 1.1408 \times 10^{-17}$) and SMR ($P = 1.0128 \times 10^{-6}$) with eQTL data in the frontal cortex. In addition, the expression of *ABCD4* is significantly associated with BFT in multiple tissues, especially in the small intestine ($P = 3.1800 \times 10^{-22}$ in TWAS, $P = 1.7848 \times 10^{-8}$ in SMR).

The section of 'Enrichment Analysis' presents the results of enrichment of GWAS signals of complex traits and different biological annotations (Figure 3D) including functional elements, chromatin states, molQTLs and tissues. We find that significant variants of metaGWAS in 'MetaGWAS_M_BFT' were significantly enriched in the functional element of 'CDS' (enrichment fold = 1.3778, $P = 9.9990 \times 10^{-5}$), chromatin states of 'weak active enhancer' in adipose (enrichment fold = 2.0848, $P = 9.9990 \times 10^{-4}$) and 'bivalent/poised TSS' in 12 out of 14 tissues. Both molQTL and heritabil-

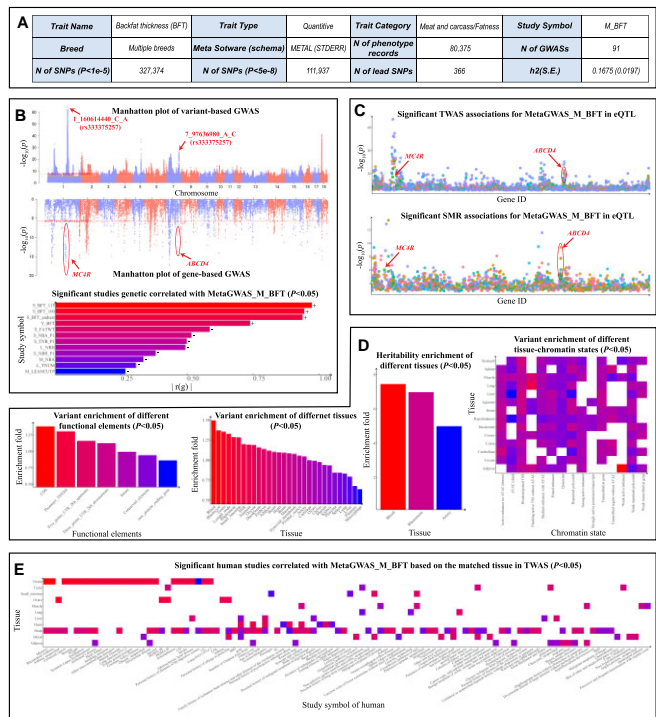


Figure 3. Case study of trait search module (MetaGWAS_M_BFT). (A) The summary information of the study. (B) The Manhattan plot of variant-based and gene-based GWAS, and the list of traits significantly genetic correlated ($P < 0.05$) with study. (C) The scatter plot of significant associations of TWAS, and SMR (corrected $P < 0.05$). (D) The bar plot and heatmap show significant results of enrichment analysis for study, containing the variant enrichment trait-relevant functional elements, variant enrichment of trait-relevant tissue, heritability enrichment of trait-relevant tissue, and variant enrichment of trait-relevant tissue-chromatin states ($P < 0.05$). (E) The heatmap of human traits significantly correlated with pig study based on TWAS ($P < 0.05$).

ity enrichment analysis showed that the tissue of 'blood' was highly relevant with BFT, with enrichment fold = 1.4952 ($P = 9.9990 \times 10^{-4}$) and 7.4395 ($P = 3.3016 \times 10^{-2}$), respectively.

The section titled 'Cross-species analysis' presents the results of comparative analyses of complex traits between humans and pigs to aid in the understanding of genetic similarities of complex traits between species. PigBiobank provides the Pearson's correlation of the absolute standardized effect sizes of homologous variants from GWAS summary. For comparison in GWAS, we obtained only eight traits of humans significantly correlated with BFT of pig at a low level (the absolute of Pearson's correlation < 0.06). By integrating PigGTEx resources with GWAS of complex traits, PigBiobank also provide the Pearson's correlation of the absolute standardized effect sizes of orthologous genes from TWAS summary for the matched tissue between species (Figure 3E). The comparison of TWAS results showed that BFT of pig was significantly correlated with 100 human traits in at least one tissue, with 14 traits in human being highly correlated ($r > 0.1500$) such as the trait of 'memory' ($r = 0.2019$, $P = 1.210 \times 10^{-3}$) and 'weight' ($r = 0.1740$, $P = 5.1600 \times 10^{-3}$).

Case study for gene and variant search module

To explore the regulatory mechanism of gene/variant towards complex traits, PigBiobank not only provide gene or vari-

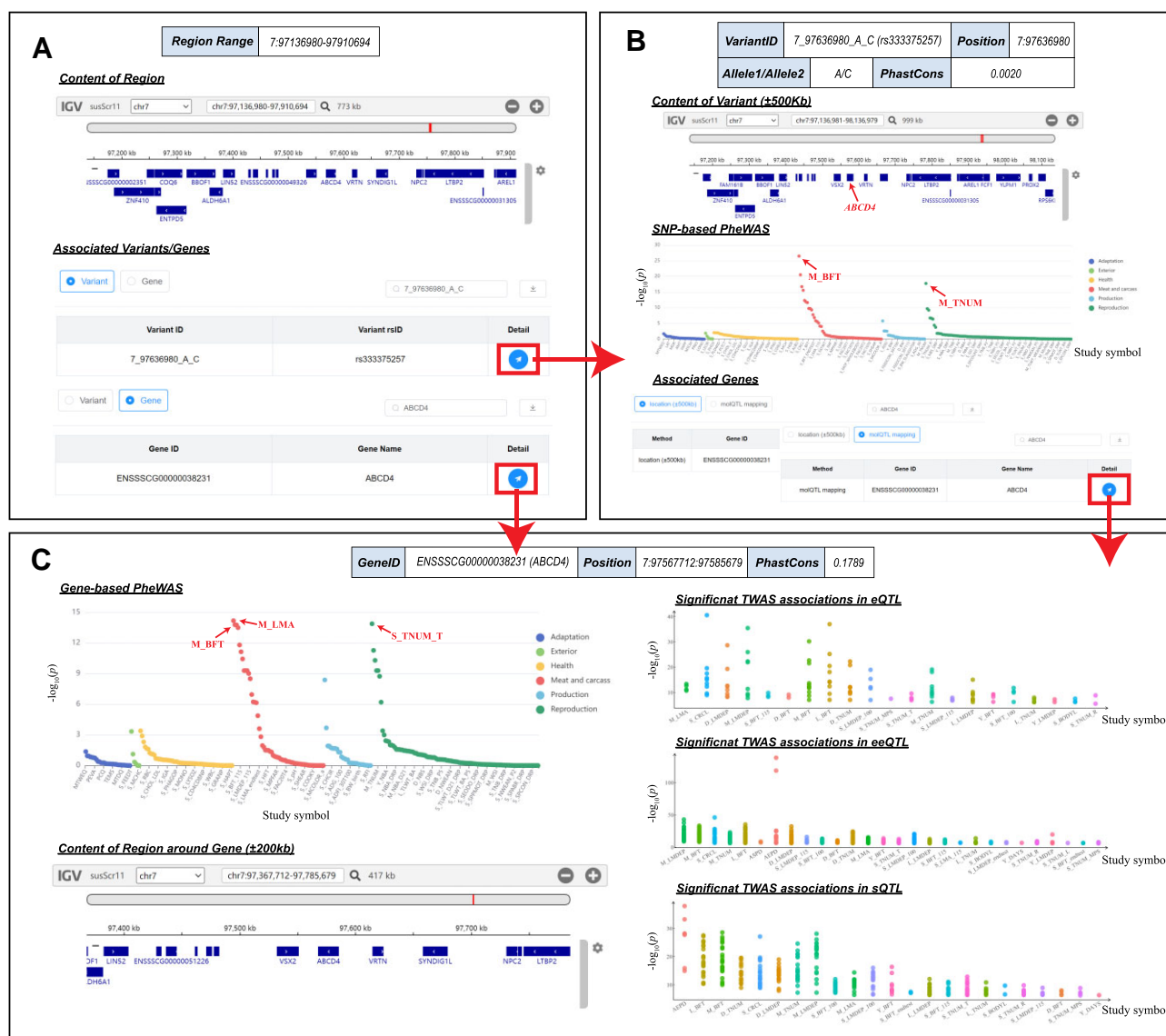


Figure 4. Case study of region search (SSC7: 97136980~97910694), variant search (rs333375257), and gene search (ABCD4). (A) The result of the region search, including the summary information of the target region, the genomics viewer of the region and the list of significant variants in GWAS within the target region, or genes overlapped within the target region. (B) The result of the variant search, containing the summary information of the target variant, the genomics viewer of the region (bp ± 500 kb), and the scatter plot of SNP-based PheWAS and the list of associated genes based on physical location (± 500 kb), or molecular QTL mapping from top to down, respectively. (C) The result of the gene search, containing the summary information of the target gene, gene-based PheWAS, the genomics viewer of the region around the gene (gene body ± 200 kb), and scatter plot of significant TWAS in eQTL, eeQTL and sQTL.

ant search module enable users to perform an exact search, but also provide region search module to narrow down a genome region to candidate target genes or variants for users to perform a fuzzy search. We presents the QTL region (SSC7: 97136980~97910694), the lead SNP (rs333375257) and gene (ABCD4) as case studies to explore the usage of gene/variant search in the PigBiobank.

In the region search module, three sections are presented: (i) 'Summary of Region', (ii) 'Content of Region' and (iii) 'Associated Gene List' (Figure 4A). The sections of 'Summary of Region' and 'Content of Region' show the information of the genomic region being queried and genes within the genomic region. The section of 'Associated Variant/Gene List' shows a list of significant GWAS variants within the genomic region, and genes whose range of gene body overlapped with the re-

gion. For instance, the region of 'SSC7: 97136980~97910694' presents a total of 2324 significant GWAS variants and 16 genes in the third section.

In the variant search module, the following four sections are presented: (i) 'Summary of Variant', (ii) 'Content of Variant', (iii) 'Summary-based Analysis', and (iv) 'Associated Genes' (Figure 4B). The section of 'Summary of Variant' shows the detailed information of the variant such as the physical position, two alleles, and the phastCons score across 100 vertebrate species. The section titled 'Content of Variant' visualizes a series of genes located around the variant using IGV. The section of 'Summary-based Analysis' shows the result of PheWAS to explore which traits are associated with this genomic variant. As an example, the variant of 'rs333375257' was significantly associated with multiple meat and carcass traits (Figure

4B). In the section of ‘Associated Genes’, PigBiobank provides a user defined list of genes either by ‘location’ or ‘molQTL mapping’. For instance, 21 and six genes can be obtained by querying ‘rs333375257’ for ‘location’ and ‘molQTL mapping’ search, respectively.

In the gene search module, the following four sections are presented: (i) ‘Summary of Gene’, (ii) ‘Content of Region around Gene’, (iii) ‘Summary-based Analysis’, and (iv) ‘Integrative Analysis’ (Figure 4C). The section of ‘Summary of Gene’ shows the information of queried gene including gene ensembl ID, gene name, the genomic range of gene body, and the phastCons score across 100 vertebrate species. The section of ‘Content of Region around Gene’ visualizes a series of genes around the queried gene using IGV. The section of ‘Summary-based Analysis’ presents the associations across all complex traits within the PigBiobank named as ‘PheWAS’ module. The section of ‘Integrative analysis’ provides diagrams and lists of associated gene-tissue-trait pairs based on TWAS, SMR, and COLOC. For instance, in PheWAS, *ABCD4* is found to be significantly associated with both the meat and carcass traits (e.g. backfat thickness) and the reproductive traits (e.g. total teat number) ($P < 5.0000 \times 10^{-8}$) (Figure 4C). In TWAS, we identify that *ABCD4* is significantly associated with a variety of traits across multiple tissues (Figure 4C). All these results illustrate the potential of PigBiobank in elucidating regulatory mechanism of complex traits, such as pleiotropy.

Discussion

To the best of our knowledge, the PigBiobank is the largest and most comprehensive database that integrates large-scale multi-omics data to deeply resolve the molecular and cellular mechanisms underlying complex traits in pigs. The PigBiobank is freely available to the public, and easy-to-accessible without logging in or registering. We would be anticipating more collaborators from around the world (pig-biobank@farmgtex.org). The first version of the PigBiobank, an encyclopedia of pig complex traits, curates a dataset of 71 885 pigs with genotypes and phenotypes from over 100 breeds, representing 264 distinct complex traits. The PigBiobank has three main features: (i) data standardization and sharing. The PigBiobank has undergone strict quality controls from raw data collection to data integrative analysis and adopted a uniform pipeline for analysis to ensure the quality of the data and the reliability of the results. Additionally, all the 264 traits in the PigBiobank have been classified into six main categories and 22 sub categories with unique and standardized indexes and names. The PigBiobank stored and shared detailed outcomes of the meta-GWAS and multi-omics integrative analysis; (ii) novel biological insights into the complex traits in pigs. By integrating the functional annotations collected from multiple consortia, such as FarmGTEx and FAANG projects, the PigBiobank provides a prototype for researchers to validate and annotate their GWAS findings, shedding light on the novel biological mechanisms underlying complex traits. Leveraging the resource of the PigBiobank, users could help identify candidate causal genes and variants underlying economically important traits in pigs, which will accelerate selective breeding to ensure food security for the growing global population in an environmentally sustainable way. (iii) Continuous update. Given that the GWAS data and molQTL data will be expanding and updated regularly in the coming years, we will keep the PigBiobank data updated con-

tinuously along with the PigGTEx project and functionality updated continuously annually. In the coming version, the PigBiobank will include more complex traits from more populations and breeds, as well as more diverse omics data. It will also provide user-friendly tools and a stable backend framework to support interactive and real-time analysis. We expect it will be a state-of-the-art, easy-to-use and open-access resource for biologically and genetically deciphering complex traits in pigs.

Data availability

The PigBiobank is freely available without registration at <http://pigbiobank.farmgtex.org>.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

The authors would like to thank the users, data providers and open-source contributors to the database. We thank National Supercomputer Center in Guangzhou China for its support in providing computing resources. We are grateful for the technical support provided by Guangdong iPig Technology Co. Ltd. We also appreciate the icons made by Surang, xnimrodx, Freepike, Becris, Marz Gallery, Smashicons, Iconic Panda, Pixel perfect, photo3idea_studio from www.flaticon.com, and Luciano Jung from www.iconfinder.com.

Author contributions: Haonan Zeng: Data curation, Visualization, Project administration, Writing-original draft. Wenjing Zhang: Formal Analysis, Validation, Writing-Original draft. Qing Lin: Formal Analysis, Writing-Original draft. Yahui Gao: Conceptualization, Writing-review & editing. Jinyan Teng: Formal Analysis, Visualization. Zhiting Xu: Formal Analysis. Xiaodian Cai: Formal Analysis. Zhanming Zhong: Formal Analysis. Jun Wu: Formal Analysis. Yuqiang Liu: Formal Analysis. Shuqi Diao: Formal Analysis. Chen Wei: Validation. Wentao Gong: Validation. Xiangchun Pan: Validation. Zedong Li: Investigation. Xiaoyu Huang: Investigation. Xifan Chen: Visualization. Jinshi Du: Visualization. Fuping Zhao: Resource. Yunxiang Zhao: Resource. Maria Ballester: Resource. Daniel Crespo-Piazuelo: Resource. Marcel Amills: Resource. Alex Clop: Resource. Peter Karlskov-Mortensen: Resource. Merete Fredholm: Resource. Pinghua Li: Resource. Ruihua Huang: Resource. Guoqing Tang: Resource. Mingzhou Li: Resource. Xiaohong Liu: Resource. Yaosheng Chen: Resource. Qin Zhang: Resource. Jiaqi Li: Resource. Funding acquisition. Xiaolong Yuan: Resource. Xiandong Ding: Resource. Conceptualization, Writing-review & editing. Lingzhao Fang: Resource, Conceptualization, Writing-review & editing. Zhe Zhang: Resource, Conceptualization, Funding acquisition, Supervision, Writing-review & editing.

Funding

National Natural Science Foundation of China [32022078]; National Key R&D Program of China [2022YFF1000900]; Local Innovative and Research Teams Project of Guangdong Province [2019BT02N630]; China Agriculture Research Sys-

tem [CARS-35]. Funding for open access charge: National Natural Science Foundation of China [32022078].

Conflict of interest statement

None declared.

References

1. FAO. (2022) *Meat Market Review: Emerging Trends and Outlook 2022*, Rome.
2. Pabst, R. (2020) The pig as a model for immunology research. *Cell Tissue Res.*, **380**, 287–304.
3. Lunney, J.K., van Goor, A., Walker, K.E., Hailstock, T., Franklin, J. and Dai, C. (2021) Importance of the pig as a human biomedical model. *Sci. Transl. Med.*, **13**, eabd5758.
4. Zhao, J., Ross, J.W., Hao, Y., Spate, L.D., Walters, E.M., Samuel, M.S., Rieke, A., Murphy, C.N. and Prather, R.S. (2009) Significant improvement in cloning efficiency of an inbred miniature pig by histone deacetylase inhibitor treatment after somatic cell nuclear transfer. *Biol. Reprod.*, **81**, 525–530.
5. Lind, N.M., Moustgaard, A., Jelsing, J., Vajta, G., Cumming, P. and Hansen, A.K. (2007) The use of pigs in neuroscience: modeling brain disorders. *Neurosci. Biobehav. Rev.*, **31**, 728–751.
6. Rogers, C.S. (2016) Genetically engineered livestock for biomedical models. *Transgenic Res.*, **25**, 345–359.
7. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
8. Li, M.J., Liu, Z., Wang, P., Wong, M.P., Nelson, M.R., Kocher, J.-P.A., Yeager, M., Sham, P.C., Chanock, S.J., Xia, Z., et al. (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
9. Watanabe, K., Stringer, S., Frei, O., Umičević Mirkov, M., Leeuw, C.d., Polderman, T.J.C., van der Sluis, S., Andreassen, O.A., Neale, B.M. and Posthuma, D. (2019) A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.*, **51**, 1339–1348.
10. Pan, S., Kang, H., Liu, X., Lin, S., Yuan, N., Zhang, Z., Bao, Y. and Jia, P. (2023) Brain Catalog: a comprehensive resource for the genetic landscape of brain-related traits. *Nucleic Acids Res.*, **51**, D835–D844.
11. Ghousaini, M., Mountjoy, E., Carmona, M., Peat, G., Schmidt, E.M., Hercules, A., Fumis, L., Miranda, A., Carvalho-Silva, D., Buniello, A., et al. (2021) Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.*, **49**, D1311–D1320.
12. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
13. Chen, S., Liu, S., Shi, S., Jiang, Y., Cao, M., Tang, Y., Li, W., Liu, J., Fang, L., Yu, Y., et al. (2022) Comparative epigenomics reveals the impact of ruminant-specific regulatory elements on complex traits. *BMC Biol.*, **20**, 273.
14. Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., Bai, L., Cai, Z., Zhao, B., Li, X., et al. (2022) A compendium of genetic regulatory effects across pig tissues. bioRxiv doi: <https://doi.org/10.1101/2022.11.11.516073>, 11 November 2022, preprint: not peer reviewed.
15. Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
16. Umans, B.D., Battle, A. and Gilad, Y. (2021) Where are the disease-associated eQTLs? *Trends Genet.*, **37**, 109–124.
17. Hu, Z.-L., Park, C.A. and Reecy, J.M. (2022) Bringing the animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res.*, **50**, D956–D961.
18. Guan, D., Bai, Z., Zhu, X., Zhong, C., Hou, Y., Lan, F., Diao, S., Yao, Y., Zhao, B., Zhu, D., et al. (2023) The ChickenGTEx pilot analysis: a reference of regulatory variants across 28 chicken tissues. bioRxiv doi: <https://doi.org/10.1101/2023.06.27.54667029>, 29 June 2023, preprint: not peer reviewed.
19. Santhanam, N., Sanchez-Roige, S., Liang, Y., Chitre, A.S., Munro, D., Chen, D., Cheng, R., Nyasimi, F., Perry, M., Gao, J., et al. (2022) RatXcan: framework for translating genetic results between species via transcriptome-wide association analyses. bioRxiv doi: <https://doi.org/10.1101/2022.06.03.494719>, 05 June 2022, preprint: not peer reviewed.
20. Xu, Z., Lin, Q., Cai, X., Zhong, Z., Li, B., Teng, J., Zeng, H., Gao, Y., Cai, Z., Wang, X., et al. (2023) Integrating large-scale meta-GWAS and PigGTEx resources to decipher the genetic basis of complex traits in pig. bioRxiv doi: <https://doi.org/10.1101/2023.10.09.561393>, 11 October 2023, preprint: not peer reviewed.
21. Browning, B.L., Zhou, Y. and Browning, S.R. (2018) A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, **103**, 338–348.
22. Fick, S.E. and Hijmans, R.J. (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.*, **37**, 4302–4315.
23. Harris, I., Osborn, T.J., Jones, P. and Lister, D. (2020) Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data.*, **7**, 109.
24. Liu, Y., Liu, R. and Chen, J.M. (2012) Retrospective retrieval of long-term consistent global leaf area index (1981–2011) from combined AVHRR and MODIS data. *J. Geophys. Res.*, **117**, G04003.
25. Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L., Kern, C., Halstead, M., Chanthavixay, G., Trakooljul, N., et al. (2021) Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat. Commun.*, **12**, 5848.
26. Yang, J., Lee, S.H., Goddard, M.E. and Visscher, P.M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
27. Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M. and Yang, J. (2019) A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.*, **51**, 1749–1755.
28. Jiang, L., Zheng, Z., Fang, H. and Yang, J. (2021) A generalized linear mixed model association tool for biobank-scale data. *Nat. Genet.*, **53**, 1616–1621.
29. Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.
30. Zhou, X. and Stephens, M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.
31. Chen, G.-B., Lee, S.H., Zhu, Z.-X., Benyamin, B. and Robinson, M.R. (2016) EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity (Edinb)*, **117**, 51–61.
32. Leeuw, C.A.d., Mooij, J.M., Heskes, T. and Posthuma, D. (2015) MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.*, **11**, e1004219.
33. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L. and Neale, B.M. (2015) LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, **47**, 291–295.
34. GTEx Consortium, Pividori, M., Rajagopal, P.S., Barbeira, A., Liang, Y., Melia, O., Bastarache, L., Park, Y., Wen, X. and Im, H.K.

- (2020) PhenomeXcan: mapping the genome to the phenome through the transcriptome. *Sci. Adv.*, **6**, eaba2083.
35. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., *et al.* (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.*, **9**, 1825.
 36. Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L. and Im, H.K. (2019) Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.*, **15**, e1007889.
 37. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., *et al.* (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.*, **48**, 481–487.
 38. Cingolani, P., Platts, A., Le Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.
 39. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.*, **47**, 1228–1235.
 40. Consortium, T.G.T.E., Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., *et al.* (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
 41. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
 42. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A. and Malinverni, R. (2016) regioneR: an R/bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289–291.
 43. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shores, N., *et al.* (2018) Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.*, **50**, 621–629.
 44. Robinson, J.T., Thorvaldsdottir, H., Turner, D. and Mesirov, J.P. (2023) igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics*, **39**, btac830.
 45. Kim, K.S., Larsen, N., Short, T., Plastow, G. and Rothschild, M.F. (2000) A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. *Mamm. Genome*, **11**, 131–135.
 46. Kim, K.-S., Reecy, J.M., Hsu, W.H., Anderson, L.L. and Rothschild, M.F. (2004) Functional and phylogenetic analyses of a melanocortin-4 receptor mutation in domestic pigs. *Domest. Anim. Endocrinol.*, **26**, 75–86.
 47. Thongkuy, S., Chuaychu, S.B., Burarnrak, P., Ruangjoy, P., Juthamane, P., Nuntapaitoon, M. and Tummaruk, P. (2020) Effect of backfat thickness during late gestation on farrowing duration, piglet birth weight, colostrum yield, milk yield and reproductive performance of sows. *Livest. Sci.*, **234**, 103983.
 48. Cheng, C., Wu, X., Zhang, X., Zhang, X. and Peng, J. (2019) Obesity of sows at late pregnancy aggravates metabolic disorder of perinatal sows and affects performance and intestinal health of piglets. *Animals*, **10**, 49.
 49. Hu, J. and Yan, P. (2022) Effects of backfat thickness on oxidative stress and inflammation of placenta in large white pigs. *Vet. Sci.*, **9**, 302.

A compendium of genetic regulatory effects across pig tissues

Received: 23 November 2022

Accepted: 13 October 2023

Published online: 4 January 2024

 Check for updates

A list of authors and their affiliations appears at the end of the paper

The Farm Animal Genotype-Tissue Expression (FarmGTEx) project has been established to develop a public resource of genetic regulatory variants in livestock, which is essential for linking genetic polymorphisms to variation in phenotypes, helping fundamental biological discovery and exploitation in animal breeding and human biomedicine. Here we show results from the pilot phase of PigGTEx by processing 5,457 RNA-sequencing and 1,602 whole-genome sequencing samples passing quality control from pigs. We build a pig genotype imputation panel and associate millions of genetic variants with five types of transcriptomic phenotypes in 34 tissues. We evaluate tissue specificity of regulatory effects and elucidate molecular mechanisms of their action using multi-omics data. Leveraging this resource, we decipher regulatory mechanisms underlying 207 pig complex phenotypes and demonstrate the similarity of pigs to humans in gene expression and the genetic regulation behind complex phenotypes, supporting the importance of pigs as a human biomedical model.

Genome-wide association studies (GWAS) reveal genomic variants associated with complex phenotypes at an unprecedented speed and scale in both plants¹ and animals², but particularly in humans^{3,4}. However, most of the variants fall in noncoding regions, putatively contributing to phenotypic variation by regulating gene activity at different biological levels^{5,6}. The systematic characterization of genetic regulatory effects on transcriptome (for example, expression quantitative trait loci (eQTLs)) across tissues, as carried out in the Genotype-Tissue Expression (GTEx) project in humans⁷, has proven to be a powerful strategy for connecting GWAS loci to gene regulatory mechanisms at large scale^{6,8,9}.

To sustain food and agriculture production while minimizing associated negative environmental impacts, it is crucial to identify molecular mechanisms that underpin complex traits of economic importance to enable biology-driven selective breeding in farm animals. However, the annotation of regulatory variants in farm animals has so far been limited by small sample size, few tissue/cell type assayed, and in restricted genetic background^{10–12}. We thus launched the international Farm Animal GTEx (FarmGTEx) project to build a comprehensive atlas of regulatory variants in domestic animal species. This resource along with the functional annotation of animal genomes project will not only facilitate fundamental biology discovery but also enhance the genetic improvement of farm animals¹³.

Pigs are an important agricultural species by supplying meat for humans, and serve as an important biomedical model for studying human development, disease and organ xenotransplantation, due to their similarity to humans in multiple attributes such as anatomical structure, physiology and immunology¹⁴. Here we report the results of the pilot PigGTEx, which is underpinned by 5,457 RNA-seq data and 1,602 whole-genome sequence (WGS) samples (Supplementary Tables 1 and 2). We test the association of transcriptomic phenotypes with 3,087,268 DNA variants in 34 pig tissues and then evaluate tissue-sharing patterns of regulatory effects. We examine multi-omics data to identify putative molecular mechanisms underlying regulatory variants and then apply this resource to dissect GWAS associations for 268 complex traits. Finally, we leverage the human GTEx resource and GWAS of 136 human complex phenotypes to assess the similarity between pigs and humans in genetic regulation of gene expression and complex phenotypes. We make the PigGTEx resources freely accessible via <http://piggtx.farmgtex.org>.

Results

Data summary

After filtering out the low-quality samples from the initial set of 9,530, we retained 7,095 RNA-seq profiles for downstream analysis (Supplementary Fig. 1 and Supplementary Note). We quantified expression

✉ e-mail: albert.tenesa@ed.ac.uk; likui@caas.cn; george.liu@usda.gov; zhezhang@scau.edu.cn; lingzhao.fang@qgg.au.dk

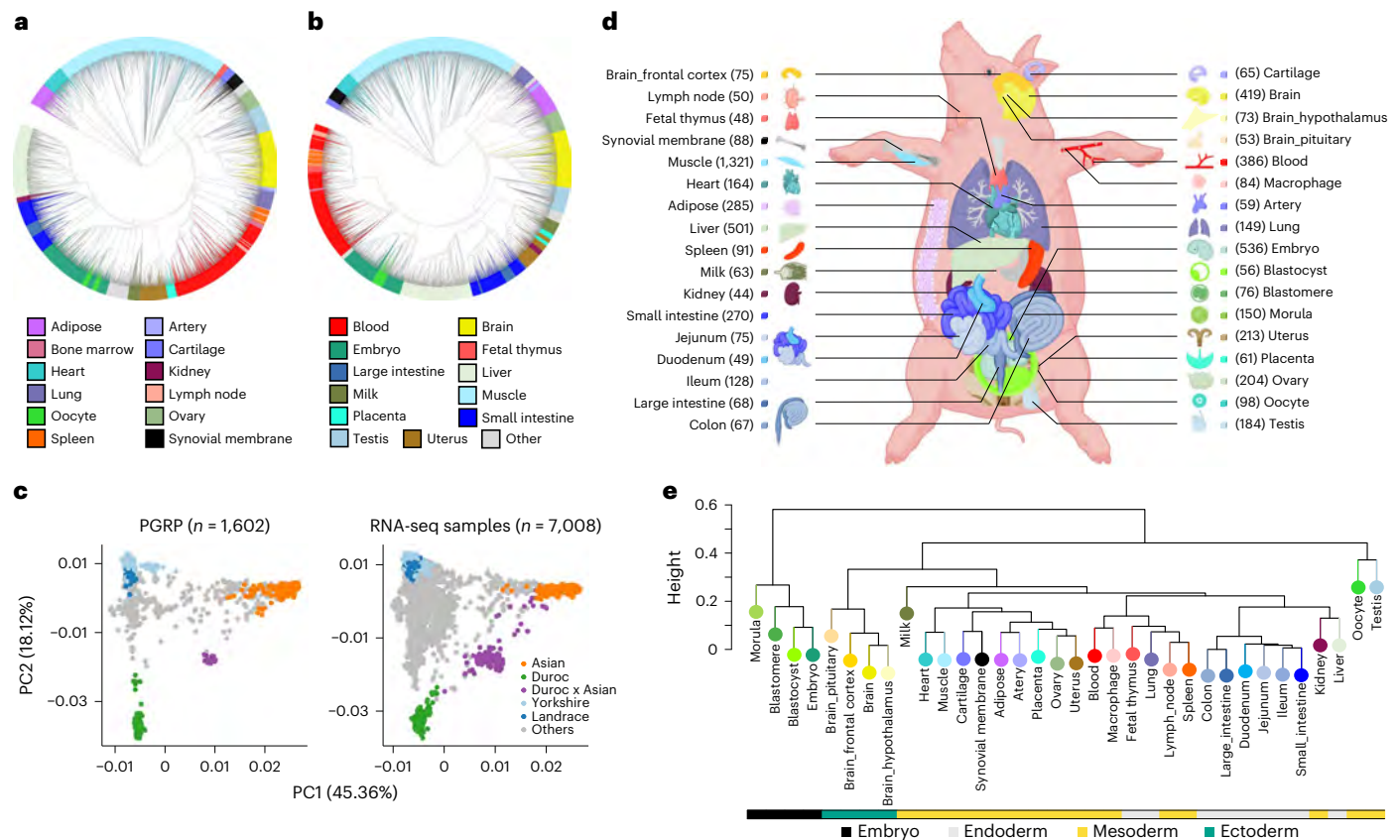


Fig. 1 | Characteristics of samples in the pilot phase of PigGTEx project.

a, Clustering of 7,095 RNA-seq samples based on the normalized expression (\log_{10} -transformed TPM) of 6,500 highly variable genes, defined as the top 20% of genes with the largest s.d. of TPM across samples. **b**, The same sample clustering as **a** but based on normalized alternative splicing values (PSI) of 6,500 highly variable spliced introns, defined as the top 13% of spliced introns with the largest s.d. of PSI across samples. **c**, Principal component analysis of samples based

on 12,207 LD-independent ($r^2 < 0.2$) SNPs. The left panel is for whole-genome sequencing samples ($n = 1,602$) in the PGRP, while the right one is for RNA-seq samples ($n = 7,008$) with successful genotype imputations. **d**, Sample sizes of 34 tissues, cell types and organ systems (all referred to as 'tissues') used for molQTLs mapping. **e**, Clustering of 34 tissues based on the median expression of all 31,871 Ensembl annotated genes (v100) across samples within tissues, representing embryo, endodermal, mesodermal and ectodermal lineages.

levels for protein-coding genes (PCG), lncRNA, exons and enhancers, and alternative splicing events in these samples. Sample clustering based on the five transcriptomic phenotypes recapitulated tissue types well (Fig. 1a,b and Supplementary Fig. 2). We called a median number of 74,347 single-nucleotide polymorphisms (SNPs) from these RNA-seq samples (Extended Data Fig. 1a,b). Leveraging a multibreed pig genomics reference panel (PGRP) consisting of 1,602 WGS samples (Supplementary Fig. 3), we imputed genotypes of RNA-seq samples with an imputation accuracy of 0.94 (concordance rate) and 0.82 (genotype correlation, r^2 ; Extended Data Fig. 1c–n and Supplementary Table 3). The population structure of the RNA-seq samples was similar to the PGRP (Fig. 1c). After removing duplicated RNA-seq samples, we retained 5,457 samples representing 34 tissues, cell types or organ systems (all referred to as 'tissues' hereafter), with at least 40 samples per tissue, for subsequent analysis (Fig. 1d–e, Extended Data Fig. 2a–e and Supplementary Table 4). We further analyzed 270 multi-omics datasets in pigs, including 245 whole-genome bisulfite sequencing (WGBS; Supplementary Figs. 4 and 5 and Supplementary Tables 5–7), 20 single-cell RNA-seq (Supplementary Fig. 6 and Supplementary Table 8) and five Hi-C samples (Supplementary Tables 9 and 10).

The gene expression atlas empowers functional annotation

Gene expression was either tissue-specific or ubiquitous (Supplementary Fig. 7a and Extended Data Fig. 3a). We detected between 145 (morula) and 5,180 (frontal cortex) tissue-specific genes across 34 tissues (Extended Data Fig. 3b and Supplementary Fig. 7b).

Tissue-specific genes showed a higher enrichment of active regulatory elements and a higher depletion of repressed polycomb regions in matching tissues than in nonmatching tissues¹⁵ (Extended Data Fig. 3c–e and Supplementary Fig. 7c,d). In addition, tissue-specific genes exhibited distinct patterns of evolutionary DNA sequence constraints across tissues (Supplementary Fig. 7e), in agreement with the hypothesis of tissue-driven evolution¹⁶. To assign function to pig genes, we performed a gene co-expression analysis in each of the 34 tissues (Supplementary Fig. 8a–c). In total, we detected 5,309 co-expression modules across tissues and assigned 25,023 genes to at least one module (Supplementary Fig. 8d–f and Supplementary Table 11). Among them, 13,266 (42.57%) genes had no functional annotation in the Gene Ontology (GO) database (Extended Data Fig. 3f and Supplementary Fig. 8d); these are referred to as 'unannotated genes' hereafter. For instance, 42 unannotated genes were co-expressed with 59 functional annotated genes in the pituitary, which were substantially enriched in neuron apoptotic processes (Extended Data Fig. 3g). Unannotated genes were less expressed, showed weaker DNA sequence conservation, lower proportion of orthologous genes and higher tissue specificity than genes with functional annotations (Extended Data Fig. 3f). The proportion of expressed unannotated genes varied across tissues, indicating differences in functional annotation between tissues (Extended Data Fig. 3h).

MolQTL mapping

In total, 93% of tested genes had significant *cis*-heritability (*cis*- h^2 ; within ± 1 Mb of transcription start sites (TSS)) estimates in at least one

tissue while accounting for hidden factors (Extended Data Fig. 2f–h and Extended Data Fig. 4a,b). We mapped molecular quantitative trait loci (molQTLs) for five molecular phenotypes, including *cis*-eQTL for PCG expression, *cis*-eeQTL for exon expression, *cis*-lncQTL for lncRNA expression, *cis*-enQTL for enhancer expression and *cis*-sQTL for alternative splicing. In total, 86%, 67%, 46%, 27% and 64% of all tested PCGs ($n = 17,431$), lncRNAs ($n = 7,374$), exons ($n = 82,678$), enhancers ($n = 3,353$) and genes with alternative splicing events ($n = 18,331$) had at least one significant variant (eVariant) detected in at least one tissue; hence, they were defined as eGenes, eLncRNAs, eExons, eEnhancers and sGenes, respectively (Supplementary Fig. 9 and Supplementary Table 12). The proportion of eGenes detected was positively correlated with sample size across tissues, similar to the other four molecular phenotypes (Fig. 2a, Extended Data Fig. 4c and Supplementary Fig. 10). The top *cis*-eQTL centered around TSS of genes (Supplementary Fig. 11a–e). Tissues with a larger sample size yielded a larger proportion of *cis*-eQTL with smaller effects (Supplementary Fig. 11f–g). PCG had the highest proportion of detected eGenes across tissues, followed by lncRNA, enhancer, splicing and finally exon (Fig. 2b). Notably, molecular phenotypes exhibited a high proportion (an average of 70%) of their own specific molQTL after taking linkage disequilibrium (LD) between SNPs into account (Fig. 2b), indicative of their distinct underlying genetic regulation. On average, 20% of eGenes, 13.5% of sGenes, 21.2% of eExons, 23.5% of eLncRNAs and 21% of eEnhancers had more than one independent eVariant across tissues, and the proportion increased with an increasing sample size of tissues (Fig. 2c and Extended Data Fig. 5a). Down-sampling analysis in three major tissues further confirmed an impact of sample size on the statistical power for *cis*-eQTL discovery (Fig. 2d). Approximately half of the independent *cis*-eQTL were located within ± 182 kb of TSS, and those with larger effect size were closer to TSS (Extended Data Fig. 5b–d). The eGenes with more independent *cis*-eQTL have a higher *cis*- h^2 , but no significant differences for the median gene expression level (Fig. 2e).

We applied four distinct strategies to validate the *cis*-eQTL. First, the summary statistics of *cis*-eQTL derived from the linear regression model¹⁷ had a strong correlation with those from a linear mixed model (Extended Data Fig. 6a–e). Second, the internal validation yielded a high replication rate (measured by π_1) of *cis*-eQTL, with an average π_1 value of 0.92 (range: 0.80–1.00) and an average of 0.56 (range 0.36–0.89) for Pearson's r between effect sizes across tissues (Fig. 2f). Third, 92%, 74%, 73% and 69% of *cis*-eQTL in blood, liver, duodenum and muscle, respectively, were replicated in independent datasets (Extended Data Fig. 6f–h). Fourth, effects derived from allele-specific expression (ASE) analysis were correlated with those from *cis*-eQTL mapping (Fig. 2g and Extended Data Fig. 6i–k). In addition, we conducted an exploratory analysis of *trans*-eQTL in 12 tissues with over 150 individuals and detected an average of 80 *trans*-eGenes (false discovery rate, FDR < 0.05) across tissues (Supplementary Fig. 12a,b). We took the muscle that had the largest sample size ($n = 1,321$) as an example to conduct an internal validation of *trans*-eQTL by randomly and evenly dividing samples into two groups. We observed that the replication rate (π_1) between the two groups was 0.4 and the Pearson's correlation of effect sizes of significant *trans*-eQTL between groups was 0.5 (Supplementary Fig. 12c).

To understand how *cis*-eQTL are shared across pig breeds, we considered muscle as an example. We divided muscle samples into eight breed groups (all referred to as 'breeds' hereafter) and performed *cis*-eQTL mapping separately (Extended Data Fig. 7a and Supplementary Table 13). Across all eight breeds, we detected 9,548 unique *cis*-eGenes, of which 97.1% could be replicated in at least two of these breeds (Fig. 2h and Extended Data Fig. 7b,c). The replication rates were higher in breeds with more samples (Extended Data Fig. 7d). For instance, the Landrace \times Yorkshire cross-breed had the largest sample size ($n = 374$) replicated on average 95.6% of the *cis*-eQTL detected in

the other seven breeds (Extended Data Fig. 7d). The *cis*-eQTL effects were positively correlated between breeds and clearly separated from other tissues (Fig. 2i and Extended Data Fig. 7e). In addition, the effects of *cis*-eQTL from the multibreed meta-analysis were correlated with those from the combined muscle population (Extended Data Fig. 7f). Compared to the single-breed meta-analysis, the combined population detected 86.2% more *cis*-eQTL (Extended Data Fig. 7g). To explore whether breed interacts with genotype to modulate expression of some genes, we conducted breed-interaction *cis*-eQTL (bieQTL) mapping. In total, 589 genes had at least one significant bieQTL in 13 tissues (Fig. 2j,k, Extended Data Fig. 7h,i and Supplementary Table 14). Furthermore, we conducted a cell-type deconvolution analysis in seven tissues, demonstrating the variation of cell-type composition across bulk tissue samples (Extended Data Fig. 8a). A total of 376 genes had at least one significant cell-type interaction *cis*-eQTL (cieQTL) in three tissues (Fig. 2l–m, Extended Data Fig. 8b,c and Supplementary Table 14). In addition, we validated half of bieQTL and cieQTL with the ASE approach¹⁸ (Fig. 2j,l and Extended Data Fig. 8d–g).

Tissue-sharing patterns of molQTL

Tissues with similar functions clustered together, and the tissue relationship was consistent across all ten data types, including the five types of molQTL and the respective molecular phenotypes (Fig. 3a,b and Extended Data Fig. 9a,d). The most easily accessible samples, that is, blood and milk cells, showed an average correlation of 0.51 *cis*-eQTL effects with other tissues. Both had the highest similarity to immune tissues, followed by intestinal tissues, and finally testis and embryonic tissues. The overall tissue-sharing of molQTL showed a U-shaped curve (Fig. 3c). Among them, *cis*-eQTL of PCG had the highest degree of tissue-sharing, followed by *cis*-lncQTL, *cis*-sQTL, *cis*-eeQTL and finally *cis*-enQTL (Fig. 3c and Extended Data Fig. 9e). An eGene tended to be regulated by *cis*-eQTL of smaller effect if it showed a higher level of tissue-sharing or was expressed in more tissues (Fig. 3d and Extended Data Fig. 9f). The higher the tissue-sharing of eGenes, the larger the minor allele frequency (MAF) of their *cis*-eQTL, and the closer the distance of their *cis*-eQTL to TSS (Fig. 3d). In addition, eGenes that were active in more tissues had a decreased PhastCons score (that is, less sequence constraint), while genes that were not eGenes (non-eGenes) in more tissues had an increased PhastCons score (Fig. 3e). The shared non-eGenes in the 34 tissues were substantially enriched in fundamental biological processes (Supplementary Table 15). We summarized four types of SNP–gene pairs and observed that 1.8% (1,166/64,250) of top *cis*-eQTL of the same eGenes had an opposite effect in at least one tissue pair, representing 3.1% (467/14,988) of all detected eGenes (Fig. 3f). Compared to other tissue pairs, blood and testis showed the highest proportion (25%) of eGenes with opposite *cis*-eQTL effects (Fig. 3g). For example, *ODF2L*, which showed the opposite direction of eQTL effect (rs329043485) between blood and testis (Fig. 3h and Extended Data Fig. 9g–h), is involved in negative regulation of cilium assembly and spermatogenesis¹⁹.

Functional annotation of molQTL

Compared to other molQTL, *cis*-sQTL had a higher enrichment for missense variants, variants with a high impact on protein sequence and variants in splice region and acceptor sites (Fig. 4a and Supplementary Fig. 13a). Although there was a significant enrichment of molQTL in exonic annotations (for example, synonymous and missense), the proportion of such variants over all the molQTL was around 5.4%, that is, 5.4% for eQTL, 5.5% for sQTL, 5.2% for eeQTL, 5.4% for lncQTL and 5.8% for enQTL. This finding was consistent with human GTEx^{7,20} and RatGTEx²¹. Looking at chromatin states, these five types of molQTL showed the highest enrichment in active promoters, followed by those proximal to TSS and ATAC islands (Fig. 4b and Supplementary Fig. 13b). The molQTL with higher causality scores showed a higher enrichment in functional features (Supplementary Fig. 13c,d). Among

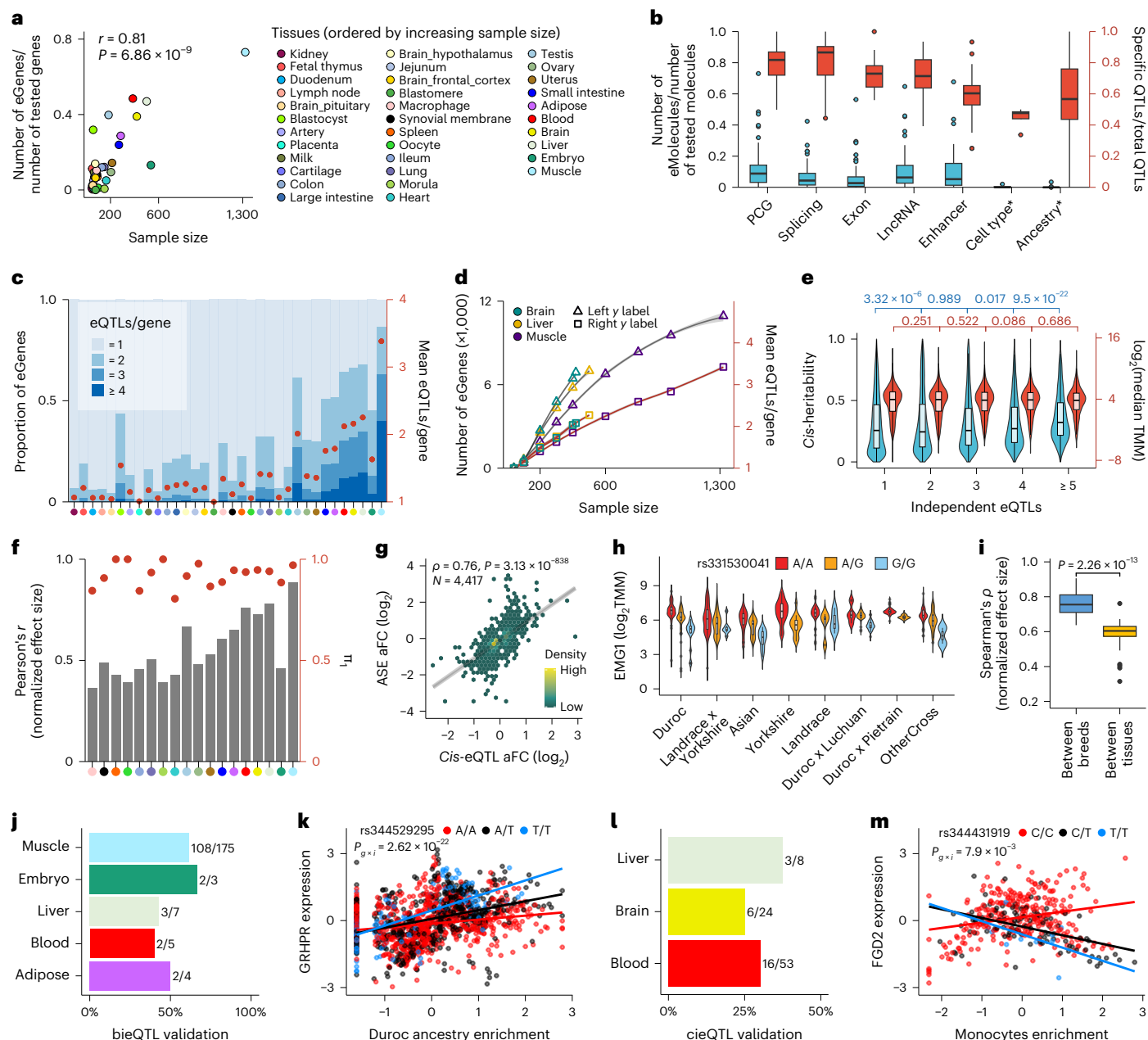


Fig. 2 | molQTL discovery. **a**, Pearson's r between the proportion of detectable eGenes and sample size across 34 tissues. **b**, Proportions of detectable eMolecule (blue) and specific molQTL (red) for different molecular phenotypes in 34 tissues. * indicates the interaction of *cis*-eQTLs (ieQTL). Cell type* and Ancestry* are for cell-type ieQTL (cieQTL) and breed/ancestry ieQTLs (bieQTL), respectively. **c**, Distribution and the average number of independent *cis*-eQTL per gene. Tissues (x axis) are ordered by increasing sample size. The color key is the same as in **a**. **d**, Number of eGenes (triangle) and average number of independent *cis*-eQTL (square). **e**, The comparison of *cis*- h^2 (blue) and median expression levels (red) of genes with different numbers of detectable independent *cis*-eQTL across tissues. The top labels show nominal P values (uncorrected for multiple testing) from one-sided Student's t tests. **f**, Internal validation of *cis*-eQTL. Bars represent Pearson's r of the normalized effects of *cis*-eQTL between validation and discovery groups. Points represent the π_1 statistic measuring the replication rate of *cis*-eQTL. **g**, Spearman's ρ of effect sizes (aFC in log₂ scale) between

cis-eQTL and ASE at matched loci ($n = 4,417$) in muscle. **h**, A *cis*-eQTL (rs331530041) of *EMG1* in muscle is shared across eight ancestry groups. **i**, Spearman's correlation of the *cis*-eQTL effects between eight breeds of the muscle (left) and between muscle and other 33 tissues (right). The P value is obtained from a two-sided Wilcoxon rank-sum test. **j**, Proportion of bieQTL that are validated with the ASE approach. The number of validated bieQTLs out of the total number of bieQTLs tested is shown to the right of each bar. **k**, Effect of eVariant (rs344529295) of *GRHR* interacted with the Duroc ancestry enrichment in muscle. The two-sided P value is calculated by the linear regression bieQTL model. The lines are fitted by a linear regression model using the `geom_smooth` function from `ggplot2` (v3.3.2) in R (v4.0.2). **l**, Proportion of cieQTL that are validated by the ASE approach. **m**, Effect of eVariant (rs344431919) of *FGD2* interacted with monocyte enrichment in blood. The two-sided P value is calculated by the linear regression cieQTL model. The lines are fitted using the same method as in **k**. aFC, allelic fold change.

all the five types of molQTL, *cis*-enQTL with high causality scores had the highest enrichment for enhancer-like chromatin states (Supplementary Fig. 13d). An average of 64% of *cis*-eQTL could potentially modify transcription factor binding sites (Supplementary Table 16). Although they

showed a weak enrichment for molQTL (except for *cis*-enQTL; Fig. 4b), enhancers had a higher enrichment for *cis*-eQTL in the matching tissue compared to nonmatching tissues (Fig. 4c). Notably, the top *cis*-eQTL tended to be enriched in promoters rather than enhancers, whereas

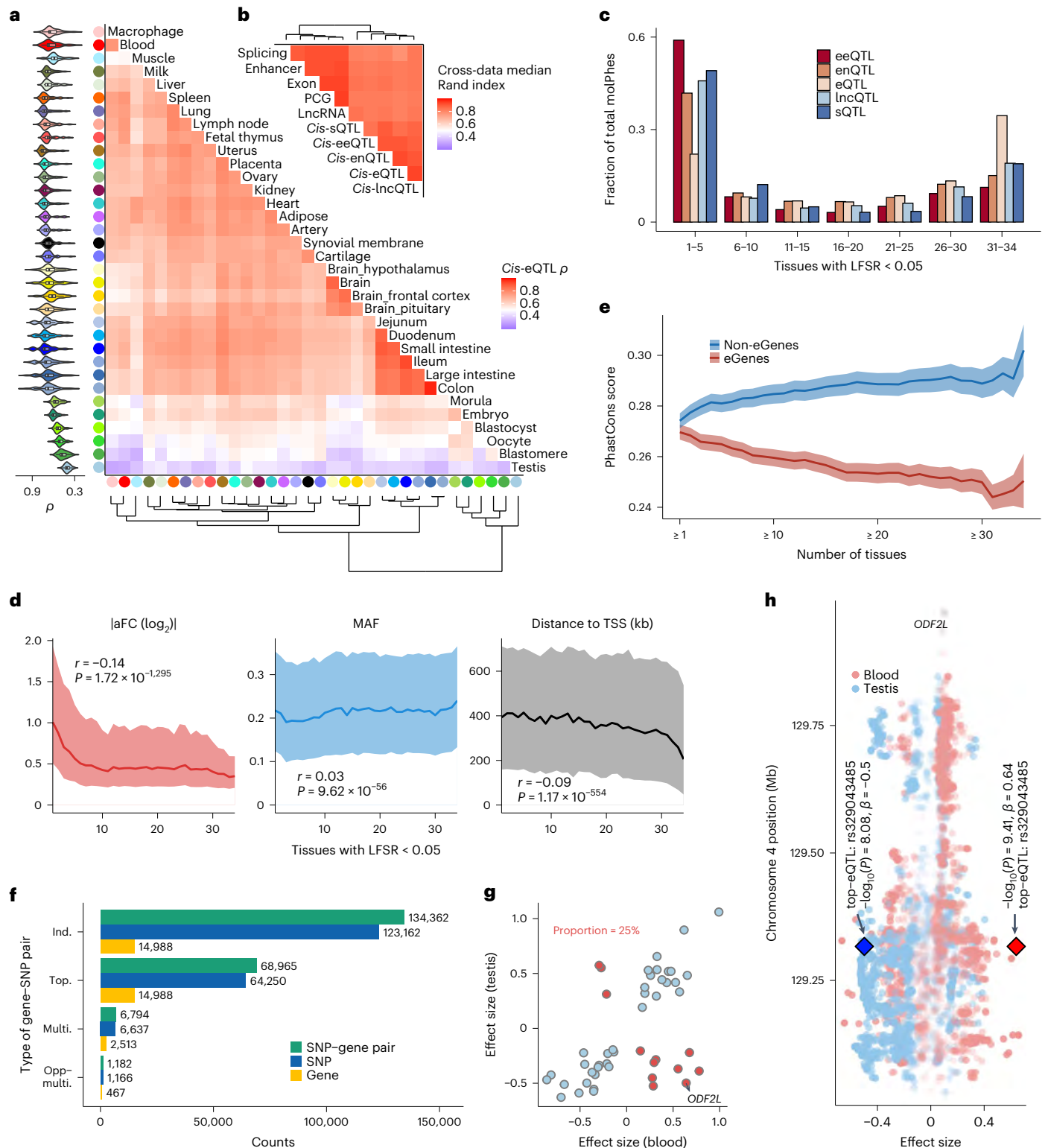


Fig. 3 | Tissue-sharing pattern of regulatory effects. a, Heatmap of tissues depicting the corresponding pairwise Spearman's correlation (ρ) of *cis*-eQTL effect sizes. Tissues are grouped by hierarchical clustering (bottom). Violin plots (left) represent Spearman's ρ between the target tissue and other tissues. **b**, Similarity (measured by the median pairwise Rand index) of tissue-clustering patterns across ten data types. **c**, The overall tissue-sharing pattern of five molQTL types at LFSR < 5% obtained by MashR (v0.2-6). **d**, Relationships between the magnitude of tissue-sharing of *cis*-eQTL and their effect sizes (aFC, left), MAFs (middle) and distances to the TSS (right). The P values are obtained by Pearson's correlation (r) test. The line and shading indicate the median and interquartile range, respectively. **e**, Conservation of DNA sequence (measured by the PhastCons score

of 100 vertebrate genomes) of eGenes and non-eGenes regarding tissue-sharing. The line and shading indicate the mean and standard error, respectively. **f**, Counts of four types of SNP-gene pairs across 34 tissues. Ind., independent *cis*-eQTL; top., top *cis*-eQTL; multi., eGenes have identical or high LD ($r^2 > 0.8$) *cis*-eQTL in any two tissues; opp-multi., eGenes have an opposite direction of *cis*-eQTL effect between any two tissues. **g**, Scatter plots of *cis*-eQTL effect sizes of 48 common multi-eGenes in blood and testis. *cis*-eQTL with the same directional effect are colored blue ($n = 36$), and those with the opposite direction are colored red ($n = 12$). **h**, The *cis*-eQTL effects of *ODF2L* on chromosome 4 in blood and testis. Diamond symbols represent the top *cis*-eQTL of *ODF2L*. The two-sided P value is calculated by the linear regression *cis*-eQTL model.

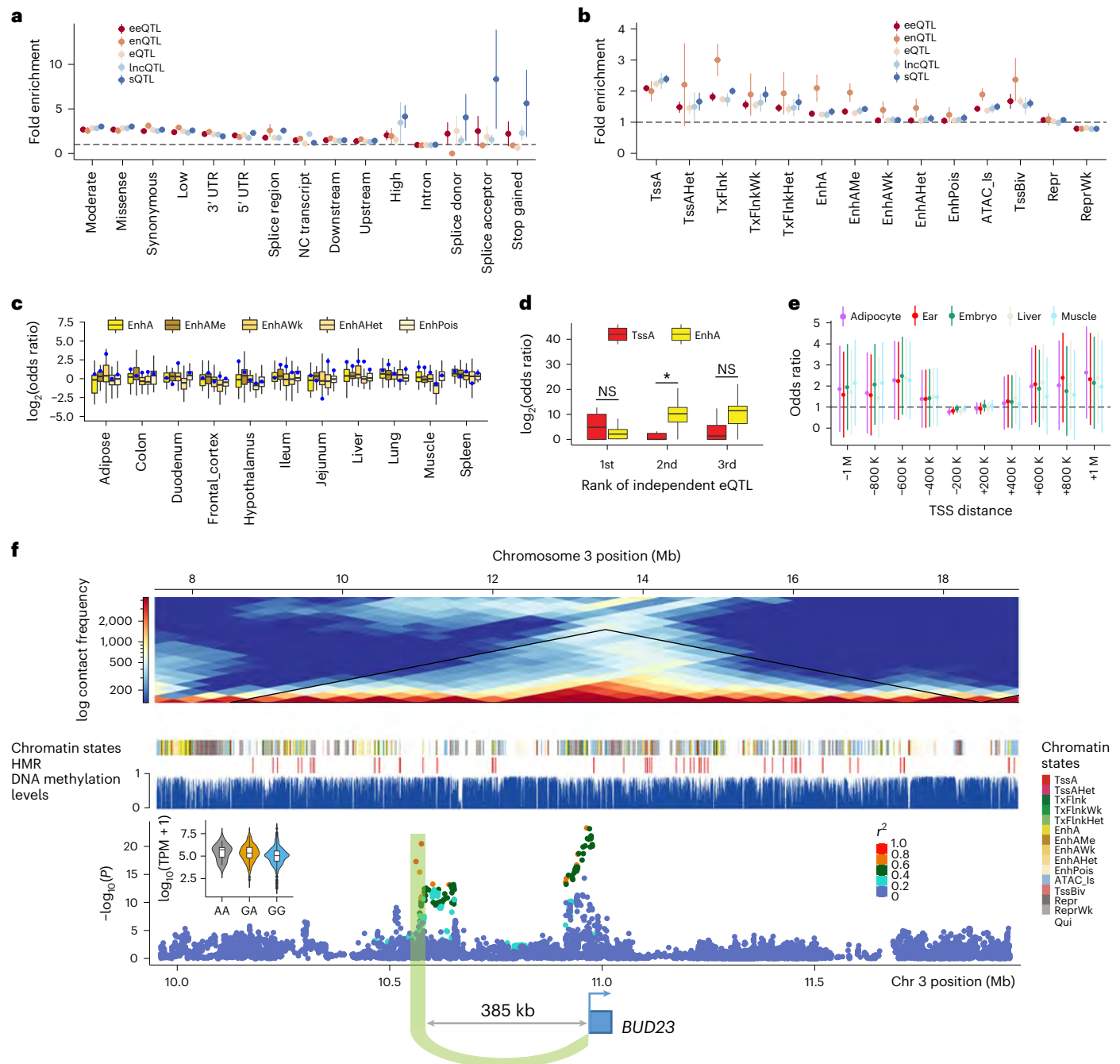


Fig. 4 | Functional characterization of regulatory variants. a, b, Fold enrichment (mean \pm s.d.) for fine-mapped molQTLs in sequence ontologies (**a**) and 14 chromatin states (**b**). **c,** Enrichment of *cis*-eQTL in five types of enhancers. Each box includes enrichment of *cis*-eQTL from 34 tissues across enhancers. Blue dots represent enrichments from matching tissues. **d,** Enrichment of top three independent *cis*-eQTL in two chromatin states. TssA is for active TSS, while EnhA is for active enhancers. The *P* values are obtained by the two-sided Student *t* test. **P* < 0.05 and NS indicates not significant. **e,** Enrichment (mean \pm s.d.) of *cis*-eQTL within the same topologically associating domain of TSS of target genes. TADs

are obtained from Hi-C data of five tissues. The *cis*-eQTL are grouped according to their distance to TSS. - and + means upstream and downstream, respectively. **f,** The landscape of *BUD23* at multiple genomic features in muscle. The top plot shows that *BUD23* and its second independent eVariant (rs790620973) are located within a TAD (the black triangle). The bottom is the Manhattan plot showing *cis*-eQTL results of *BUD23*. The violin plot shows the expression levels (\log_{10} -transformed TPM) of *BUD23* across three genotypes (AA, *n* = 9; GA, *n* = 131; GG, *n* = 1,181) of this eVariant in muscle. The two-sided *P* value is obtained from the linear regression *cis*-eQTL model.

the reverse was observed for the second- and third-ranked *cis*-eQTL (Fig. 4d). In addition, molQTL showed tissue-specific enrichment for hypomethylated regions (HMRs) and allele-specific methylation loci (Supplementary Fig. 13e). In muscle, 2,016 *cis*-eQTL, 4,694 *cis*-eeQTL, 524 *cis*-lncQTL, 5,174 *cis*-enQTL and 1,590 *cis*-sQTL were mediated by methylation QTL (Supplementary Fig. 13f,g and Supplementary Table 17). The long-distance *cis*-eQTL were substantially enriched in

the same topologically associating domain (TAD) as TSS of target genes after accounting for the *cis*-eQTL-TSS distance (Fig. 4e). This suggests that long-range *cis*-eQTL may affect gene expression by mediating 3D genome interactions²². For instance, in muscle, the second independent *cis*-eQTL of *BUD23* was 385 kb upstream of its TSS, and located within the same TAD of the TSS, as well as was surrounded by HMRs and enhancers (Fig. 4f).

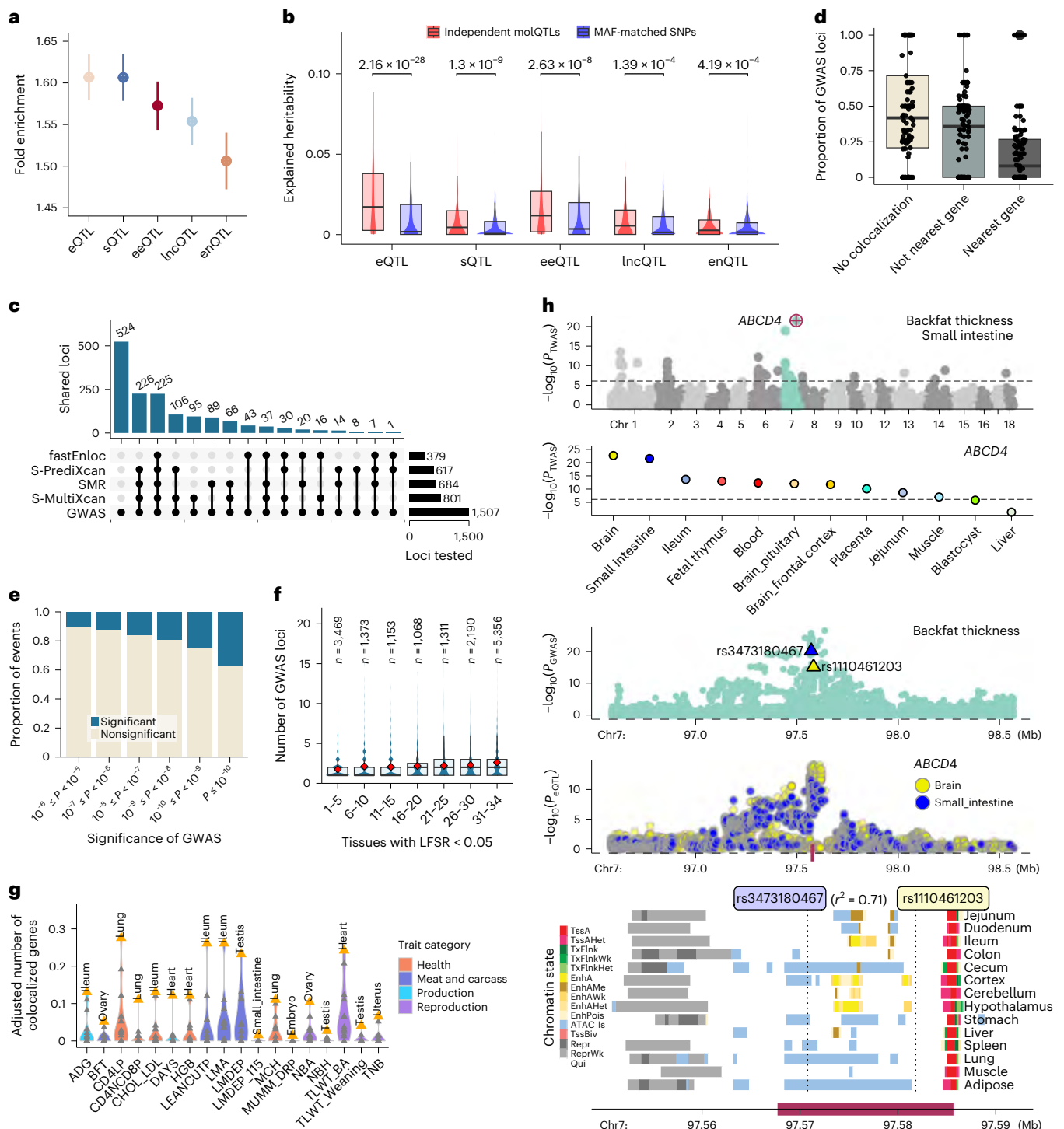


Fig. 5 | Interpreting GWAS loci of complex traits using molQTL. a, Enrichment (mean and 95% confidence interval) of GWAS variants with five types of molQTL in 34 tissues. **b**, Heritability of 16 complex traits of pig explained by independent molQTLs and those MAF-matched SNPs across 34 tissues. The top numerical labels are the nominal P values (uncorrected for multiple testing) based on the two-sided paired Student's t test. **c**, Number of GWAS loci linked to eGenes through fastENloc, SMR, S-PrediXcan and S-MultiXcan. The bottom point-line combinations of the upset plot represent the intersections of GWAS loci linked to eGenes by different methods. **d**, Proportion of three types of GWAS loci regarding the colocalization results, where 105 GWAS traits are shown in each category. No colocalization, GWAS loci that are not colocalized with any eGenes in 34 tissues. Not nearest gene, GWAS loci whose colocalized eGenes are not nearest genes to GWAS lead SNPs. Nearest gene, GWAS loci whose colocalized eGenes are the nearest ones. Each dot represents a complex trait. **e**, Proportion of significant

colocalizations of GWAS loci with *cis*-eQTL at various significance levels of GWAS. **f**, The number of colocalized GWAS loci per eGene across 105 traits above. eGenes are classified into seven groups regarding the tissue-sharing pattern. Diamond indicates the mean value. **g**, The number of colocalized genes adjusted for tissue sample size and eGene discovery ratio in 14 tissues across 18 GWAS traits (detailed abbreviations in Supplementary Table 18). Top tissues are labeled. **h**, The association of *ABCD4* with the average BFT. The top Manhattan plot represents the TWAS results of BFT in the small intestine, followed by the TWAS results of *ABCD4* for BFT in 12 tissues being tested. The two following Manhattan plots show the colocalization of BFT GWAS (top) and *cis*-eQTL (bottom) of *ABCD4* on chromosome 7 (chr 7) in both the brain and small intestine. The blue and yellow triangles indicate the top variants of *ABCD4* in the small intestine (rs3473180467) and brain (rs1110461203), respectively. These two variants are in high LD ($r^2 = 0.71$). The bottom panel is for chromatin states around *ABCD4*.

Interpreting GWAS loci with molQTL

To study the regulatory mechanisms underlying complex traits in pigs, we examined 268 GWAS summary statistics of 207 complex traits (Supplementary Table 18) and found that GWAS signals were enriched in molQTL (Fig. 5a and Supplementary Fig. 14a–e). Among them, *cis*-eQTL/*cis*-sQTL showed the highest enrichment (–1.61-fold, *s.e.* = 0.014), followed by *cis*-eeQTL (1.57-fold, *s.e.* = 0.015), *cis*-lncQTL (1.55-fold, *s.e.* = 0.014) and *cis*-enQTL (1.51-fold, *s.e.* = 0.017; Fig. 5a and Supplementary Fig. 14f). Averaging across 198 traits, approximately half of the heritability was mediated by PCG expression and alternative splicing, followed by exon expression (46.4%), enhancer expression (29.5%) and lncRNA expression (28.5%; Supplementary Fig. 14g). The amounts of heritability of complex traits explained by molQTL were higher than those explained by MAF-matched random SNPs (Fig. 5b and Supplementary Fig. 14h).

Furthermore, we employed four complementary approaches to detect shared regulatory variants/genes associated with both molecular phenotypes and complex traits, including colocalization via fastENLOC²³, Mendelian randomization via SMR²⁴, single-tissue transcriptome-wide association studies (TWAS) via S-PrediXcan²⁵ and multi-tissue TWAS via S-MultiXcan²⁶. Of 1,507 significant GWAS loci that were tested in the *cis*-eQTL mapping, 983 (65%) were interpreted with *cis*-eQTL in at least one tissue (Fig. 5c and Supplementary Table 19). Among them, only 33% were colocalized with the nearest genes of the lead GWAS SNP (Fig. 5d). GWAS loci mapped with higher significance levels were more likely to be colocalized with *cis*-eQTL (Fig. 5e). The eGenes shared by more tissues tended to be colocalized with more GWAS loci (Fig. 5f). The number of colocalization events of a trait was determined by the statistical power of both GWAS and *cis*-eQTL mapping (Supplementary Fig. 14i–o).

To prioritize tissues relevant for complex trait variation, we defined a ‘tissue relevance score’ through the number of colocalization events adjusted by sample size and eGene discovery ratio of a tissue (Supplementary Table 20). We only considered 14 tissues with over 100 samples and found that, for instance, the ileum was the most relevant tissue for both average daily gain (ADG) and loin muscle area (Fig. 5g). For instance, *ABCD4* was the top associated gene in the small intestine TWAS of the average backfat thickness (BFT; Fig. 5h). It also had a significant association with BFT in the brain. The GWAS loci of BFT were colocalized with *cis*-eQTL of *ABCD4* in both the brain and small intestine. Although these lead SNPs were different in these two tissues, they had a relatively high LD ($r^2 = 0.71$), potentially tagging the same underlying causal variant. The fine-mapped SNP (rs1114012229) of the BFT GWAS was in a high LD ($r^2 = 0.85$) with the fine-mapped SNP (rs1107405934) of the *ABCD4* eQTL (Supplementary Fig. 15a). In addition, rs1107405934 was specifically associated with the expression of *ABCD4* in both intestinal tissues and the brain (Supplementary Fig. 15b, c).

Furthermore, we employed the same GWAS integrative analysis for other molQTL (Supplementary Tables 21–24). Around

80% (1,204/1,507) of significant GWAS loci could be explained by at least one molQTL in the 34 tissues. Of note, 8.2%, 3.8%, 3.5%, 1.9% and 0.4% of all 1,507 GWAS loci were only explained by *cis*-eQTL, *cis*-sQTL, *cis*-eeQTL, *cis*-lncQTL and *cis*-enQTL, respectively (Extended Data Fig. 10a, b). For example, a GWAS signal of ADG on chromosome 13 was only colocalized with *cis*-eQTL of *CFAP298-TCPIOL* in the colon, but not with its *cis*-sQTL or *cis*-eeQTL (Extended Data Fig. 10c). The GWAS signal for BFT on chromosome 15 was exclusively colocalized with *cis*-sQTL of *MYO7B* in small intestine, while the GWAS signal of litter weight was exclusively colocalized with *cis*-eeQTL of *FBXL12* in uterus (Extended Data Fig. 10d–e). In addition, 63% of GWAS loci were colocalized with more than one type of molQTL (Extended Data Fig. 10a and Supplementary Fig. 16). In addition, we detected 512 lncRNA-PCG-trait trios with significant pleiotropic associations (Supplementary Table 25 and Extended Data Fig. 10f).

The shared genetic regulation between humans and pigs

By examining GTEx (v8) in humans⁷, we found that one-to-one orthologous genes ($n = 15,944$) contributed to an average of 82% and 87% of overall expression across 17 common tissues in pigs and humans, respectively (Supplementary Fig. 17a, b). The visualization of variation in gene expression among all 12,453 samples clearly recapitulated tissue types rather than species (Supplementary Fig. 17c–h). The number of tissues in which an eGene was active was correlated between species (Supplementary Fig. 17i). The eGenes in a pig tissue generally had a higher enrichment for eGenes in the matching tissue in humans compared to other tissues (Fig. 6a). Furthermore, we observed a significant correlation ($r = 0.56$) of averaged eQTL effect between humans and pigs (Fig. 6b), which was higher than that ($r = 0.24$) observed between humans and rats previously²¹. In general, matching tissues had a higher correlation of eQTL effect compared to nonmatching tissues (Supplementary Fig. 18a, b and Supplementary Table 26). We observed a significant but weak correlation ($r = 0.09$) of *cis*- h^2 between humans and pigs (Supplementary Fig. 18c), similar to that between humans and rats ($r = 0.10$)²¹. In addition, tissue-specific expression of genes was more similar between pigs and humans than that between cattle and humans (Supplementary Fig. 19a–c). Similarly, the eQTL effects of orthologous genes in pigs were more correlated with those in humans than with those in cattle (Supplementary Fig. 19d–f).

We divided orthologous genes into four groups (that is, ‘neither’, ‘human-specific’, ‘pig-specific’ and ‘shared’) in each of the 17 matching tissues and observed a significant difference in expression levels among them. The shared eGenes had a lower tissue specificity in expression levels and regulatory effects, compared to genes in the other three groups (Fig. 6c and Supplementary Fig. 18d). A total of 783 eGenes were active in all tissues in both species, which were substantially enriched in metabolic processes (Supplementary Table 27). A total of

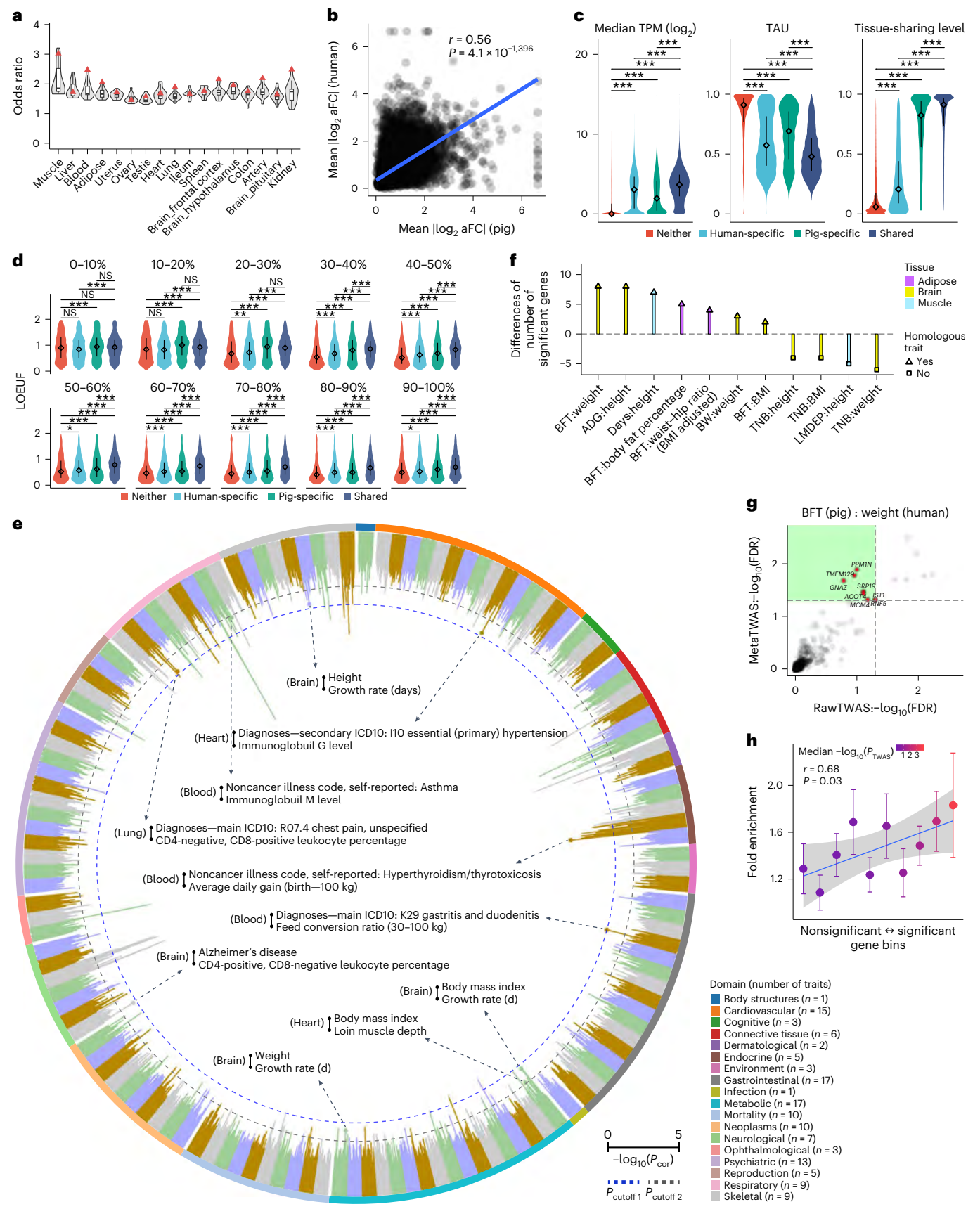
Fig. 6 | Conservation of gene expression, *cis*-eQTL and complex trait

genetics between pigs and humans. **a**, Enrichment (Fisher’s exact test) of pig eGenes with human eGenes across 17 matching tissues. Red triangles: matching tissues. **b**, Pearson’s correlation of eQTL effect size in orthologous genes ($n = 15,944$) between pigs and humans. **c**, Expression levels, TAU values and tissue-sharing levels for four groups of orthologous genes across 17 tissues in pigs. Neither, 3,993 non-eGenes in both species; human-specific, 8,174 eGenes; pig-specific, 3,882 eGenes; shared, 10,574 eGenes in both species. Two-sided Wilcoxon rank-sum test, *** $P < 0.001$. Diamond, median; error bar, upper/lower quartiles. **d**, LOEUF in the four groups of orthologous genes in ten evenly spaced expression level bins. One-sided Wilcoxon rank-sum test, NS $P > 0.05$, * $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$. The diamond and error bar are the same as in **c**. **e**, Significance ($-\log_{10}(P)$) of Pearson’s r of orthologous gene effect size between pig ($n = 268$) and human ($n = 136$) traits derived from TWAS. Each bar represents a pig–human trait pair in the same tissue ($n = 11$) and the within-domain blocks

of color correspond to different human traits. The number of tested genes for each of the pairs is shown in Supplementary Table 30. The text in the middle of the circle represents the significant examples of pig–human trait pairs in different thresholds. For each example, it includes human trait (top), pig trait (bottom) and TWAS tissue (left). P_{cutoff1} : FDR $< 10\%$ across all tested combinations. P_{cutoff2} : Bonferroni-corrected $P < 5\%$ within each trait–tissue pair of humans. **f**, Differences in the number of significant genes (FDR $< 5\%$) from cross-species (pig and human) meta-TWAS, compared to those from human TWAS. Supplementary Tables 18 and 29 present a detailed description of pig traits and human traits, respectively. **g**, FDR of discovered genes in human TWAS (RawTWAS) and cross-species meta-TWAS in the brain for BFT (pig) and weight (human). **h**, Pearson’s r between TWAS significances (color bar) of genes in pig BFT and their heritability enrichments (mean \pm *s.e.*) in human weight. The orthologous genes were divided into ten evenly spaced bins by sorting the P values of TWAS in the brain of pig BFT. Shading: standard error of the fitting line.

194 genes were not eGenes in any tissues in both species, and these were substantially enriched in essential biological functions (Supplementary Table 28). Expression levels of genes were negatively correlated with LOEUF scores, which was consistent across the four groups of

genes (Supplementary Fig. 18e). Among them, ‘Shared’ eGenes had the weakest negative correlation of expression levels and LOEUF scores, while ‘neither’ eGenes had the strongest negative correlation (Supplementary Fig. 18e). Of specific note, although they had the



highest expression levels, ‘Shared’ eGenes showed the strongest tolerance to loss of function mutations among the four gene groups (Fig. 6d). Compared to other genes, eGenes shared in both species had the lowest evolutionary DNA sequence constraints, whereas shared non-eGenes showed the opposite trend (Supplementary Fig. 18f). The expression levels of most genes were weakly or even not correlated with their PhastCons scores, eQTL detection and *cis-h*² estimates across tissues (Supplementary Fig. 18g–i).

To investigate whether the regulatory mechanism of complex phenotypes was conserved between humans and pigs, we compared the effect sizes of orthologous genes between 268 pig and 136 human complex phenotypes based on the summary statistics of TWAS (Supplementary Table 29). We observed a clear deviation (Wilcoxon rank-sum test $P = 2.16 \times 10^{-62}$) of the observed *P* values of TWAS correlations from the permutation-based null distribution (Supplementary Fig. 20a), and a total of 89 pig–human trait pairs were significant (FDR < 0.1; Supplementary Table 30, Fig. 6e and Supplementary Fig. 20b–e). We then chose several well-recognized homologous trait pairs between humans and pigs to perform the meta-TWAS, with several nonhomologous trait pairs as negative controls. For homologous trait pairs, cross-species meta-TWAS improved the discovery of trait-associated genes in humans (Fig. 6f). For instance, cross-species meta-TWAS analysis of pig average BFT and human body weight (BW) revealed eight new genes (FDR < 0.05) associated with BW in humans (Fig. 6g). Based on GWAS of 3,302 traits in humans²⁷, genome-wide association studies (PheWAS) showed that five of these eight genes were associated with other BW-relevant traits, such as height, birth weight and BMI (Supplementary Table 31). Furthermore, gene groups with higher significance in the pig BFT TWAS showed a higher enrichment for heritability of human BW (Fig. 6h).

Discussion

The pilot PigGTEx offers a deep survey of genetic regulatory effects across a wide range of tissues, representing a substantial advance in the understanding of the gene regulation landscape in pigs. This multi-tissue catalog of regulatory variants further advances our understanding of biological mechanisms underlying complex traits of economic importance in pigs. On average, about 80% of GWAS loci tested in pigs are linked to candidate target genes by molQTL in the PigGTEx, comparable with 78% of GWAS loci linked by GTEx in humans⁷. The PigGTEx will eventually enhance genetic improvement programs through the development of advanced biology-driven genomic prediction models that depend on informative SNPs²⁸. We also demonstrate the level of similarity between pigs and humans in gene expression, gene regulation and complex trait genetics. This extensive comparison of the pig and human genomes at multiple biological levels will be instructive for deciding which human diseases and complex traits make the pig the most suitable animal model.

Although a fraction of regulatory effects are shared across tissues, we note that some tissues, like the testis and those from early developmental stages, are distinct from other primary tissues. Due to the differences in sample size and other biological factors (for example, breed and cell-type composition) across tissue types in the current phase of PigGTEx, underrepresented tissues at multiple developmental stages are still required to gain a more comprehensive view of tissue-specific gene regulation and to refine the tissue-trait map in pigs. To elucidate gene regulation at single-cell resolution, we conducted an exploratory analysis to discover cell-type-interaction regulatory effects through an in silico cell-type deconvolution¹⁸. The cieQTL identified for several cell types indicate that a vast majority of cell-type-specific *cis*-QTL remain to be detected^{29,30}. Compared to *cis*-eQTL, *trans*-eQTL often have smaller effect sizes and thus require hundreds of thousands of samples to be discovered^{22,31}. Although integrating multi-omics data provides insight into the molecular

mechanisms underlying regulatory variants, experimental follow-ups are necessary to functionally validate and characterize these regulatory variants at large scale^{32,33}.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-023-01585-7>.

References

1. Tibbs Cortes, L., Zhang, Z. & Yu, J. Status and prospects of genome-wide association studies in plants. *Plant Genome* **14**, e20077 (2021).
2. Hu, Z. L., Park, C. A. & Reecy, J. M. Bringing the animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res.* **50**, D956–D961 (2022).
3. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).
4. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
5. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
6. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).
7. Aguet, F. et al. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
8. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
9. Barbeira, A. N. et al. Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *Genome Biol.* **22**, 49 (2021).
10. Velez-Irizarry, D. et al. Genetic control of longissimus dorsi muscle gene expression variation and joint analysis with phenotypic quantitative trait loci in pigs. *BMC Genomics* **20**, 3 (2019).
11. Criado-Mesas, L. et al. Identification of eQTLs associated with lipid metabolism in longissimus dorsi muscle of pigs with different genetic backgrounds. *Sci. Rep.* **10**, 9845 (2020).
12. Liu, Y. et al. Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. *Genet. Sel. Evol.* **52**, 59 (2020).
13. Clark, E. L. et al. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biol.* **21**, 285 (2020).
14. Lunney, J. K. et al. Importance of the pig as a human biomedical model. *Sci. Transl. Med.* **13**, eabd5758 (2021).
15. Pan, Z. et al. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat. Commun.* **12**, 5848 (2021).
16. Gu, X. & Su, Z. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc. Natl Acad. Sci. USA* **104**, 2779–2784 (2007).
17. Taylor-Weiner, A. et al. Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).
18. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
19. De Saram, P., Iqbal, A., Murdoch, J. N. & Wilkinson, C. J. BCAP is a centriolar satellite protein and inhibitor of ciliogenesis. *J. Cell Sci.* **130**, 3360–3373 (2017).
20. Flynn, E. & Lappalainen, T. Functional characterization of genetic variant effects on expression. *Annu. Rev. Biomed. Data Sci.* **5**, 119–139 (2022).

21. Munro, D. et al. The regulatory landscape of multiple brain regions in outbred heterogeneous stock rats. *Nucleic Acids Res.* **50**, 10882–10895 (2022).
22. Vösa, U. et al. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
23. Pividori, M. et al. 2020 PhenomeXcan: mapping the genome to the phenome through the transcriptome. *Sci. Adv.* **6**, eaba2083.
24. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
25. Barbeira A. N. et al. 2018 Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9** (1825).
26. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* **15**, e1007889 (2019).
27. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
28. Xiang, R. et al. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. *Nat. Commun.* **12**, 860 (2021).
29. Schmiedel, B. J. et al. Single-cell eQTL analysis of activated T cell subsets reveals activation and cell type-dependent effects of disease-risk variants. *Sci. Immunol.* **7**, eabm2508 (2022).
30. Nathan, A. et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120–128 (2022).
31. Wong, E. S. et al. Interplay of *cis* and *trans* mechanisms driving transcription factor binding and gene expression evolution. *Nat. Commun.* **8**, 1092 (2017).
32. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
33. Freimer J. W. et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks *Nat. Genet.* **54** 1133–1144 .

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

Jinyan Teng^{1,54}, Yahui Gao^{1,2,3,54}, Hongwei Yin^{4,54}, Zhonghao Bai^{5,6,54}, Shuli Liu^{2,7,54}, Haonan Zeng^{1,54}, The PigGTEx Consortium*, Lijing Bai⁴, Zexi Cai⁵, Bingru Zhao⁸, Xiujin Li⁹, Zhiting Xu¹, Qing Lin¹, Zhangyuan Pan^{10,11}, Wenjing Yang^{8,10}, Xiaoshan Yu⁶, Dailu Guan¹⁰, Yali Hou¹², Brittney N. Keel¹³, Gary A. Rohrer¹³, Amanda K. Lindholm-Perry¹³, William T. Oliver¹³, Maria Ballester¹⁴, Daniel Crespo-Piazuelo¹⁴, Raquel Quintanilla¹⁴, Oriol Canela-Xandri⁶, Konrad Rawlik¹⁵, Charley Xia^{16,17}, Yuelin Yao^{6,18}, Qianyi Zhao⁴, Wenye Yao^{4,19}, Liu Yang⁴, Houcheng Li⁵, Huicong Zhang⁵, Wang Liao⁶, Tianshuo Chen⁶, Peter Karlsson-Mortensen²⁰, Merete Fredholm²⁰, Marcel Amills^{21,22}, Alex Clop^{21,23}, Elisabetta Giuffra²⁴, Jun Wu¹, Xiaodian Cai¹, Shuqi Diao¹, Xiangchun Pan¹, Chen Wei¹, Jinghui Li¹⁰, Hao Cheng¹⁰, Sheng Wang²⁵, Guosheng Su⁵, Goutam Sahana⁵, Mogens Sandø Lund⁵, Jack C. M. Dekkers²⁶, Luke Kramer²⁶, Christopher K. Tuggle²⁶, Ryan Corbett²⁶, Martien A. M. Groenen¹⁹, Ole Madsen¹⁹, Marta Godia^{19,21}, Dominique Rocha²⁴, Mathieu Charles²⁷, Cong-jun Li², Hubert Pausch²⁸, Xiaoxiang Hu²⁹, Laurent Frantz^{30,31}, Yonglun Luo^{32,33,34}, Lin Lin^{32,33}, Zhongyin Zhou²⁵, Zhe Zhang³⁵, Zitao Chen³⁵, Leilei Cui^{36,37,38}, Ruidong Xiang^{39,40}, Xia Shen^{41,42,43}, Pinghua Li⁴⁴, Ruihua Huang⁴⁴, Guoqing Tang⁴⁵, Mingzhou Li⁴⁵, Yunxiang Zhao⁴⁶, Guoqiang Yi⁴, Zhonglin Tang⁴, Jicai Jiang⁴⁷, Fuping Zhao¹¹, Xiaolong Yuan¹, Xiaohong Liu⁴⁸, Yaosheng Chen⁴⁸, Xuwen Xu⁴⁹, Shuhong Zhao⁴⁹, Pengju Zhao⁵⁰, Chris Haley^{6,51}, Huaijun Zhou¹⁰, Qishan Wang³⁵, Yuchun Pan³⁵, Xiangdong Ding⁸, Li Ma³, Jiaqi Li¹, Pau Navarro^{6,51}, Qin Zhang⁵², Bingjie Li⁵³, Albert Tenesa^{6,51}, Kui Li⁴, George E. Liu², Zhe Zhang¹ & Lingzhao Fang^{5,6}✉

¹State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University (SCAU), Guangzhou, China.

²Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service (ARS), U.S. Department of Agriculture (USDA), Beltsville, MD, USA. ³Department of Animal and Avian Sciences, University of Maryland, College Park, MD, USA.

⁴Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-Omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, China. ⁵Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark. ⁶MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, UK. ⁷School of Life Sciences, Westlake University, Hangzhou, China. ⁸College of Animal Science and Technology, China Agricultural University, Beijing, China.

⁹Guangdong Provincial Key Laboratory of Waterfowl Healthy Breeding, College of Animal Science and Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, China. ¹⁰Department of Animal Science, University of California, Davis, Davis, CA, USA. ¹¹Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China. ¹²Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing, China. ¹³ARS, USDA, U.S. Meat Animal Research Center, Clay Center, NE, USA. ¹⁴Animal Breeding and Genetics Programme,

Institut de Recerca i Tecnologia Agroalimentàries (IRTA), Torre Marimon, Caldes de Montbui, Spain. ¹⁵Baillie Gifford Pandemic Science Hub, University of Edinburgh, Edinburgh, UK. ¹⁶Lothian Birth Cohort studies, University of Edinburgh, Edinburgh, UK. ¹⁷Department of Psychology, University of Edinburgh, Edinburgh, UK. ¹⁸School of Informatics, The University of Edinburgh, Edinburgh, UK. ¹⁹Animal Breeding and Genomics, Wageningen University and Research, Wageningen, The Netherlands. ²⁰Animal Genetics, Bioinformatics and Breeding, Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark. ²¹Department of Animal Genetics, Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, Campus de la Universitat Autònoma de Barcelona, Bellaterra, Spain. ²²Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, Bellaterra, Spain. ²³Consejo Superior de Investigaciones Científicas, Barcelona, Spain. ²⁴Paris-Saclay University, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France. ²⁵State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ²⁶Department of Animal Science, Iowa State University, Ames, IA, USA. ²⁷Paris-Saclay University, INRAE, AgroParisTech, GABI, SIGENAE, Jouy-en-Josas, France. ²⁸Animal Genomics, ETH Zurich, Universitaetstrasse 2, Zurich, Switzerland. ²⁹State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing, China. ³⁰Palaeogenomics Group, Department of Veterinary Sciences, Ludwig Maximilian University, Munich, Germany. ³¹School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK. ³²Department of Biomedicine, Aarhus University, Aarhus, Denmark. ³³Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark. ³⁴Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Research, Qingdao, China. ³⁵Department of Animal Science, College of Animal Sciences, Zhejiang University, Hangzhou, China. ³⁶School of Life Sciences, Nanchang University, Nanchang, China. ³⁷Human Aging Research Institute and School of Life Science, Nanchang University, and Jiangxi Key Laboratory of Human Aging, Jiangxi, China. ³⁸UCL Genetics Institute, University College London, London, UK. ³⁹Faculty of Veterinary and Agricultural Science, The University of Melbourne, Parkville, Victoria, Australia. ⁴⁰Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria, Australia. ⁴¹State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China. ⁴²Center for Intelligent Medicine Research, Greater Bay Area Institute of Precision Medicine, Fudan University, Guangzhou, China. ⁴³Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh, UK. ⁴⁴Institute of Swine Science, Nanjing Agricultural University, Nanjing, China. ⁴⁵Farm Animal Genetic Resources Exploration and Innovation Key Laboratory of Sichuan Province, Sichuan Agricultural University, Chengdu, China. ⁴⁶College of Animal Science and Technology, Guangxi University, Nanning, China. ⁴⁷Department of Animal Science, North Carolina State University, Raleigh, NC, USA. ⁴⁸State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. ⁴⁹Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education and College of Animal Science and Technology, Huazhong Agricultural University, Wuhan, China. ⁵⁰Hainan Institute, Zhejiang University, Yongyou Industry Park, Yazhou Bay Sci-Tech City, Sanya, China. ⁵¹The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, UK. ⁵²College of Animal Science and Technology, Shandong Agricultural University, Tai'an, China. ⁵³Scotland's Rural College (SRUC), Roslin Institute Building, Midlothian, UK. ⁵⁴These authors contributed equally: Jinyan Teng, Yahui Gao, Hongwei Yin, Zhonghao Bai, Shuli Liu, Haonan Zeng. *A list of authors and their affiliations appears at the end of the paper. ✉ e-mail: albert.tenesa@ed.ac.uk; likui@caas.cn; george.liu@usda.gov; zhezhang@scau.edu.cn; lingzhao.fang@qgg.au.dk

The PigGTEx Consortium

Jinyan Teng^{1,54}, Yahui Gao^{1,2,3,54}, Hongwei Yin^{4,54}, Zhonghao Bai^{5,6,54}, Shuli Liu^{2,7,54}, Haonan Zeng^{1,54}, Lijing Bai⁴, Zexi Cai⁵, Bingru Zhao⁸, Xiujin Li⁹, Zhiting Xu¹, Qing Lin¹, Zhangyuan Pan^{10,11}, Wenjing Yang^{8,10}, Xiaoshan Yu⁶, Dailu Guan¹⁰, Yali Hou¹², Brittney N. Keel¹³, Gary A. Rohrer¹³, Amanda K. Lindholm-Perry¹³, William T. Oliver¹³, Maria Ballester¹⁴, Daniel Crespo-Piazuelo¹⁴, Raquel Quintanilla¹⁴, Oriol Canela-Xandri⁶, Konrad Rawlik⁵³, Charley Xia^{16,17}, Yuelin Yao^{6,18}, Qianyi Zhao⁴, Wenye Yao^{4,19}, Liu Yang⁴, Houcheng Li⁵, Huicong Zhang⁵, Wang Liao⁶, Tianshuo Chen⁶, Peter Karlskov-Mortensen²⁰, Merete Fredholm²⁰, Marcel Amills^{21,22}, Alex Clop^{21,23}, Elisabetta Giuffra²⁴, Jun Wu¹, Xiaodian Cai¹, Shuqi Diao¹, Xiangchun Pan¹, Chen Wei¹, Jinghui Li¹⁰, Hao Cheng¹⁰, Sheng Wang²⁵, Guosheng Su⁵, Goutam Sahana⁵, Mogens Sandø Lund⁵, Jack C. M. Dekkers²⁶, Luke Kramer²⁶, Christopher K. Tuggle²⁶, Ryan Corbett²⁶, Martien A. M. Groenen¹⁹, Ole Madsen¹⁹, Marta Godia^{19,21}, Dominique Rocha²⁴, Mathieu Charles⁵², Cong-jun Li², Hubert Pausch²⁷, Xiaoxiang Hu²⁸, Laurent Frantz^{29,30}, Yonglun Luo^{31,32,33}, Lin Lin^{31,32}, Zhongyin Zhou²⁵, Zhe Zhang³⁴, Zitao Chen³⁴, Leilei Cui^{35,36,37}, Ruidong Xiang^{38,39}, Xia Shen^{40,41,42}, Pinghua Li⁴³, Ruihua Huang⁴³, Guoqing Tang⁴⁴, Mingzhou Li⁴⁴, Yunxiang Zhao⁴⁵, Guoqiang Yi⁴, Zhonglin Tang⁴, Jicai Jiang⁴⁶, Fuping Zhao¹¹, Xiaolong Yuan¹, Xiaohong Liu⁴⁷, Yaosheng Chen⁴⁷, Xuewen Xu⁴⁸, Shuhong Zhao⁴⁸, Pengju Zhao⁴⁹, Chris Haley^{6,15}, Huaijun Zhou¹⁰, Qishan Wang³⁴, Yuchun Pan³⁴, Xiangdong Ding⁸, Li Ma³, Jiaqi Li¹, Pau Navarro^{6,15}, Qin Zhang⁵⁰, Bingjie Li⁵¹, Albert Tenesa^{6,15}, Kui Li⁴, George E. Liu², Zhe Zhang¹ & Lingzhao Fang^{5,6}

Methods

Ethics

It is not applicable because no biological samples were collected and no animal handling was performed for this study.

RNA-seq data analysis and molecular phenotype quantification

In total, we gathered 11,323 publicly accessible raw RNA-seq datasets, representing 9,530 distinct samples (downloaded from NCBI SRA by 26 February 2021), of which 98.13% were generated using the Illumina platform. We removed 121 embargoed RNA-seq samples and then processed all the remaining RNA-seq samples using a uniform pipeline. Briefly, we first trimmed adaptors and discarded reads with poor quality using Trimmomatic (v0.39)³⁴. We then aligned clean reads to the Sscrofa11.1 (v100) pig reference genome using STAR (v2.7.0)³⁵. We kept 8,262 samples with more than 500K clean reads and uniquely mapping rates $\geq 60\%$ for subsequent analysis (Supplementary Table 1). We extracted the raw read counts of 31,871 Ensembl (Sscrofa11.1 v100) genes by featureCounts (v1.5.2)³⁶ and obtained their normalized expression (that is, transcripts per million (TPM)) using Stringtie (v2.1.1)³⁷. We removed 544 samples in which less than 20% of all annotated genes were expressed (TPM ≥ 0.1), resulting in 7,597 samples. We then visualized the variance in gene expression among samples using *t*-distributed stochastic neighbor embedding (*t*-SNE)³⁸. After filtering out outliers within each of the tissues, we eventually kept 7,095 samples for subsequent analysis (Supplementary Table 1). We employed MEGA (vX)³⁹ to build a neighbor-joining tree of these samples based on TPM and then visualized it by iTOL (v6)⁴⁰.

For PCG expression, we considered 21,280 PCGs from the Ensembl annotation (Sscrofa11.1 v100). For exon expression of PCGs, we extracted raw read counts of 290,536 exons by featureCounts (v1.5.2)³⁶ and normalized them as TPM. To explore enhancer expression, we downloaded the previously predicted enhancers (strong active enhancers, EnhA) from 14 pig tissues¹⁵. We merged these enhancer regions across tissues using bedtools (v2.30.0)⁴¹, resulting in 158,998 nonredundant enhancer regions. To control the potential contamination of transcribed genes, we only focused on transcribed enhancers that were not overlapped with any known gene regions (including protein-coding gene, lncRNA, pseudogene, tRNA, miRNA and snoRNA)^{42–44}, resulting in 3,679 enhancers. We obtained raw read counts of these nonredundant enhancer regions from all 7,095 RNA-seq samples by featureCounts (v1.5.2)³⁶, followed by TPM normalization. For lncRNA expression, we obtained 17,162 lncRNAs predicted from 33 Iso-Seq datasets, representing ten tissues from four animals by using FEELnc⁴⁵. We applied the same approach to extract and normalize lncRNA expression as above.

For alternative splicing, we used Leafcutter (v0.2.9)⁴⁶ to quantify excision levels of introns and then to identify splicing events within each tissue as described in the following: (1) converting aligned bam files from STAR (v2.7.0) into junction files using the script bam2junc.sh; (2) generating intron clusters using the script leafcutter_cluster.py, and then mapping them to genes by the map_clusters_to_genes.R script with exon coordinates extracted from the Ensembl annotation file (v100); (3) discarding introns without any read count in more than 50% of samples or with fewer than $\max(10, 0.1n)$ unique values, where n is the sample size; (4) filtering out introns with low complexity: $\sum_i (|z_i| < 0.25) \geq n-3$ and $\sum_i (|z_i| > 6) \leq 3$, where z_i is the *i*th cluster read fraction across individuals; (5) using prepare_phenotype_table.py script to normalize filtered counts and convert them into BED format, where start/end positions correspond to the TSS of corresponding genes. Furthermore, we normalized excision levels of introns as percent spliced-in (PSI) values.

MolQTL mapping

For molQTL mapping within each of the 34 tissues, we only considered SNPs with MAF $\geq 5\%$ and minor allele count ≥ 6 , resulting in an average of 2,705,637 SNPs (ranging from 1,815,729 in synovial membrane to

3,004,852 in muscle). We computed genotype PCs based on the filtered SNPs within each of the tissues using SNPRelate (v1.26.0)⁴⁷. We used the top five and ten genotype PCs to account for the population structure among samples in tissues with <200 and ≥ 200 samples, respectively (Extended Data Fig. 2f). To account for technical confounders among RNA-seq samples, we used the probabilistic estimation of expression residual (PEER) method, implemented in PEER (v1.0) R package⁴⁸, to estimate a set of latent covariates within each tissue based on gene expression matrices. We obtained a total of 60 PEER factors in each tissue and assessed their relative contributions (that is, factor weight variance) to gene expression variation using the PEER_getAlpha function. We decided to use the top ten PEER factors for each tissue as covariates when conducting molQTL mapping for PCG, exon, lncRNA and enhancer expression (Extended Data Fig. 2g). For *cis*-sQTL mapping, we estimated and fitted ten PEER factors from the splicing quantifications of genes within each tissue. To understand whether known covariates can be captured by PEER factors, we fitted a linear regression model to estimate the proportion of variance in known confounders that were explained by the top ten PEER factors.

For *cis*-eQTL mapping, we first normalized the PCGs expression across samples within each tissue using the trimmed mean of M-value (TMM) method, implemented in edgeR⁴⁹, followed by inverse normal transformation of the TMM. We performed *cis*-eQTL mapping using a linear regression model, implemented in TensorQTL (v1.0.3)¹⁷, while accounting for the estimated covariates. Within each tissue, we filtered out genes with TPM < 0.1 and/or raw read counts < 6 in more than 80% of samples. We defined the *cis*-window of PCG as ± 1 Mb of TSS and obtained the nominal *P* values of *cis*-eQTL with the parameter mode *cis_nominal* in TensorQTL. We then employed two layers of multiple testing corrections based on the permutation approach⁵⁰, implemented in the TensorQTL. In the first layer, we applied an adaptive permutation approach to calculate the empirical *P* values of variants within each gene and obtained the permutation *P* value of the lead variant for each gene. In the second layer, we conducted the Benjamini–Hochberg correction for the permutation *P* values of lead variants across all tested genes and considered genes with FDR $< 5\%$ as the genome-wide significant eGenes and genes without significant *cis*-eQTL as non-eGenes. To identify significant *cis*-eQTL associated with eGenes, we defined the empirical *P* value of the gene that was closest to an FDR of 0.05 as the genome-wide empirical *P* value threshold (pt). We obtained the gene-level threshold for each gene from the beta distribution by qbeta (pt, beta_shape1, beta_shape2) in R (v4.0.2), where beta_shape1 and beta_shape2 were derived using TensorQTL. We considered SNPs with a nominal *P* value below the gene-level threshold as significant *cis*-eQTL for a given gene–tissue pair.

Similarly, we normalized the expression of exons, lncRNAs and enhancers to inverse normal transformed TMM across samples and excluded lowly expressed elements using the same approach as for PCG. We conducted *cis*-QTL mapping for exons (*cis*-eeQTL), lncRNAs (*cis*-lncQTL) and enhancers (*cis*-enQTL) using TensorQTL. For *cis*-eeQTL mapping, we defined the *cis*-window of an exon as the ± 1 Mb region of its source gene's TSS. For exons, lncRNA and enhancer *cis*-QTL mapping, we defined the *cis*-window as the ± 1 Mb region of the TSS of the source gene, of its TSS and its TSS, respectively. We declared significant *cis*-QTL for exons, lncRNAs and enhancers using the same approach as done for the *cis*-eQTL mapping. We defined exons, lncRNAs and enhancers with at least one significant *cis*-QTL as eExon, eLncRNA and eEnhancer, respectively.

We performed *cis*-sQTL mapping for genes with splicing quantifications (PSI values) and tested SNPs within ± 1 Mb of TSS using TensorQTL (v1.0.3)¹⁷ while accounting for the estimated covariates. To compute the empirical *P* value of *cis*-sQTL, we grouped all intron clusters of a gene with the parameter: --phenotype_groups option in the permutation mode of TensorQTL (v1.0.3)¹⁷. We defined sGene and significant *cis*-sQTL using the same approach as used for *cis*-eQTL

mapping. We refer to the eGene, eExon, eLncRNA and eEnhancer above, as well as sGene collectively as eMolecule.

Conditionally independent molQTL mapping

To identify the multiple independent *cis*-QTL signals of a given eMolecular, we applied a forward-backward stepwise regression approach⁷, using TensorQTL (v1.0.3) with the parameter: `--mode cis-independent`¹⁷. We set the gene-level significance threshold to be the maximum β -adjusted *P* value for eMolecules within each tissue after correcting for multiple testing as described above. At each iteration, we scanned the new *cis*-QTL after adjusting for all previously discovered *cis*-QTL and covariates. In addition, we further employed SuSiE-inf (v1.2)⁵¹ to fine-map the potential causal *cis*-QTL for each eMolecule.

The tissue-sharing patterns of molQTL

To understand the shared or specific genetic regulatory mechanisms between tissues, we performed a meta-analysis of molQTL across all 34 tissues using MashR (v0.2–6)⁵² and METASOFT (v2.0.1)⁵³ as described above. For MashR (v0.2–6), we only considered the *z* scores from TensorQTL (v1.0.3; slope/slope_se) of the top *cis*-molQTL. We obtained the estimated effect sizes (that is, posterior means) and the corresponding significance levels (that is, local false sign rate (LFSR)) from the mash function. We defined a molQTL with LFSR < 0.05 as active in a given tissue. To estimate the pairwise tissue similarity with regard to genetic regulation of gene expression, we calculated the pairwise Spearman's correlation of effect size estimates of *cis*-molQTL between any tissue pairs, focusing on SNPs with LFSR < 0.05 in at least one tissue. For METASOFT (v2.0.1), we used summary statistics (that is, slope and slope_se) from TensorQTL (v1.0.3) of molQTL across all tissues. We estimated the meta-analytic effect size using a fixed effect model and calculated *M* values (posterior probabilities) using the MCMC method. We considered a molQTL with *M* > 0.7 active in tissue. To evaluate the similarity of tissue-clustering patterns across different data types (that is, PCG expression, splicing quantifications, exon expression, lncRNA expression, enhancer expression, *cis*-eQTL, *cis*-sQTL, *cis*-lncQTL, *cis*-eeQTL and *cis*-enQTL), we performed *k*-means clustering using the *k*-means function in the stats R package (v4.0.2), in which parameter *k* was allowed to range from 2 to 20 and the maximum number of iterations was 1,000,000. We calculated the pairwise Rand index to measure the clustering similarity using the rand.index function in the fossil (v0.4.0) R package (v4.0.2)⁵⁴.

GWAS summary statistics

To investigate the regulatory mechanisms underpinning complex traits in pigs, we systematically integrated the identified molQTL with summary statistics of 268 meta-GWAS from 207 complex traits of economic importance, representing five trait domains (Supplementary Table 18). In total, we performed 2,056 separate GWAS and conducted the meta-GWAS analysis for the same traits across different populations based on GWAS summary statistics using METAL (v2011-03-25)⁵⁵, resulting in 268 meta-GWAS results. To perform the integrative analysis of GWAS and molQTL, we overlapped significant GWAS loci with the 3,087,268 SNPs tested in the molQTL mapping, resulting in 1,507 GWAS loci with lead SNP *P* < 1×10^{-5} .

Enrichment of molQTL and trait-associated variants

To examine whether molQTL was enriched among the significant GWAS variants, we applied three distinct approaches as described in the following. First, we used a simple overlapping approach to examine whether a significant molQTL is more likely to be a significant trait-SNP as described in ref. 9 Briefly, for each tissue, we kept SNPs with the most significant nominal *P* value for a gene and scaled *P* values to a comparable level ($\lambda = 10$) across 34 tissues. We selected the minimum *P* value of each SNP in the 34 tissues as the background set, from which we extracted *P* values for SNPs that overlapped with significant GWAS loci.

Second, we applied QTLEnrich (v2)⁷ to quantify the enrichment degree between significant molQTL and GWAS loci. We only used summary statistics of 198 GWAS for which $\geq 80\%$ of SNPs were also tested in the molQTL mapping. Third, we applied the mediated expression score regression method to estimate the heritability of complex trait that was mediated by the *cis*-genetic component of different molecular phenotypes (h^2_{med})⁵⁶.

Cis-molQTL-GWAS colocalization

To identify shared genetic variants between the molecular phenotypes and complex traits, we performed a colocalization analysis of molQTL and GWAS loci using fastENLOC (v1.0)²³. Briefly, we obtained the probabilistic annotation of molQTL from the DAP-G (v1.0.0)⁵⁷ and then used the summarize_dap2enloc.pl script to generate the annotation file of multi-tissue molQTLs. We generated approximate LD blocks across the entire genome based on the PGRP using PLINK (v1.90)⁵⁸. We applied the TORUS tool to generate the posterior inclusion probability of each LD block based on GWAS *z* scores⁵⁹, followed by the colocalization analysis with fastENLOC (v1.0). We obtained the regional colocalization probability (RCP) of each LD-independent genomic region using a natural Bayesian hierarchical model⁶⁰ and considered a gene with RCP > 0.9 as significant. To identify the trait-relevant tissues, we calculated a 'relevance score' between a tissue and a trait by dividing the number of colocalized genes by the product of sample size and eGene proportion in this tissue. We only considered 14 tissues with ≥ 100 samples.

TWAS of complex traits

To explore whether the overall *cis*-genetic component of a molecular phenotype is associated with complex traits, we conducted single- and multi-tissue TWAS using S-PrediXcan²⁵ and S-MultiXcan in MetaXcan (v0.6.11)²⁶, respectively, based on the summary statistics of the meta-GWAS. Briefly, we employed the nested cross-validated elastic net model implemented in S-PrediXcan to predict the five types of molecular phenotypes in all 34 tissues. To train the predictive model, we used the confounder-corrected expression or PSI values as phenotypes and SNPs within the *cis*-windows of genes as genotypes. We kept only predictive models with cross-validated correlation $\rho > 0.1$ and prediction performance *P* < 0.05 for further TWAS analysis. We ran S-PrediXcan on all 268 GWAS to obtain gene–trait associations at a single-tissue level. Based on results from S-PrediXcan, we ran S-MultiXcan to integrate predictions from multiple tissues, yielding the multi-tissue TWAS results. We applied Bonferroni correction and considered a corrected *P* < 0.05 as significant.

MR analysis between molQTL and GWAS loci

We conducted MR analysis to infer the causality between molecular phenotypes and complex traits using the SMR (v1.03)²⁴. We first converted the summary statistics of molQTL from TensorQTL (v1.0.3) to BESD format using SMR with the options: `--fastqtl-nominal-format --make-besd`. We only considered eMolecules with top nominal *P* value < 1×10^{-5} for the SMR test. We defined gene–trait pairs to pass the SMR test if the Benjamini–Hochberg-adjusted *PSMR* < 0.05 and *PGWAS* < 1×10^{-5} . For gene–trait pairs that passed the SMR test, we performed the heterogeneity in dependent instruments (HEIDI) test, with $P_{\text{HEIDI}} \geq 0.05$ reflecting that we could not reject a single causal variant with effects on both molecular phenotype and complex trait. As a *cis*-regulator, lncRNA can regulate the expression of neighboring PCGs and then can influence complex traits. To understand this etiology of complex traits, we performed an integrative SMR analysis that used three layers of summary-level information from *cis*-lncQTL, *cis*-eQTL and GWAS. We used the summary statistics of *cis*-lncQTL and *cis*-eQTL as the exposure and the outcome input for SMR (v1.03)⁶¹, respectively, which detected pleiotropic effects between lncRNA and PCG expression. We used Bonferroni correction within each tissue and defined a corrected *P* < 0.05 as significant.

Comparative analysis between pigs and humans

To explore the genetic similarity of complex traits between pigs and humans, we performed a comparative analysis of TWAS summary statistics. We downloaded public human GWAS summary statistics for 136 complex traits, representing 18 trait domains (Supplementary Table 29). Based on the predictive models in human GTEx v8 (ref. 62), we applied the S-PrediXcan to conduct TWAS for all 136 complex traits across 49 human tissues. We only kept TWAS results from 11 major tissues in humans that had matched tissues with ≥ 100 samples in pigs. We only considered 15,944 one-to-one orthologous genes. For a trait pair, we calculated the Pearson's correlation of absolute effect size estimated of orthologous genes between pigs and humans within the matching tissue. We applied Benjamini–Hochberg correction for P values of all tested correlations and defined an FDR $< 10\%$ as significant. To investigate whether GTEx-like resources can facilitate cross-species gene mapping of complex traits through borrowing 'information' at the level of orthologous genes instead of individual variants, we performed a cross-species meta-TWAS analysis through modifying a multi-ancestry meta-TWAS method in humans⁶³. We calculated the z statistics of meta-TWAS as follows: $z_{\text{meta}} = \frac{n_i z_{\text{TWAS},i} + n_j z_{\text{TWAS},j}}{\sqrt{n_i^2 + n_j^2}}$,

where $z_{\text{TWAS},i}$ and $z_{\text{TWAS},j}$ were the z statistics from pig TWAS and human TWAS results, respectively; n_i and n_j were the population size of pig TWAS and human TWAS, respectively. If the tested trait is a case–control study, we adjusted the sample size as $4/(\frac{1}{n_{\text{cases}}} + \frac{1}{n_{\text{controls}}})$. We chose

several well-recognized homologous trait pairs between humans and pigs to perform the meta-TWAS, and we also selected several nonhomologous trait pairs as negative controls. We divided orthologous genes into ten bins sorted by P values of pig TWAS and estimated the heritability enrichment of different gene bins in homologous trait of humans using LD score regression implemented in LDSC⁶⁴. We performed the PheWAS based on 4,756 GWAS, including 3,302 traits in GWAS ATLAS²⁷.

Statistics and reproducibility

No statistical method was used to predetermine the sample size. The details of data exclusions for each specific analysis are available in the Methods section. For all the boxplots, the horizontal lines inside the boxes show the medians. Box bounds show the lower quartile ($Q1$, the 25th percentile) and the upper quartile ($Q3$, the 75th percentile). Whiskers are minima ($Q1 - 1.5 \times \text{IQR}$) and maxima ($Q3 + 1.5 \times \text{IQR}$), where IQR is the interquartile range ($Q3 - Q1$). Outliers are shown in the boxplots unless otherwise stated. The experiments were not randomized, as all the datasets are publicly available from observational studies. The investigators were not blinded to allocation during experiments and outcome assessment, as the data were not from controlled randomized studies.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All raw data analyzed in this study are publicly available for download without restrictions from SRA (<https://www.ncbi.nlm.nih.gov/sra/>) and BIGD (<https://bigd.big.ac.cn/bioproject/>) databases. Details of RNA-seq, WGS, WGBS, single-cell RNA-seq and Hi-C datasets can be found in Supplementary Tables 1, 2, 5, 8 and 9, respectively. All the WGS data newly generated in this study are available under CNCB GSA (<https://ngdc.cncb.ac.cn/>) accessions PRJCA016120, PRJCA016130, PRJCA017284, PRJCA016012 and PRJCA016216. All processed data and the full summary statistics of molQTL mapping are available at <http://piggtex.farmgtex.org/>.

Code availability

All the computational scripts and codes for RNA-seq, WGS, WGBS, single-cell RNA-seq and Hi-C dataset analyses, as well as the respective

quality control, molecular phenotype normalization, genotype imputation, molQTL mapping, functional enrichment, colocalization, SMR and TWAS, are available at the FarmGTEx GitHub website (<https://github.com/FarmGTEx/PigGTEx-Pipeline-v0>, <https://doi.org/10.6084/m9.figshare.24247771>)⁶⁵.

References

- Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
- Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
- Van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
- Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
- Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).
- Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- Chen, H. et al. A Pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* **173**, 386–399 (2018).
- Ren, B. Enhancers make non-coding RNA. *Nature* **465**, 173–174 (2010).
- Zhang, Z. et al. HeRA: an atlas of enhancer RNAs across human tissues. *Nucleic Acids Res.* **49**, D932–D938 (2021).
- Wucher, V. et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* **45**, e57 (2017).
- Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
- Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
- Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
- Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
- Aguet, F. et al. Molecular quantitative trait loci. *Nat. Rev. Methods Primers* **3**, 4 (2023).
- Cui, R. et al. Improving fine-mapping by modeling infinitesimal effects. Preprint at *bioRxiv* 10.1101/2022.10.21.513123 (2022).
- Urbat, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
- Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet.* **8**, e1002555 (2012).
- Vavrek, M. J. Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontol. Electron.* **14**, 16 (2011).
- Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
- Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

57. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
58. Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Science*. **296**, 2225–2229 (2002).
59. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).
60. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, 1–25 (2017).
61. Wu, Y. et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* **9**, 918 (2018).
62. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
63. Bhattacharya, A. et al. Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: lessons from the Global Biobank Meta-analysis Initiative. *Cell Genomics* **2**, 100180 (2022).
64. Bulik-Sullivan, B. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
65. Teng, J. & FarmGTEx. FarmGTEx/PigGTEx-Pipeline-v0. GitHub. github.com/FarmGTEx/PigGTEx-Pipeline-v0 (2023).

Acknowledgements

Zhe Zhang (SCAU) acknowledges funding from the National Natural Science Foundation of China (32022078), the National Key R&D Program of China (2022YFF1000900) and the Local Innovative and Research Teams Project of Guangdong Province (2019BT02N630), and support from National Supercomputer Center in Guangzhou, China. Y.C., Zhe Zhang (SCAU), Jiaqi Li, X. Liu, X.D. and S.Z. acknowledge funding from the China Agriculture Research System (CARS-35). L. Fang acknowledges funding from HDR-UK under award HDR-9004 and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement 801215. G.E.L. was supported by United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) AFRI under grant numbers 2019-67015-29321 and 2021-67015-33409 and the appropriated project 8042-31000-112-00-D, 'Accelerating Genetic Improvement of Ruminants Through Enhanced Genome Assembly, Annotation, and Selection' of the USDA Agricultural Research Service (ARS). This research used resources provided by the SCINet project of the USDA ARS under project 0500-00093-001-00-D. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer. A.T. acknowledges funding from the BBSRC through program grants BBS/E/D/10002070 and BBS/E/D/30002275, MRC research grant MR/P015514/1 and HDR-UK award HDR-9004. P.N. and C.H. were supported by the Medical Research Council, UK (grant MC_UU_00007/10). O.C.-X. was supported by MR/R025851/1. M.B. and D.C.-P. belonged to a Consolidated Research Group AGAUR, ref. 2017SGR-1719, and D.C.-P. was supported by the GENE-SWITCH project (<https://www.gene-switch.eu>), which is funded by the European Union's Horizon 2020 research and innovation program under the grant agreement 817998. R.X. was supported by the Australian Research Council's Discovery Projects (DP200100499). L.M. was supported in part by AFRI under grants 2020-67015-31398 and 2021-67015-33409 from the USDA NIFA. B.N.K. and G.A.R. were supported by appropriated project 3040-31000-099-000-D, 'Identifying

Genomic Solutions to Improve Efficiency of Swine Production' of the ARS of the USDA. A.K.L.P. and W.T.O. were supported by appropriated project 3040-31000-102-000-D, 'Optimizing Nutrient Management and Efficiency of Beef Cattle and Swine' of the ARS of the USDA. Z.P., D.G. and H. Zhou, and computational resource were supported in part by Agriculture and Food Research Initiative Competitive grants 2018-67015-27501 and 2015-67015-22940. All the funders had no role in study design, data collection and analysis and decision to publish or prepare the manuscript.

We thank all the researchers who have contributed to the publicly available data used in this research. We thank the valuable comments and suggestions from D. Speed, G. Paul Ramstein (QGG, Aarhus University, Denmark), M. E. Goddard (The University of Melbourne, Australia), C. Ponting (IGC, The University of Edinburgh, UK) and G. Larson (The University of Oxford, UK). Figure 1d was created with BioRender.com. For the purpose of open access, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission.

Author contributions

L. Fang, Zhe Zhang (SCAU), G.E.L., A.T. and K.L. conceived and designed the project. Y.G., S.L., X. Li, H.Y., B.Z., W. Yang, W. Yao, Y.Y., H.L., H. Zhang and X.P. performed bioinformatic analyses of RNA-seq data analysis. H.Y., S.D., L.B., S.W., D.G., L.Y. and Z.Chen conducted whole-genome sequence data analysis. Y.G., Q. Zhao and Z.P. performed omics data analysis. J.T. conducted genotype imputation and molQTL mapping. Z.X., H. Zeng, C.W., W.L., T.C. and X. Yu prepared the summary statistics of GWAS in pigs and humans. J.T., Q.L., X.C. and J.W. integrated molQTL with GWAS. Z.B., J.T., C.X. and Jinghui Li led the comparison of PigGTEx and human GTEx. B.N.K., G.A.R., A.K.L.P., W.T.O., M.B., D.C., M.C. and L.K. contributed to the validation and functional annotation of molQTL. P.N., Y.H., B.L., Z. Cai, P.Z., D.R., C.L., H.P., X.H., L. Frantz, Y.L., L.L., L.C., J.J., R.H., Z.T., M.L., S.Z. and Y.C. contributed to the critical interpretation of analytical results before and during manuscript preparation. H. Zeng, J.T., Zhe Zhang (SCAU) and L. Fang built the PigGTEx web portal. L. Fang, Zhe Zhang (SCAU), G.E.L., K.L., M.B., R.Q., O.C.-X., K.R., P.K.M., M.F., M.A., A.C., E.G., H.C., G. Su, G. Sahana, M.S.L., J.C.M.D., C.K.T., R.C., M.A.M.G., O.M., M.G., Z. Zhou, Z. Zhang, R.X., X.S., P.L., G.T., Y.Z., G.Y., F.Z., P.N., X. Yuan, X. Liu, L.M., H.S., X.X., Q.W., X.D., H. Zhou, Jiaqi Li, C.H., Y.P., B.L. and Q. Zhang contributed to the data and computational resources. L. Fang, J.T., Y.G. and Z.B. drafted the manuscript. All authors read, edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

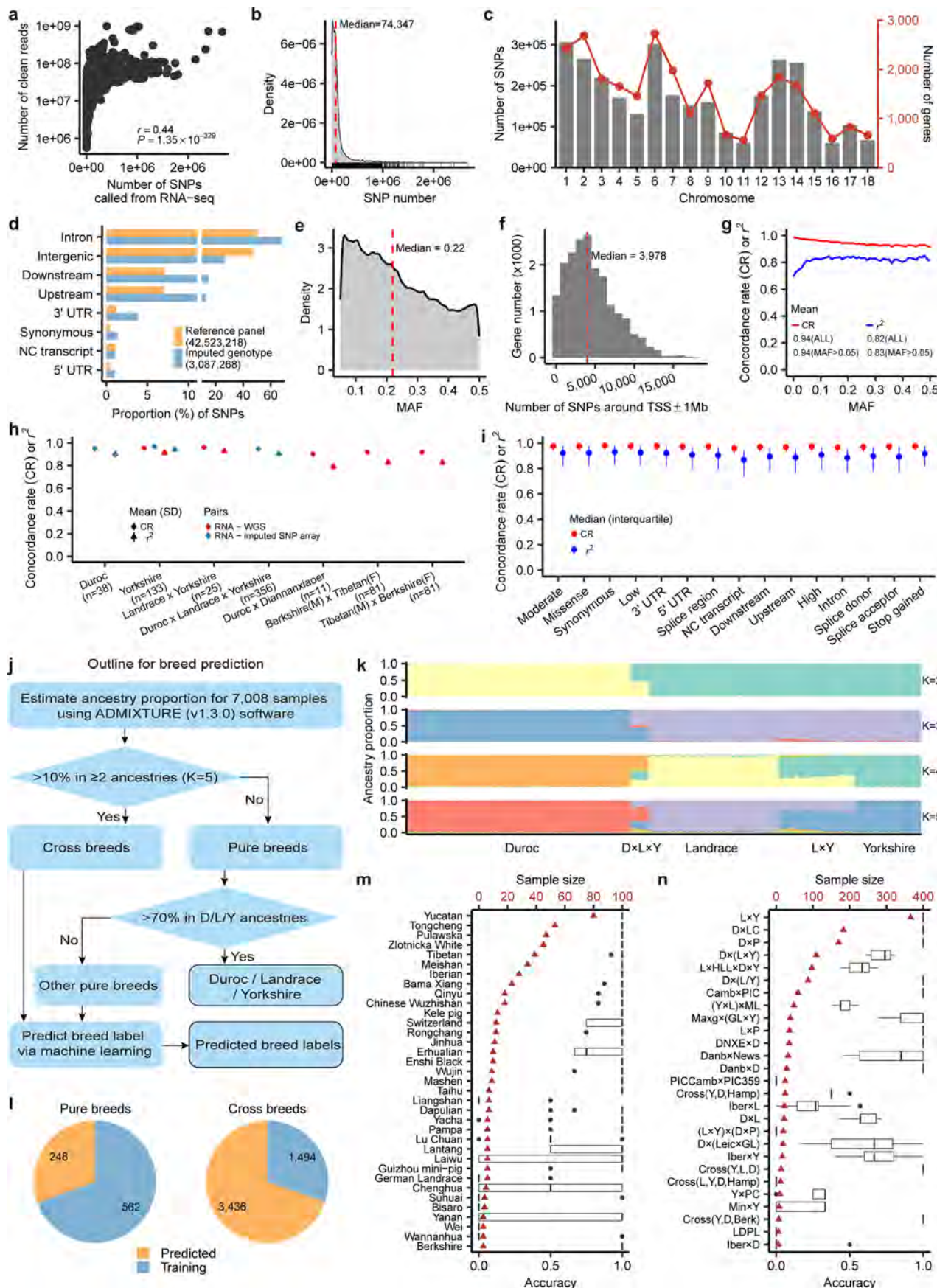
Extended data is available for this paper at <https://doi.org/10.1038/s41588-023-01585-7>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-023-01585-7>.

Correspondence and requests for materials should be addressed to Albert Tenesa, Kui Li, George E. Liu, Zhe Zhang or Lingzhao Fang.

Peer review information *Nature Genetics* thanks Wei Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

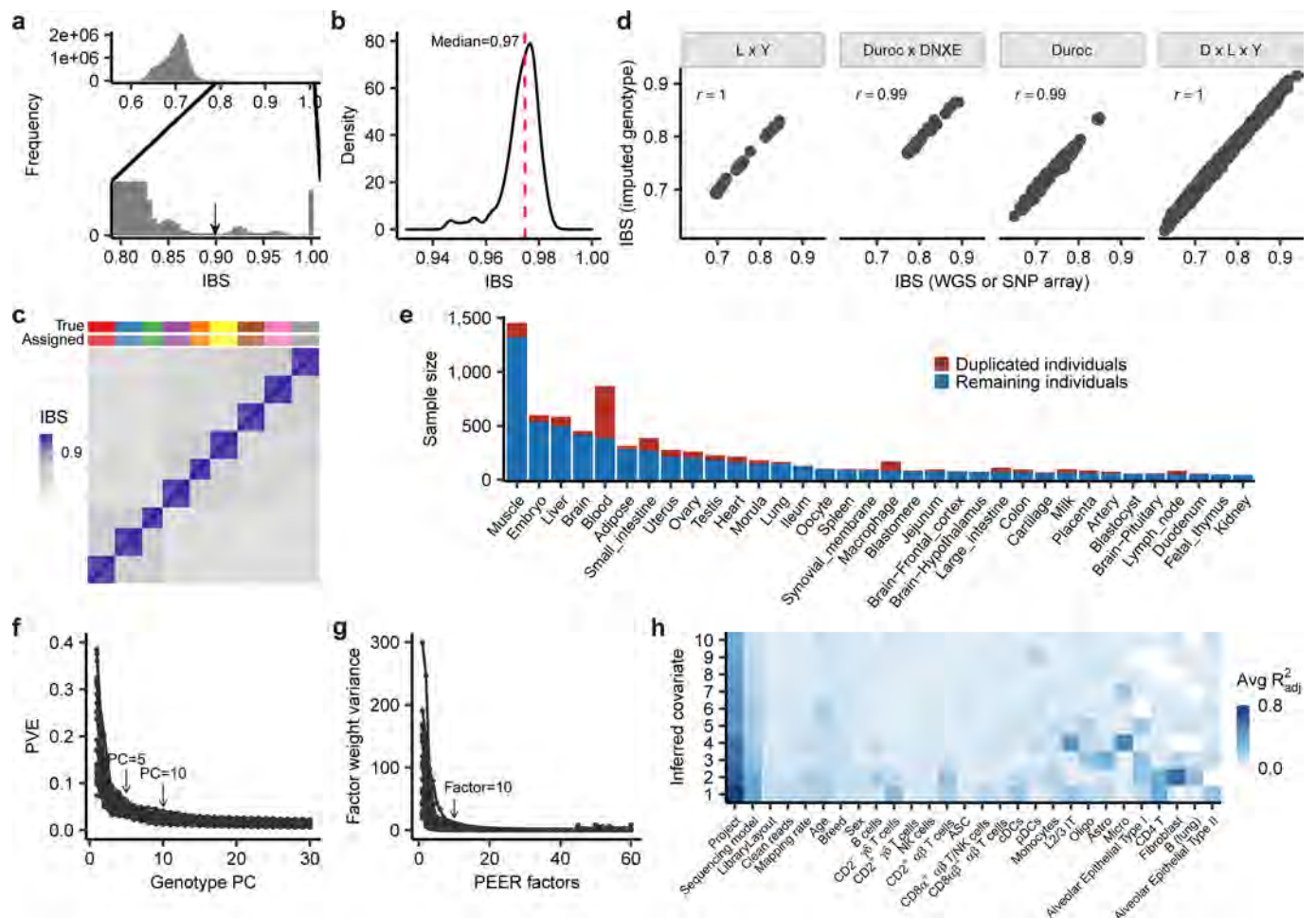


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Genotype calling and imputation and breed prediction.

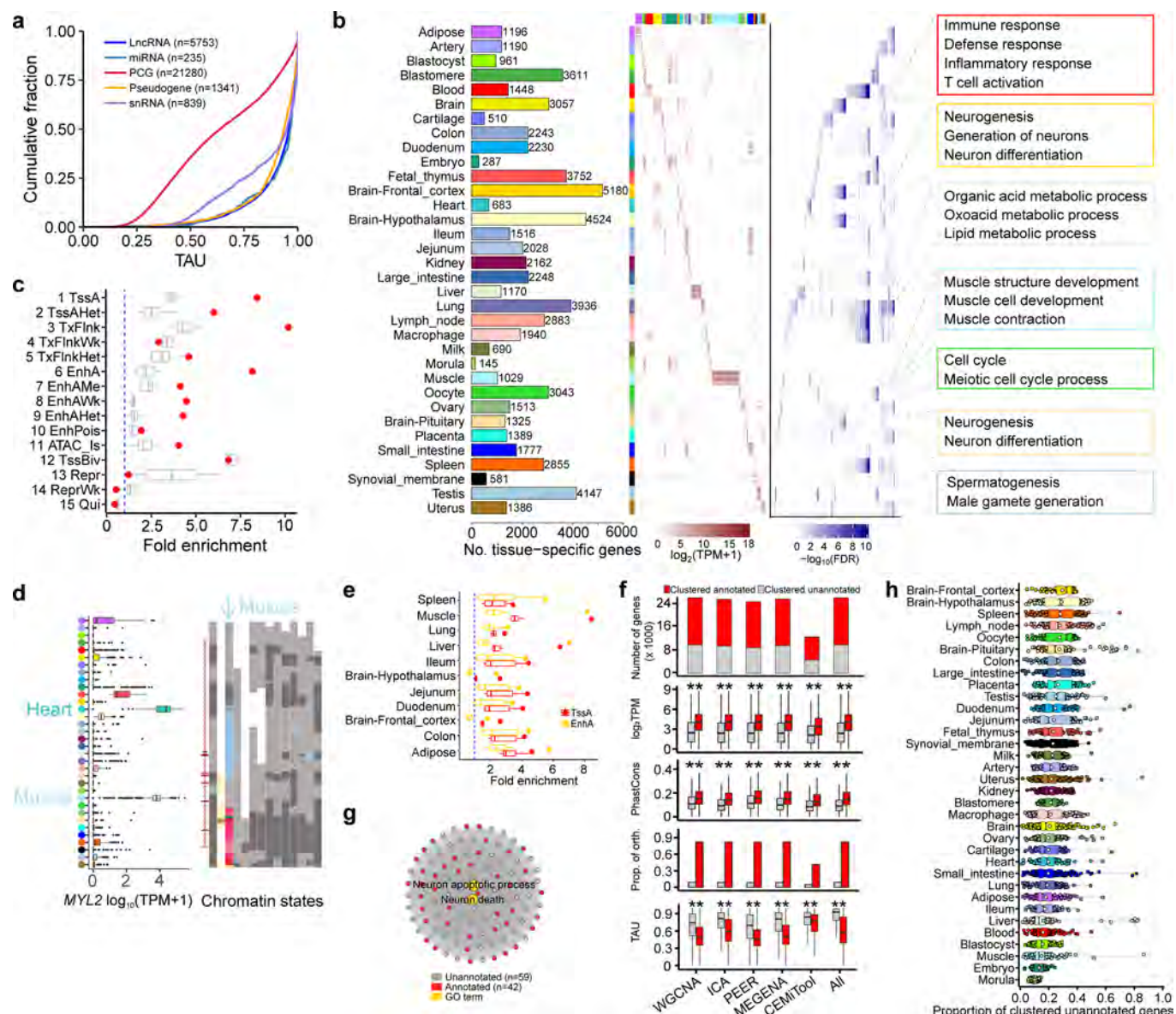
a, Pearson's correlation (r) between number of clean reads and number of called SNPs across 7,095 RNA-Seq samples. The P -value is obtained by Pearson's r test. **b**, Distribution of the number of SNPs called from 7,095 RNA-Seq samples. **c**, Number of imputed SNPs (left, gray bars) from 7,008 RNA-Seq samples across 18 pig chromosomes after quality control ($DR^2 \geq 0.85$, minor allele frequency ≥ 0.05). The red point represents the number of genes (right) in each chromosome in the Sscrofa11.1 assembly (Ensembl v100). **d**, Distribution of 42,523,218 SNPs from the Pig Genomics Reference Panel (PGRP) and 3,087,268 imputed SNPs used for molecular QTL (molQTL) mapping across eight genomic features. **e**, Minor allele frequency (MAF) of imputed SNPs in 7,008 RNA-Seq samples. **f**, Distribution of the number of imputed SNPs around 1 Mb of transcript start site (TSS) of 18,911 protein-coding genes. **g**, Concordance rate (CR) and squared correlation (r^2) of imputed and observed genotypes in 50 evenly spaced MAF bins

based on individuals that are not present in the PGRP. 'ALL' represents the entire variants. **h**, CR and r^2 of imputed genotypes from RNA-Seq only and those directly called from whole-genome sequence (WGS) data (red), and imputed genotypes (blue) from SNP array, respectively, in the same individuals. Point and whisker are mean and standard deviation, respectively. Labels of x-axis are breeds and number of individuals. **i**, CR and r^2 (median and interquartile) of imputed and observed genotypes in different genomic features. Point and whisker are median and interquartile, respectively. **j**, The overall pipeline utilized to predict missing breed labels for RNA-Seq samples. **k**, Estimated ancestry proportion of Duroc ($n = 485$), Landrace ($n = 280$), Yorkshire ($n = 145$), Landrace \times Yorkshire ($n = 165$) and Duroc \times Landrace \times Yorkshire ($n = 40$) samples. **l**, Distribution of sample size of training and prediction sets in pure and cross breeds. **m, n**, Accuracy of breed prediction for pure breeds (**m**) and cross breeds (**n**) measured by cross-validation. The red triangle represents the sample size of the target breed.



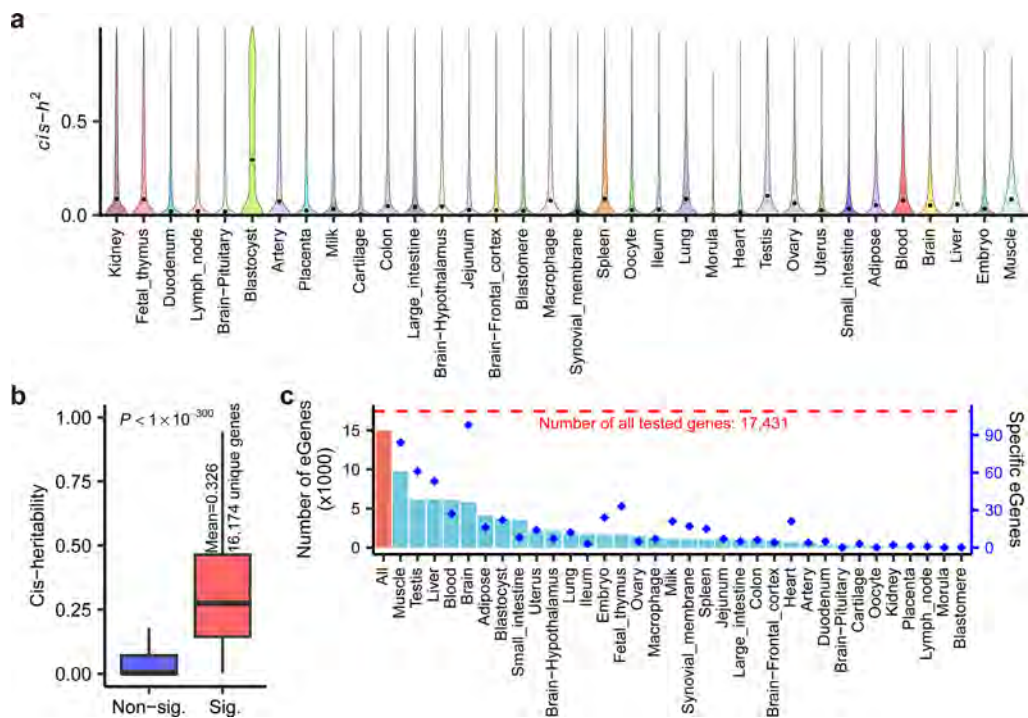
Extended Data Fig. 2 | Detection of duplicated individuals and confounders of RNA-Seq samples. **a**, Distribution of identity-by-state (IBS) distances among 7,008 RNA-Seq samples, which are calculated using 12,207 LD-independent SNPs ($r^2 < 0.2$). **b**, Density of IBS distances that were computed using genotypes derived from RNA-Seq only and whole-genome sequence (WGS) or SNP array data in the same individuals ($n = 227$). **c**, Heatmap of IBS distance of 25 RNA-Seq samples from 9 individuals. The same color on the top of panel represents samples from the same individuals. True: true individual label; Assigned: assigned individual label using an IBS distance cutoff of 0.9. **d**, Pearson's correlation (r) between IBS distance calculated from imputed genotypes and those calculated from WGS or SNP array data across four different populations. L×Y: Landrace and Yorkshire

cross breed ($n = 25$); Duroc×DNXE: Duroc and Diannanxiaoer cross breed ($n = 11$); Duroc: Duroc pure breed ($n = 37$); D×L×Y: composite population with 1/4 Duroc, 1/2 Landrace and 1/4 Yorkshire ($n = 179$). **e**, Duplicated and remaining individuals in each of the 34 pig tissues used for molecular QTL mapping. Sample pairs with IBS > 0.9 were considered as duplicated individuals. **f**, Proportion of variance explained (PVE) by genotype principal components (PC) in each of 34 tissues (lines). **g**, Factor weight variance of probabilistic estimation of expression residual (PEER) factors in each of 34 tissues (lines). **h**, Proportion of variance (adjusted R^2) of known confounders captured by the top 10 inferred PEER factors, calculated using the *lm* function in R (v4.0.2).



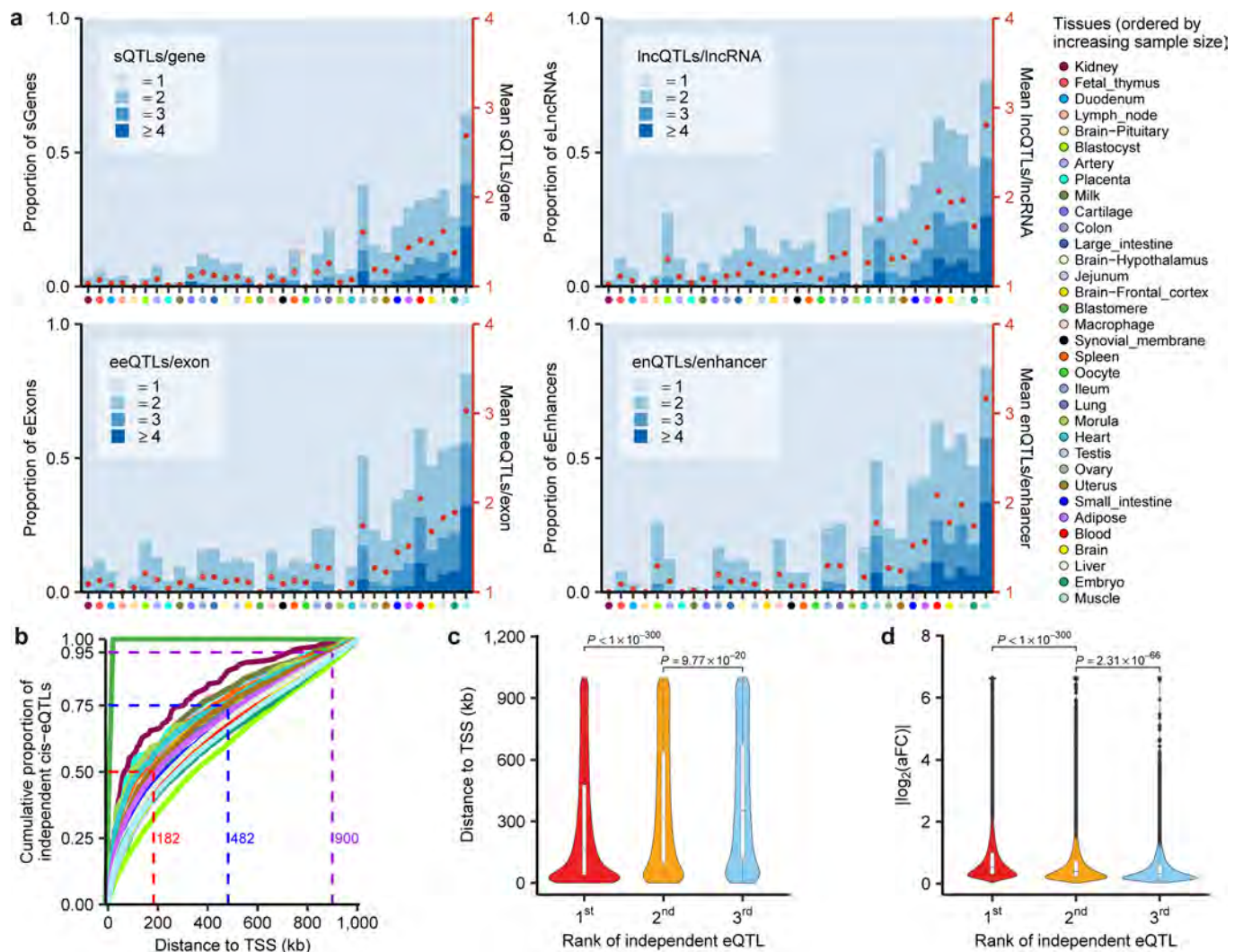
Extended Data Fig. 3 | The pig gene expression atlas. **a**, Tissue-specific expression of five transcript types reflected by the TAU score. PCG: protein-coding genes. **b**, Gene numbers (left), expression pattern (middle, transcripts per million, TPM), and enriched Gene Ontology (GO) terms (right) of tissue-specific genes in 34 tissues. **c**, Enrichment of muscle-specific genes in 15 tissues across 14 pig tissues¹⁶. The red dots represent respective chromatin states in muscle. The blue line indicates enrichment fold = 1. **d**, Expression profiles of MYL2 gene across 34 tissues (left). The tissue color key is the same as in (b). Chromatin state distribution (right) around MYL2 in 14 pig tissues¹⁶. In brief, red is for promoters, yellow for enhancers, blue for open chromatin and gray for repressed regions. **e**, Enrichment of tissue-specific genes for two active chromatin states across 11 tissues, which have both chromatin states and gene expression data. The dots represent enrichments from matching tissues. TssA is for active TSS (promoter), and EnhA for active enhancers. **f**, Comparison of genes with and without functional annotation (referred to as ‘annotated

genes’ and ‘unannotated genes’, respectively) in gene co-expression modules at different biological layers. The gene co-repression analysis was conducted using five complementary methods, including WGCNA, ICA, PEER, MEGENA and CEMiTool. ‘All’ shows the combined results from the five methods. The functional annotation was based on the Gene Ontology database (version 2022-01-18). The plots from top to bottom include Gene Ontology counts, expression level, PhastCons score from 100 vertebrate genomes, proportion of orthologous genes in humans and TAU values. Significant differences between annotated and unannotated genes were obtained using a two-sided Student *t*-test. ** means $P < 0.01$. **g**, An example of gene co-expression module in the pituitary, which includes 59 unannotated and 42 annotated genes, respectively. The functional annotated genes are significantly ($P = 8 \times 10^{-3}$) enriched in neuron apoptotic processes. The gray edges between genes represent Pearson’s correlations of expression across all 53 samples in the pituitary. **h**, The proportion of unannotated genes in each gene co-expression modules across 34 tissues.



Extended Data Fig. 4 | Cis-heritability of gene expression across 34 pig tissues. a, Distribution of estimated *cis*-heritability ($cis-h^2$) of gene expression across 34 tissues. The black point represents the median of $cis-h^2$ of all tested genes in a tissue. **b**, Box plot showing the $cis-h^2$ estimates of genes across 34 tissues that are significant (likelihood ratio test $P < 0.05$) or non-significant,

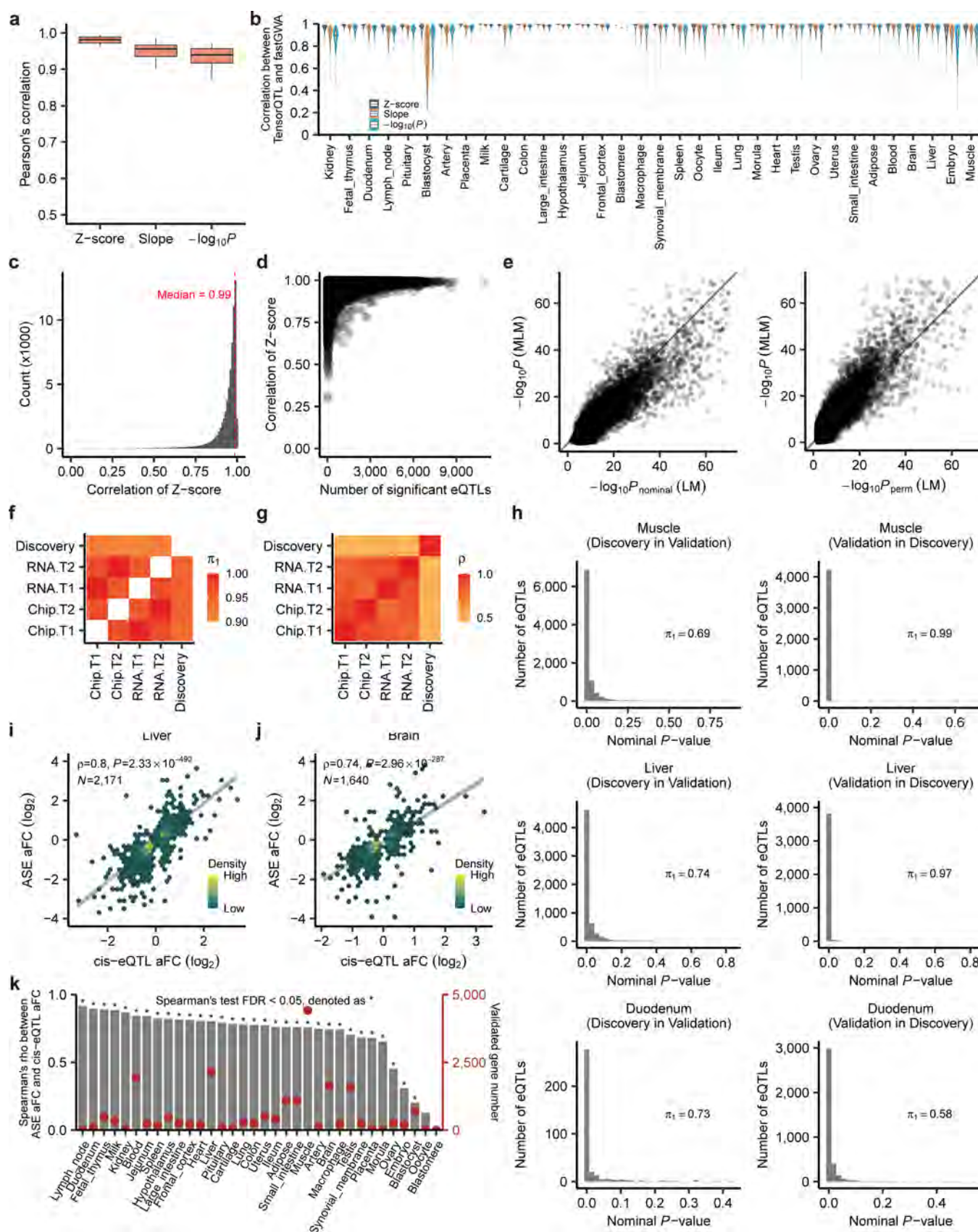
where 16,174 (93%) unique genes have significant *cis*-heritability in at least one tissue. The P value was calculated by two-sided Student t -test. **c**, The number of eGenes in each tested tissue, with 86% of the tested genes (red bar, left) are eGenes in at least one tissue. The blue points represent the number of tissue-specific eGenes.



Extended Data Fig. 5 | Conditionally independent molecular QTLs (molQTL).

a, Distribution and average number (red dots, right y-axis) of conditionally independent *cis*-QTL per eMolecules across 34 tissues. Tissues (x-axis) are ordered by increasing sample size. **b**, Cumulative proportion of distance to the transcription start site (TSS) of target genes for conditionally independent

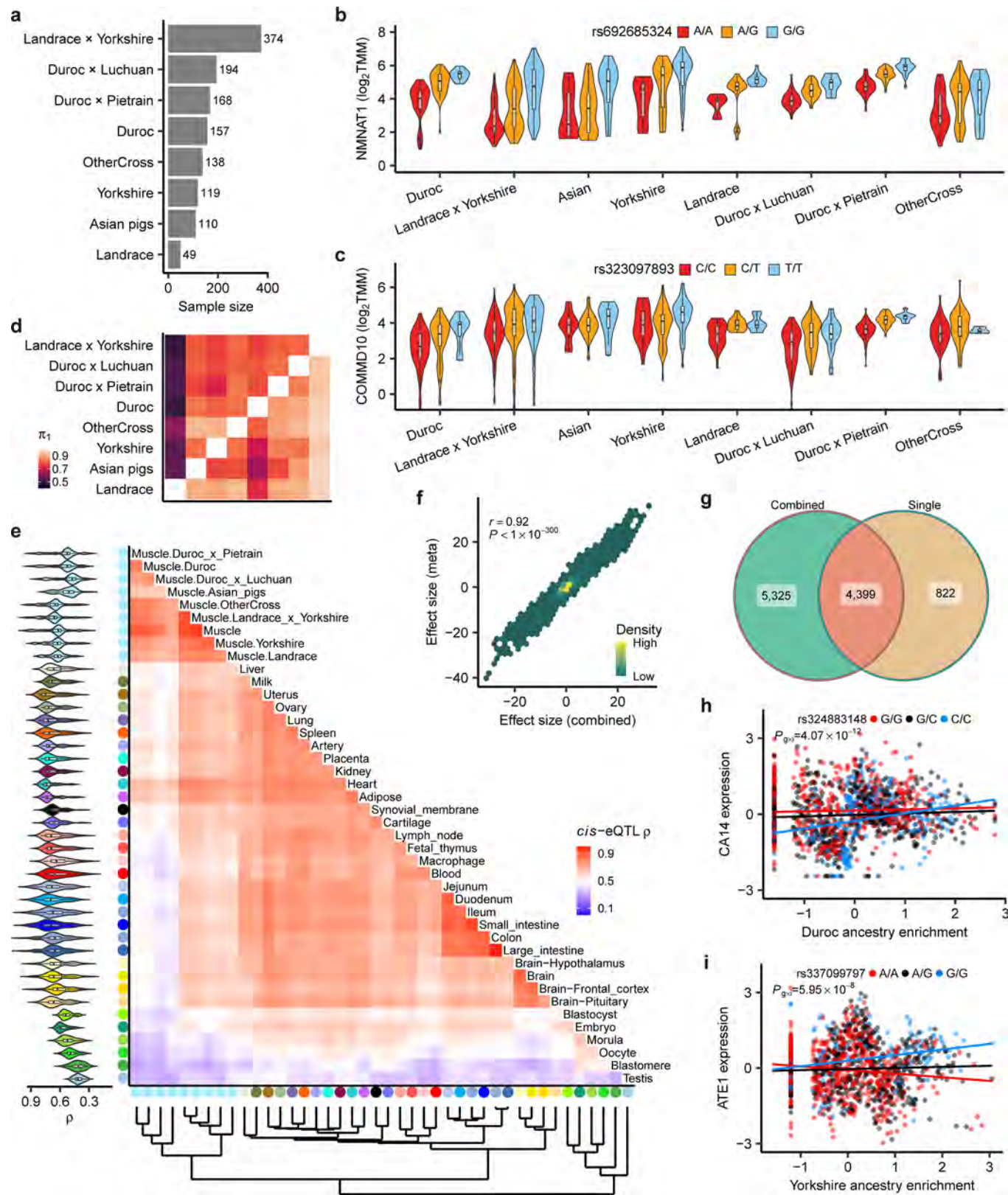
cis-eQTL in each of 34 tissues. The meanings of the colors of curved lines are the same as the color key in panel (a). **c,d**, Comparison of distance to TSS (**c**) and effect size ($|\log_2(\text{aFC})|$) (**d**) among top three independent *cis*-eQTL per eGene across 34 tissues. The aFC is for allelic fold change. The P values were obtained by the two-sided Wilcoxon rank-sum test.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Validation of *cis*-eQTL. **a**, Pearson's correlation of combined summary statistics (for example, Z-score, slope and *P*-value ($-\log_{10}$ scale)) of *cis*-eQTL for all the eGenes across 34 tissues between TensorQTL (linear model, LM) and fastGWA (mixed linear model, MLM). **b**, Pearson's correlation of summary statistics for each eGene in each tissue between LM and MLM. **c**, Distribution of the Pearson's correlations of Z-score between LM and MLM. **d**, Relationship between correlations of Z-score and the number of significant eQTL across all the eGenes. **e**, Correlation of *P* values derived from MLM and nominal (left) or permutation-corrected (right) *P* derived from LM for the lead eQTL of all the eGenes. **f**, Replication rates (π_1) of blood *cis*-eQTL between the PigGTEx discovery population ($n = 386$, Discovery) and the external datasets ($n = 179$). For π_1 calculation, rows are discovery populations, and columns are replication populations. The external datasets include whole-blood-cell RNA-Seq data and SNP Chip array (Chip) from 179 animals at two developmental

stages (T1 and T2). The prefix 'RNA' and 'Chip' indicate imputed genotypes from RNA-Seq and SNP array, respectively. **g**, Spearman's correlation (ρ) of effect size (z-scores) for blood *cis*-eQTL among the same populations above. **h**, Replication rates (π_1) of PigGTEx *cis*-eQTL in external validation datasets of three tissues, including muscle ($n_{\text{PigGTEx}} = 1,321$, $n_{\text{external}} = 100$), liver ($n_{\text{PigGTEx}} = 501$, $n_{\text{external}} = 100$) and duodenum ($n_{\text{PigGTEx}} = 49$, $n_{\text{external}} = 100$). The x-axis is the nominal *P*-value of *cis*-eQTL detected from dataset₂ and is significant in dataset₁ (that is, dataset₁ in dataset₂). **i, j**, Spearman's correlation (ρ) of effect sizes (allelic fold change, aFC in \log_2 scale) between *cis*-eQTL and matched allele-specific expression (ASE) loci in the liver (**i**) and brain (**j**). *N* indicates number of tested loci. The lines are fitted by a linear regression model using the *geom_smooth* function from ggplot2 (v3.3.2) in R (v4.0.2). The shading represents the standard error of the fitting line. **k**, Spearman's correlation (ρ) of effect sizes between *cis*-eQTL and matched ASE loci across 34 tissues. Red dots indicate number of tested loci (right y-axis).

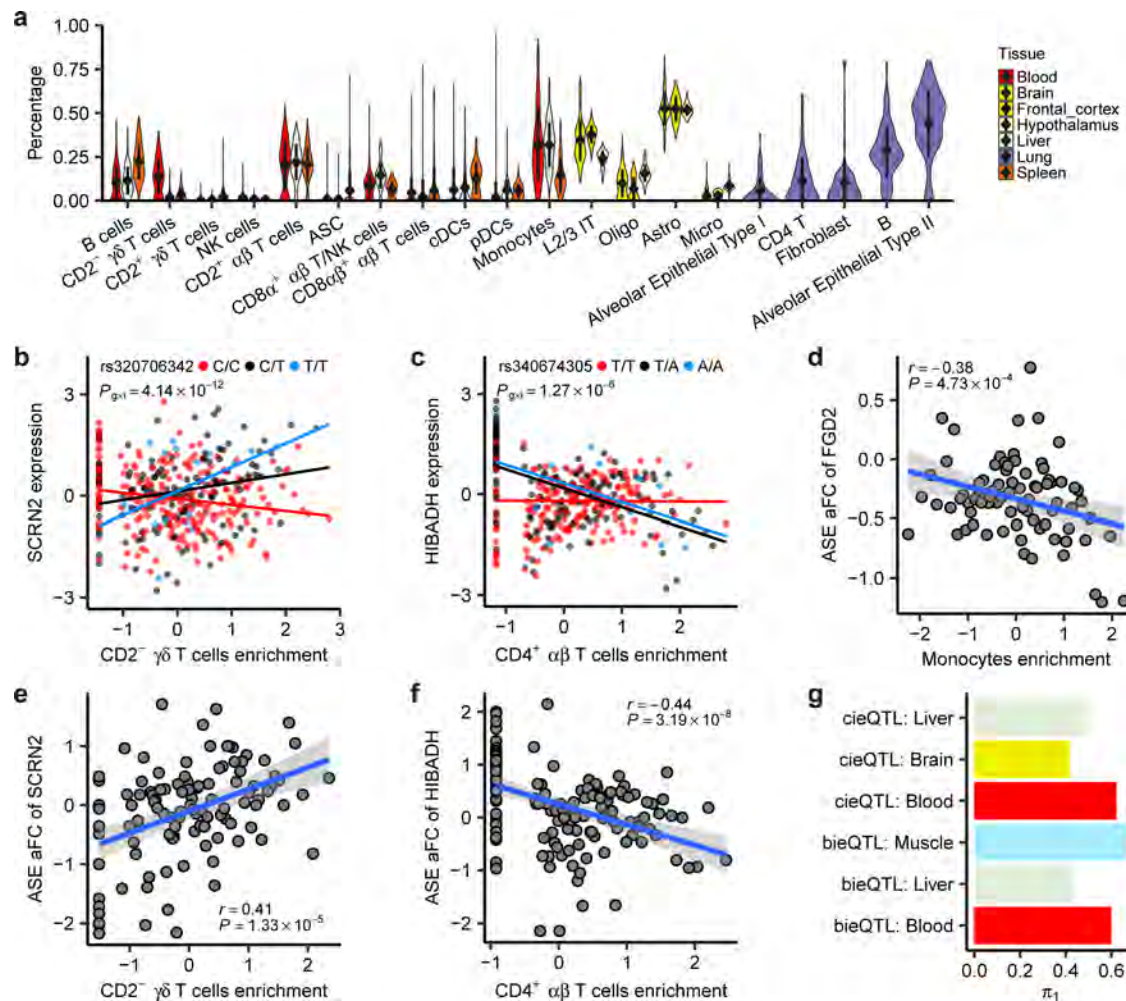


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Breed sharing and interaction *cis*-eQTL (bieQTL).

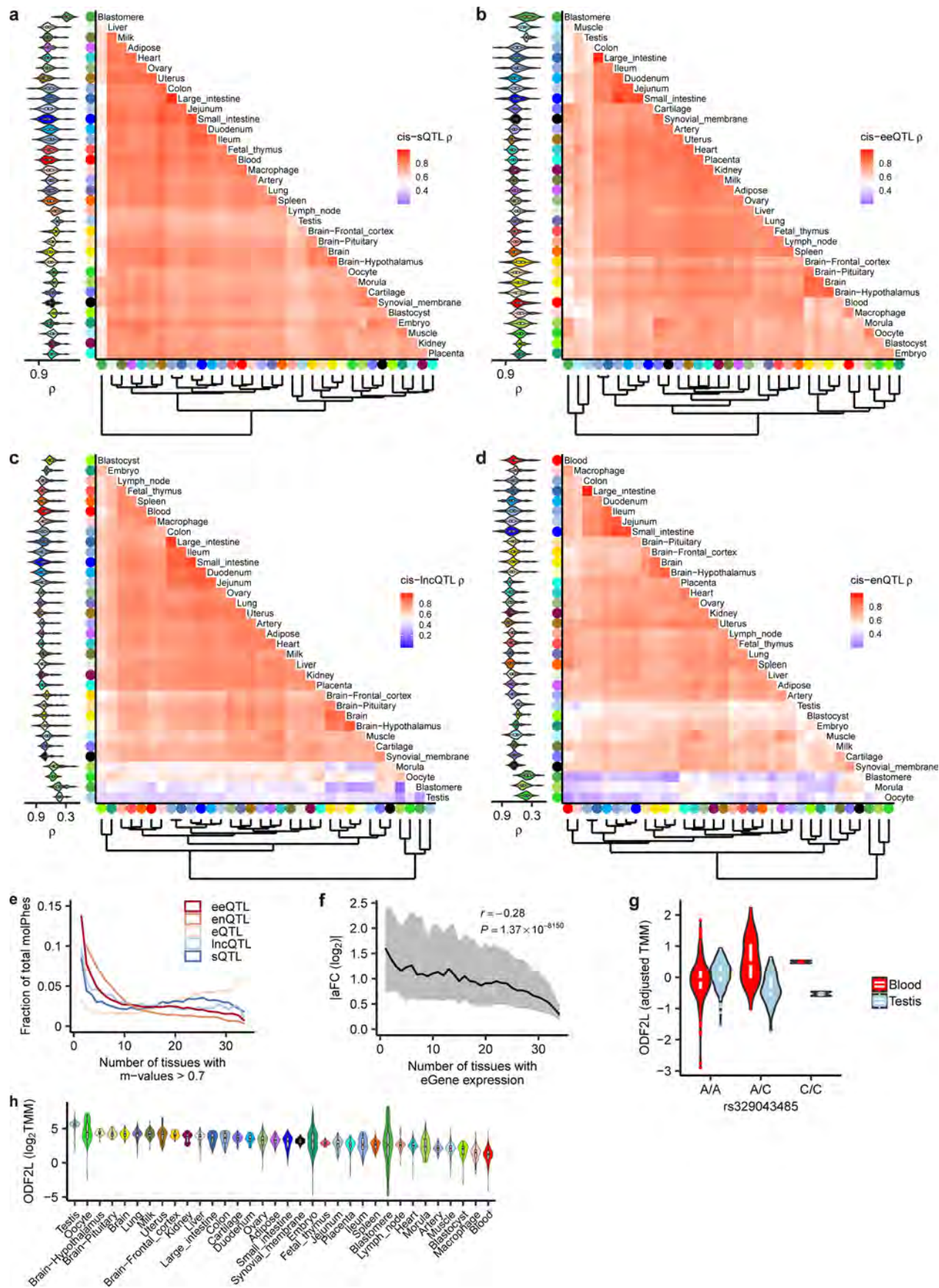
a, Sample size of muscle RNA-Seq data across eight breed groups. **b,c**, Expression levels of *NMNAT1* (**b**) and *COMMD10* (**c**) at three genotypes of *cis*-eQTL in muscle across eight breed groups. **d**, The *cis*-eQTL discovered in each breed group (rows) that can be replicated (π_1) across all other breed groups (columns). **e**, The heatmap of tissues regarding the pairwise Spearman's correlation (ρ) of *cis*-eQTL effect sizes. Tissues are grouped by hierarchical clustering (bottom). Violin plot (left) represents Spearman's correlation between the target group and the rest. **f**, Pearson's correlation (r) of effect size between *cis*-eQTL from the multi-breed

meta-analysis (y-axis) and those from the combined muscle population (x-axis). The P value was obtained from Pearson's r test. **g**, Overlap of *cis*-eQTL detected from the combined muscle population (Combined) and those detected in single-breed (Single) *cis*-eQTL mapping (shared in at least two breeds). **h,i**, Examples of bieQTL in muscle. Each dot in (**h**, *CA14*) and (**i**, *ATE1*) represents an individual and is colored by three genotypes. Gene expression levels and ancestry enrichment scores are inverse normal transformed. The two-sided P value is calculated by the linear regression bieQTL model. The lines are fitted by a linear regression model using the *geom_smooth* function from ggplot2 (v3.3.2) in R (v4.0.2).



Extended Data Fig. 8 | Cell-type enrichment and interaction cis-eQTL (cieQTL). **a**, Distribution of enrichment scores (percentage) of major cell types in samples of seven tested tissues (brain: $n = 415$, frontal cortex: $n = 75$, hypothalamus: $n = 73$, lung: $n = 149$, blood: $n = 386$, liver: $n = 501$, and spleen: $n = 91$). Each point and whisker indicate the mean value and standard deviation, respectively. **b,c**, Examples of cieQTL in blood. Each dot in (**b**, *SCRN2*) and (**c**, *HIBADH*) represents an individual and is colored by three genotypes. Gene expression levels and cell-type enrichment scores are inverse normal transformed. The two-sided P value was calculated by the linear regression

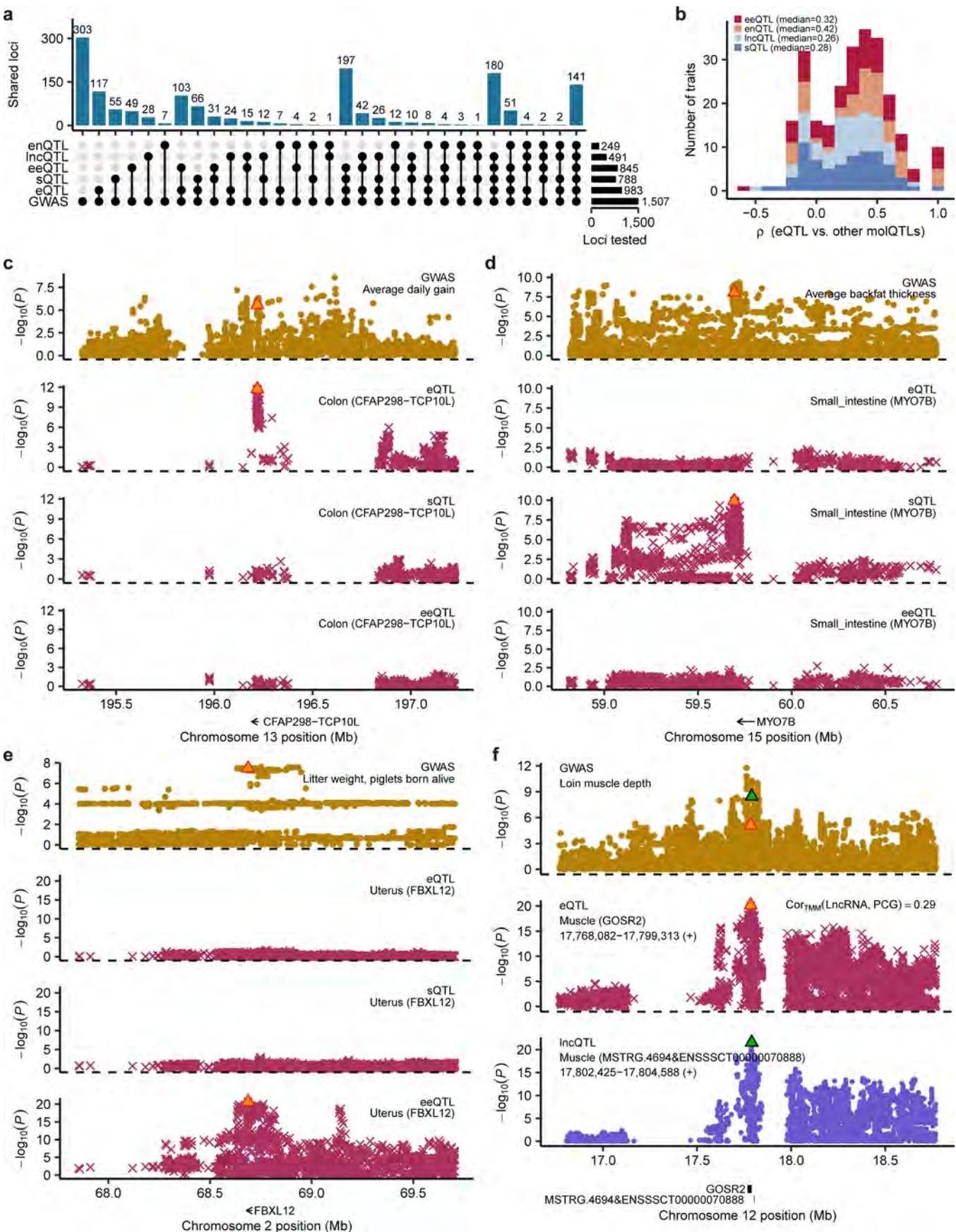
cieQTL model. The lines are fitted by a linear regression model using the *geom_smooth* function from *ggplot2* (v3.3.2) in R (v4.0.2). **d–f**, Pearson's correlation (r) between allele-specific expression (ASE) effect sizes (allelic fold change, aFC) and specific cell-type enrichment scores for *FGD2* with monocytes (**d**), *SCRN2* with CD2⁺ γδ T cells (**e**) and *HIBADH* with CD4⁺ αβ T cells in the blood (**f**). The lines are fitted by a linear regression model using the *geom_smooth* function from *ggplot2* (v3.3.2) in R (v4.0.2). The shading represents the standard error of the fitting line. **g**, ASE validation rate (π_1) of breed/cell-type interaction QTL (bieQTL and cieQTL) across tissues with ≥ 5 detectable bieQTL or cieQTL.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Tissue-sharing and specificity patterns of molecular QTL (molQTL). **a–d**, The heatmap of tissues regarding the pairwise Spearman's correlation (ρ) of molQTL effect sizes, that is, *cis*-sQTL (**a**), *cis*-eeQTL (**b**), *cis*-lncQTL (**c**) and *cis*-enQTL (**d**). Tissues are grouped by the hierarchical clustering (bottom). Violin plot (left) represents Spearman's correlations between the target tissue and the rest. **e**, Distribution of number of tissues having METASOFT activity (m -value > 0.7) for each of molQTL. MolPhe: molecular phenotype. **f**, Pearson's correlation (r) between number of tissues an eGene expressed in

(transcript per million, TPM > 0.1) and its *cis*-eQTL effect sizes ($|aFC(\log_2)|$). The aFC is for allelic fold change. The line and shading indicate the median and interquartile range, respectively. **g**, Expression levels (adjusted TMM) of *ODF2L* at three genotypes of top *cis*-eQTL (rs329043485) in blood and testis. TMM: trimmed mean of M-value normalized expression levels. There are 337, 47 and 2 samples for A/A, A/C and C/C genotypes in blood, respectively, and 148, 34 and 2 in testis, respectively. **h**, Expression levels (\log_2 TMM) of *ODF2L* across 34 tissues. Tissues are ordered (from smallest to largest) by the median expression values.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Complementarity of molecular QTL (molQTL) in interpreting GWAS loci. **a**, Number of GWAS loci linked to *cis*-eQTL, *cis*-sQTL, *cis*-eeQTL, *cis*-lncQTL and *cis*-enQTL in 34 tissues based on four different integrative methods, including colocalization (fastEnloc), Mendelian randomization (SMR), single-tissue transcriptome-wide association studies (TWAS, S-PrediXcan) and multi-tissue TWAS (S-MultiXcan). The bottom point-line combinations of the Upset plot represent the intersections of GWAS loci linked to eGenes by different types of molecular phenotypes. **b**, Distribution of rank correlations between tissue-relevance-scores derived from *cis*-eQTL and those from *cis*-sQTL, *cis*-lncQTL, *cis*-eeQTL and *cis*-enQTL across 86 GWAS traits with significant colocalizations for at least one molecular phenotype. **c**, Significant SMR signals ($P_{\text{SMR}} = 9.16 \times 10^{-5}$, $P_{\text{HEIDI}} = 0.9$) between GWAS loci of average daily gain (ADG) and *cis*-eQTL of *CFAP298-TCP10L* in colon, but not for its *cis*-sQTL or *cis*-eeQTL. The orange triangle represents the top *cis*-eQTL of *CFAP298-TCP10L*. **d**, Significant

SMR signals ($P_{\text{SMR}} = 1.78 \times 10^{-5}$, $P_{\text{HEIDI}} = 0.07$) between GWAS loci of the average backfat thickness (BFT) and *cis*-sQTL of *MYO7B* in the small intestine, but not for its *cis*-eQTL or *cis*-eeQTL. **e**, Significant SMR signals ($P_{\text{SMR}} = 1.78 \times 10^{-6}$, $P_{\text{HEIDI}} = 0.97$) between GWAS loci of litter weight (LW, piglets born alive) and *cis*-eeQTL of *FBXL12* in the uterus, but not for its *cis*-eQTL or *cis*-sQTL. **f**, Significant SMR signals ($P_{\text{SMR}(\text{lncQTL-GWAS})} = 4.49 \times 10^{-7}$, $P_{\text{SMR}(\text{eQTL-GWAS})} = 5.45 \times 10^{-5}$, $P_{\text{SMR}(\text{lncQTL-eQTL})} = 4.62 \times 10^{-7}$) among GWAS loci of loin muscle depth (LMD), *cis*-lncQTL of *MSTRG.4694&ENSSSCT00000070888*, and *cis*-eQTL of *GOSR2* in the muscle. *MSTRG.4694&ENSSSCT00000070888* is a lncRNA gene located on the 3112 bp downstream of *GOSR2*, where the Pearson's correlation of their normalized expression levels (trimmed mean of M-value, TMM) is 0.29 in muscle. The orange and green triangles in the top GWAS Manhattan plot represent the top molQTL of *GOSR2* and *MSTRG.4694&ENSSSCT00000070888*, respectively.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input type="checkbox"/>	<input checked="" type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input type="checkbox"/>	<input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input type="checkbox"/>	<input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	All raw data analyzed in this study are publicly available for download without restrictions from SRA (https://www.ncbi.nlm.nih.gov/sra/) and BIGD (https://bigd.big.ac.cn/bioproject/) databases using the wget function in Linux. Details of RNA-Seq, WGS, WGBS, single-cell RNA-Seq and Hi-C datasets can be found in Supplementary Table 1, 2, 5, 8 and 9, respectively.
Data analysis	<p>All the computational scripts and codes (with software version) for RNA-Seq, WGS, WGBS, single-cell RNA-Seq and Hi-C datasets analyses, as well as the respective quality control, molecular phenotype normalization, genotype imputation, molQTL mapping, functional enrichment, colocalization, SMR and TWAS are available at the FarmGTEx GitHub website (https://github.com/FarmGTEx/PigGTEx-Pipeline-v0, https://doi.org/10.6084/m9.figshare.24247771).</p> <p>For RNA-Seq data analysis, we used Trimmomatic (v0.39), STAR (v2.7.0), Stringtie (v2.1.1), featureCounts (v1.5.2), Leafcutter (v0.2.9), GATK (v4.0.8.1), phASER (v1.1.1), and Beagle (v5.1) for quality control, mapping, gene expression quantification, alternative splicing, SNP calling, ASE analysis, and genotype imputation, respectively. For sample clustering, we used MEGA (vX) and then visualized with iTOL (v6). For tissue-specific gene expression, we used limma (v3.51.2). For gene co-expression analysis, we used WGCNA (v1.69), ICA (v1.0.2), PEER (v1.3), MEGENA (v1.3.7), and CEMiTool (v1.8.3). For gene functional enrichment analysis, we used clusterProfiler (v4.0) and visualized it using Gephi (v0.9.2).</p> <p>For WGS analysis, we used Trimmomatic (v0.39), BWA-MEM (v0.7.5a-r405), Picard (v2.21.2), GATK (v4.1.4.1), and Beagle (v5.1) for quality control, mapping, marked duplicated reads, variants calling, and phasing, respectively. We used PLINK (v1.90) to do LD pruning.</p> <p>For WGBS, we used FastQC (v0.11.9), Trim Galore (v0.4.5), Bismark (v0.19.0), and SMART2 (v2.2.8), Methpipe (v4.1.1), and FastQTL (v2.184) for quality evaluation, quality control, read mapping and DNA methylation level extraction, hypomethylation region detection, allele-specific</p>

methylation loci analysis and methylation QTL mapping, respectively.

For Hi-C, we used Trim Galore (v0.6.7), BWA (v0.7.17), Juicer (v1.6), Arrowhead (v1.22.01), hicConvertFormat (v3.7.1), pyGenomeTracks (v3.6) for quality control, read mapping, Hi-C contact matrix construction, TAD identification, format conversion, and visualization, respectively.

For single-cell RNA-Seq, we used Seurat (v3.0.2), Azimuth (v0.4.0) and CIBERSORTx online tool (v1) for data processing, cell type annotation and cell type deconvolution, respectively.

We removed SNPs with MAF < 0.01 and/or missing rate > 0.9 using bcftools (v1.9) and employed Beagle (v5.1) to phase the filtered variants and impute sporadically missing genotypes.

For QTL mapping, we used TensorQTL (v1.0.3), aFC (v0.3), dap-g (v1.0.0), METASOFT (v2.0.1), MashR (v0.2-6), and GCTA (v1.93.0) for cis-QTL mapping, effect size estimation, fine-mapping, meta-analysis, tissue-sharing pattern estimation, and cis-QTL mapping with mixed linear model, respectively. We estimated the genetic parameters using the restricted maximum likelihood (REML) method implemented in GCTA (v1.93.0).

We computed genotype PCs based on the filtered SNPs within each of the tissues using SNPrelate (v1.26.0). To account for technical confounders among RNA-Seq samples (e.g., hidden batch effects and other technical or biological factors), we used the Probabilistic Estimation of Expression Residuals (PEER) method, implemented in peer R package (v4.0.2), to estimate a set of latent covariates within each of the 34 tissues based on gene expression matrices. We computed the mappability of each locus in the reference genome using GenMap (v1.3.0). We removed SNPs in repeat regions annotated by the UCSC RepeatMasker track.

We first used imputed genotypes to estimate the ancestry composition of all RNA-Seq samples across tissues using ADMIXTURE (v1.3.0). We estimated the effect size (aFC) of the top ieQTL of ieGenes from ASE data using the script phaser_cis_var.py in phASER (v1.1.1).

For integrative analysis between GWAS and molQTL, we used S-PrediXcan and S-MultiXcan in MetaXcan (v0.6.11) for single-tissue and multi-tissue TWAS analysis, SMR (v1.03) for Mendelian Randomization analysis, and fastENLOC (v1.0) for colocalization. We performed a meta-analysis of molQTL across all 34 tissues using MashR (v0.2-6) and METASOFT (v2.0.1). We calculated the pairwise Rand index to measure the clustering similarity using the rand.index function in the fossil (v0.4.0) R package (v4.0.2).

We performed 2,056 separate GWAS, and conducted the meta-GWAS analysis for the same traits across different populations based on GWAS summary statistics using METAL (v2011-03-25).

For enrichment analysis, we used TORUS (v1) and ClusterProfiler (v4.0) for molQTLs and genes functional annotation, respectively.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All raw data analyzed in this study are publicly available for download without restrictions from SRA (<https://www.ncbi.nlm.nih.gov/sra/>) and BIGD (<https://bigd.big.ac.cn/bioproject/>) databases. Details of RNA-Seq, WGS, WGBS, single-cell RNA-Seq and Hi-C datasets can be found in Supplementary Tables 1, 2, 5, 8 and 9, respectively. All WGS data generated in this study are available under CNGB GSA (<https://ngdc.cncb.ac.cn/>) accessions: PRJCA016120, PRJCA016130, PRJCA017284, PRJCA016012, and PRJCA016216. All processed data and the full summary statistics of molQTL mapping are available at <http://piggtex.farmgtex.org/>.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No power calculation was needed in advance in this study. In total, we analyzed all 11,323 RNA-Seq runs (downloaded by March, 2021) from SRA (https://www.ncbi.nlm.nih.gov/sra/), and BIGD databases (https://bigd.big.ac.cn/bioproject/), yielding 9,530 unique RNA-Seq samples. After filtering the samples with low quality (see below), all the remaining samples have been used for analysis.
Data exclusions	<p>Full details of data exclusions for each analysis can be found in the Methods as well.</p> <p>We filtered out RNA-Seq samples with clean read counts $\leq 500K$ or uniquely mapping rates $< 60\%$, resulting in 8,262 samples. We further excluded samples with obvious clustering errors (e.g., samples labeled as liver that were not clustered with other liver samples), resulting in 7,095 samples for subsequent analysis.</p> <p>For cis-QTLs detection, we excluded tissues with less than 40 individuals, resulting in 34 tissues for cis-QTL mapping.</p>
Replication	<p>To validate the cis-eQTLs, we applied four distinct strategies including linear mixed model, internal validation, external validation, and ASE validation.</p> <p>First, we observed that the summary statistics of cis-eQTL derived from the linear regression model in TensorQTL had a strong correlation (an average Pearson's r of 0.91 across tissues) with those from a linear mixed model.</p> <p>Second, we performed an internal validation in 18 tissues with over 80 samples by randomly dividing samples into two equal groups, and then conducting cis-eQTL mapping separately in both subgroups. We observed a high replication rate (an average π_1 of 0.92) for cis-eQTL discovery.</p> <p>Third, we found that 92%, 74%, 73%, and 69% of cis-eQTL in blood, liver, duodenum, and muscle, respectively, were replicated in independent datasets.</p> <p>Fourth, we further found that effects (allelic fold changes, aFC) derived from allele specific expression (ASE) analysis were significantly correlated with those from cis-eQTL mapping consistently across tissues. For instance, in muscle, ASE-derived effects of 4,417 SNPs were significantly correlated (Spearman's $\rho = 0.76$, $P < 1e-300$) with their cis-eQTL effects.</p>
Randomization	All the datasets are from observation studies and we used all samples publicly available after data exclusions listed above. Therefore, Randomization were not relevant in this study. Samples were grouped by tissue types.
Blinding	In this study, we re-analyzed all the publicly available RNA-seq data using a uniform pipeline, followed by the population-based association studies and validated the findings in independent populations. The blinding study design may be not applicable in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

RESEARCH

Open Access



Genome-wide association analysis of heifer livability and early first calving in Holstein cattle

Yahui Gao^{1,2}, Alexis Marceau¹, Victoria Iqbal¹, Jose Antonio Torres-Vázquez¹, Mahesh Neupane², Jicai Jiang³, George E. Liu² and Li Ma^{1*}

Abstract

Background The survival and fertility of heifers are critical factors for the success of dairy farms. The mortality of heifers poses a significant challenge to the management and profitability of the dairy industry. In dairy farming, achieving early first calving of heifers is also essential for optimal productivity and sustainability. Recently, Council on Dairy Cattle Breeding (CDCB) and USDA have developed new evaluations of heifer health and fertility traits. However, the genetic basis of these traits has yet to be thoroughly studied.

Results Leveraging the extensive U.S dairy genomic database maintained at CDCB, we conducted large-scale GWAS analyses of two heifer traits, livability and early first calving. Despite the large sample size, we found no major QTL for heifer livability. However, we identified a major QTL in the bovine MHC region associated with early first calving. Our GO analysis based on nearby genes detected 91 significant GO terms with a large proportion related to the immune system. This QTL in the MHC region was also confirmed in the analysis of 27 K bull with imputed sequence variants. Since these traits have few major QTL, we evaluated the genome-wide distribution of GWAS signals across different functional genomics categories. For heifer livability, we observed significant enrichment in promotor and enhancer-related regions. For early calving, we found more associations in active TSS, active Elements, and Insulator. We also identified significant enrichment of CDS and conserved variants in the GWAS results of both traits. By linking GWAS results and transcriptome data from the CattleGTEx project via TWAS, we detected four and 23 significant gene-trait association pairs for heifer livability and early calving, respectively. Interestingly, we discovered six genes for early calving in the Bovine MHC region, including two genes in lymph node tissue and one gene each in blood, adipose, hypothalamus, and leukocyte.

Conclusion Our large-scale GWAS analyses of two heifer traits identified a major QTL in the bovine MHC region for early first calving. Additional functional enrichment and TWAS analyses confirmed the MHC QTL with relevant biological evidence. Our results revealed the complex genetic basis of heifer health and fertility traits and indicated a potential connection between the immune system and reproduction in cattle.

Keywords Dairy Cattle, Heifer, Fertility, Disease, GWAS, TWAS

*Correspondence:

Li Ma

lima@umd.edu

¹Department of Animal and Avian Sciences, University of Maryland, Room 2123, 8127 Regents Drive, College Park, MD 20742, USA

²Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, MD 20705, USA

³Department of Animal Science, North Carolina State University, Raleigh 27695, USA



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Heifers are young female cows that have not yet given birth to a calf and are the future of the dairy herd. Heifer health and fertility are crucial for the success of a dairy farm, as it directly affects milk production and the sustainability of a farm [1]. Healthy heifers are more likely to produce more milk once they begin lactating. In addition, heifers need to be healthy and well-cared for to conceive and carry a calf to term successfully. Unhealthy heifers may have difficulty getting pregnant, leading to lower milk production and fewer replacement cows for the farmer [2]. Diseased heifers may also require costly veterinary treatments that can add up quickly and cut into profits. Early first calving in heifers is also important for dairy farming, particularly for economic and environmental considerations, because it can reduce unproductive periods and increase lifetime production, faster generation turnover and selection progress, and improve reproductive efficiency [3]. In summary, heifer health and fertility are critical for a profitable and sustainable dairy operation [1].

The health and fertility of cows are complex issues that involve various factors, including nutrition, environment, physiology, and genetics [4–6]. Compared to cows, heifers generally encounter more challenges as heifers have not yet reached sexual maturity and are not yet capable of smooth reproduction and production. Moreover, heifers typically have additional nutritional requirements than cows, as they are still growing and developing. Despite the complexity of cattle health and fertility, many GWAS studies have been conducted to identify genomic regions and genes associated with health and fertility-related traits in cattle [7–12]. For instance, the bovine MHC region has been associated with cow livability and immune system-related diseases [7]. The *ABCC9* and *GC* genes have been associated with pregnancy rate [8], while *ARRDC3* was associated with growth and calving traits [4]. Heifer fertility and health traits are less studied than cows, mainly due to limited data availability.

Although the heritability of fertility and health traits tends to be relatively low, CDCB and the USDA Animal Genomics and Improvement Lab have been evaluating fertility and health-related traits using the large volume of data collected from the dairy industry (<https://uscdcb.com/>). Recently, they added heifer livability and early first calving to the evaluation system [13, 14]. Heifer livability represents the expected survival percentage of an animal's female offspring from 2 days after birth up to 18 months of age in a herd with average management conditions. Larger, positive values of heifer livability are more favorable. It measures a heifer's overall resistance to causes leading to mortality. Since the most common reasons for heifer death are digestive and respiratory diseases [15], heifer livability is primarily related to the

resistance to these diseases and other causes of death. The heritability of heifer mortality has been estimated to be less than 1% in many studies [13, 16, 17]. Early first calving (EFC) is defined as the age at first calving. As a heifer fertility trait, the heritability of EFC is only 2–3% [14, 18, 19]. As part of the genetic evaluation process, traits have been corrected for management effect by CDCB, resulting in a PTA (Predicted Transmission Ability) that can be used directly for genetic studies. In this research, we aim to identify genes and genomic regions associated with these heifer traits using the large amount of genotype and phenotype data from the US dairy genomic database. To further boost power, we also included transcriptome and other functional genomics data for fine-mapping and validation.

Results

Large-sample GWAS of heifer livability and early calving

Our large-sample GWAS started with a discovery population of 3,649,734 genotyped Holstein cattle (336,386 bulls and 3,313,348 cows). After calculating deregressed predicted transmitting ability (PTA) as phenotype and editing, we included 510,318 and 768,645 animals for the GWAS of heifer livability and early first calving, respectively. All of the animals were imputed to 79 K SNPs, and we retained 73,554 SNPs after QC editing. We applied SLEMM [20] to perform the GWAS analyses that can efficiently run large-scale mixed models and incorporate variational residual variances for differential reliabilities of deregressed PTAs. As a result, we found only one QTL region for each of the two traits (Fig. 1A and B C, and 1D). Nonetheless, for both traits, the *P* values for the majority of SNPs showed no inflation of test statistics and good quality of the results (Fig. 1A C). After removing SNPs with low minor allele frequency (MAF), only one SNP (ARS-BFGL-NGS-105563, $P=1.28\text{e-}07$) passed the Bonferroni-corrected threshold for early calving (Table 1). Interestingly, this SNP is located near the Bovine MHC region on BTA 23 that encodes many fundamental molecules for regulating the immune response [21].

Despite the few QTL regions detected in the initial GWAS, we evaluated all SNPs passing the suggestive significance levels for functional annotation analyses. For heifer livability, we obtained 118 genes located within or overlapping the vicinity of leading SNPs (<1 Mb) using BioMart in the Ensembl database (Ensembl Genes 106; Table S1). Several genes close to the top SNPs exhibited biological relevance for cow livability, including *CHCHD7* and *PLAG1*, which are related to growth and development [22] and the *LYN* gene related to the regulation of innate and adaptive immune responses [23]. We also performed GO analysis by KOBAS [24] to determine the potential biological functions of these genes.

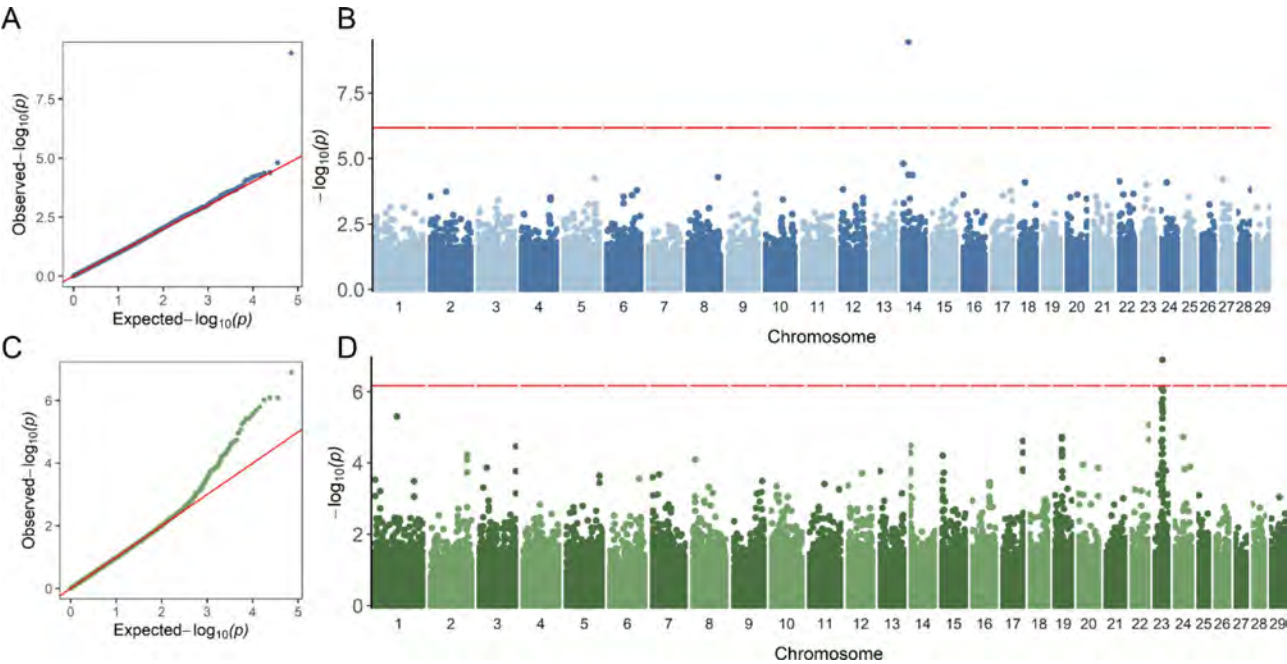


Fig. 1 Large-scale GWAS results of heifer livability and early calving based on the 79 K SNPs. **(A)** Quantile–quantile (QQ) plot for heifer livability. **(B)** Manhattan plot for heifer livability. **(C)** QQ plot for early calving. **(D)** Manhattan plot for early calving. The red horizontal lines correspond to the genome-wide significance threshold

Table 1 GWAS results for two heifer traits based on 79 K SNPs

Trait	Chr	SNP	Position	P	MAF
Heifer Livability	5	BovineHD0500029553	102,805,184	5.66E-05	0.0988
	8	BovineHD0800030830	101,970,883	5.16E-05	0.4321
	14	Hapmap25183-BTC-049425	5,880,036	1.59E-05	0.4835
	14	BovineHD1400007185	23,060,870	3.53E-10	0.0011
	14	BovineHD1400007271	23,389,588	4.16E-05	0.0011
	14	BTA-107899-no-rs	36,267,667	4.32E-05	0.0020
	18	BovineHD1800007256	23,555,637	8.14E-05	0.3387
	22	ARS-BFGL-NGS-100995	5,061,438	7.42E-05	0.2269
	23	BovineHD2300005239	20,146,079	9.79E-05	0.4162
	24	BTA-57516-no-rs	20,681,407	8.31E-05	0.3521
Early Calving	27	BovineHD2700002926	10,675,111	6.26E-05	0.3911
	1	Hapmap38109-BTA-36588	73,971,461	4.95E-06	0.1178
	22	ARS-BFGL-NGS-67185	60,478,622	8.48E-06	0.0260
	23	BovineHD2300007231	26,926,436	8.13E-07	0.1781
	23	BTA-27247-no-rs	26,934,192	2.07E-06	0.4998
	23	BovineHD2300007469	27,446,664	8.13E-07	0.2150
	23	BovineHD2300007953	28,526,405	2.64E-06	0.3401
	23	BovineHD2300008056	28,785,343	5.63E-06	0.4924
	23	BovineHD2300008081	28,825,626	3.31E-06	0.1532
	23	ARS-BFGL-NGS-105563	29,018,391	1.28E-07	0.4395
	23	BovineHD2300008507	29,958,908	1.61E-06	0.4158
	23	ARS-BFGL-NGS-104394	30,013,004	3.85E-06	0.4494
	23	Hapmap36280-SCAFFOLD155216_10397	30,176,828	3.85E-06	0.4495
	23	BovineHD2300008966	31,163,980	9.45E-07	0.1188

As a result, 139 significant GO terms ($P < 0.05$) were found, with the top relevant terms being mineral absorption, homeostasis, metabolic process, and development (Table S2). According to existing studies and the cattle QTL database [25], in the upstream and downstream 1 Mb range of the top SNPs, many QTLs were previously associated with milk production, body type, and disease related traits in dairy cattle (Table S3).

For early first calving, we identified 596 genes within or near the associated SNPs (Table S4). Notably, the top associated SNPs were located within or near the bovine MHC region on BTA 23, indicating potential connections between the immune system and early first calving [26]. Many nearby genes were involved with immune functions and relevant biology for early calving, including *ABCF1*, *ABHD16A*, *AGER*, *BOLA-NC1*, *BTN1A1*, *LTA*, *LTB*, etc. We also performed the GO analysis based on these genes and detected 91 significant GO terms ($P < 0.05$) with a large proportion associated with immune processes (Table S5). Finally, previously reported QTL within 1 Mb of associated SNPs were associated with milk production, reproduction, body type, and disease-related traits in cattle (Table S6).

Sequence-level GWAS and fine mapping of heifer livability and early calving in 27,235 bulls

To refine the GWAS results, we conducted additional GWAS analyses with imputed sequence data for heifer livability and early first calving in 27,235 bulls that have

highly accurate phenotypes. We used 3,148,506 imputed sequence SNPs as genotype and de-regressed PTAs as phenotype. After editing and filtering on reliability, we included 11,562 and 10,700 bulls for heifer livability and early calving, respectively. The QTL regions discovered in the large-sample GWAS were validated for both traits at the nominal significance level (Fig. 2A and B C, and 2D). Interestingly, sequence-level GWAS found some additional associations compared to low-density SNP data (Fig. 2B and D). As shown in Tables 2 and 16 SNPs passed the genome-wide threshold for heifer livability, and two SNPs passed the threshold for early calving. By checking the 1-Mb regions surrounding these associated SNPs, we identified many genes that were also detected in the large-sample GWAS, namely *MOG*, *OR12D2E*, *OR12D3*, *OR2H1*, *OR5V1*, *OR5V1C*, *OR5V2*, *TRIM10*, *TRIM15* (Table S7).

Functional enrichment analysis

We analyzed the enrichment of GWAS signals across SNPs in different functional genomic regions based on the 27 K bulls and imputed sequence data. We first categorized sequence variants into 14 groups based on the locations of 14 chromatin states reported previously [27], i.e., CTCF/Active_TSS, Active_TSS, CTCF/Promoter, Active_Promoter, Flanking_TSS, Promoter, Poised_Promoter, Active_Enhancer, CTCF/Enhancer, Primed_Enhancer, Active_Element, Insulator, Polycomb_Repressed, and Low_Signal. For heifer livability,

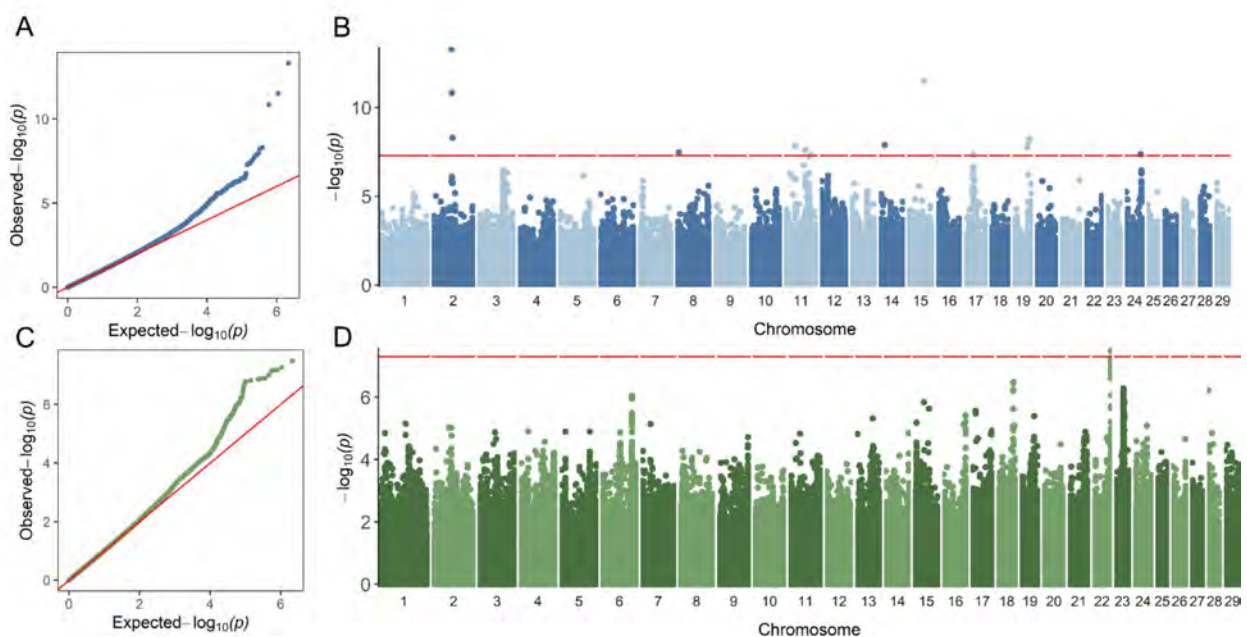


Fig. 2 GWAS results of heifer livability ($n = 11,562$) and early calving ($n = 10,700$) based on bulls with imputed sequence variants. (A) Quantile-quantile (QQ) plot for heifer livability. (B) Manhattan plot for heifer livability. (C) QQ plot for early calving. (D) Manhattan plot for early calving. The red horizontal lines correspond to the genome-wide significance threshold

Table 2 GWAS results based on imputed sequence variants for heifer livability (n = 11,562) and early calving (n = 10,700)

Trait	Chr	Position	P
Heifer Livability	2	60,792,638	4.90E-14
	2	63,123,798	1.44E-11
	2	63,150,519	1.44E-11
	2	64,387,597	5.00E-09
	8	6,002,999	3.24E-08
	11	33,641,469	1.43E-08
	11	67,900,672	2.39E-08
	11	67,903,861	2.39E-08
	11	87,363,160	4.45E-08
	14	16,826,158	1.27E-08
	15	54,473,598	3.05E-12
	17	27,784,375	4.32E-08
	19	48,236,260	1.73E-08
	19	50,258,942	1.07E-08
	19	54,797,534	5.91E-09
	24	52,033,790	4.15E-08
Early Calving	22	60,394,806	3.26E-08
	22	60,422,561	3.26E-08

we observed significant enrichment of variants in Active_Promoter, Promoter, CTCF/Enhancer, Primed_Enhancer, and Active_Element (Fig. 3). For early calving, we observed significant enrichment of associated variants in Active_TSS, Active_Element, and Insulator (Fig. 3).

We further investigated the enrichment of variants concerning their genomic locations (conserved) and genic annotations (CDS, intron, and UTR) inferred by SnpEff [28]. As a result, we observed significant enrichment of CDS and conserved variants in the GWAS results of both traits. For heifer livability, we observed enrichment of intron variants (Fig. 3). And for early calving, we observed significant enrichment of variants in the UTR regions (Fig. 3).

Transcriptome-wide association study (TWAS)

TWAS seeks to identify trait-associated genes by testing for the association between a phenotype and the genetic components of gene expression levels [29]. By linking our GWAS results and existing transcriptome data from the CattleGTEx project [30] via a TWAS analysis, we detected four and 23 significant gene-trait association pairs for heifer livability and early calving, respectively (Fig. 4). Interestingly, we discovered six genes overlapped with 27 K bulls GWAS results for early calving in the Bovine MHC region, including two genes in lymph node tissue and one gene in blood, adipose, hypothalamus, and leukocyte (Table 3). In addition, the expression of *OR12D2* in adipose was significantly associated with early calving (Table 3), consistent with previous findings that *OR12D2* is linked with MHC [31].

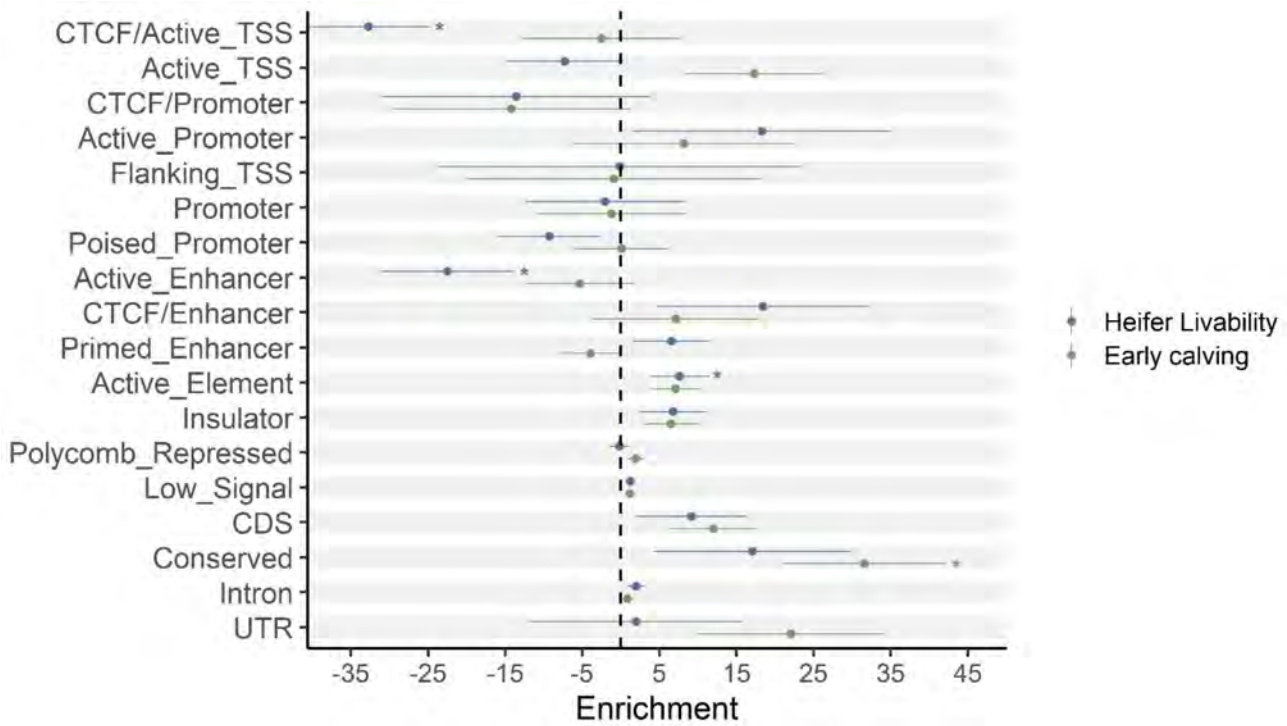


Fig. 3 Enrichment of fine-mapping variants across functional annotations

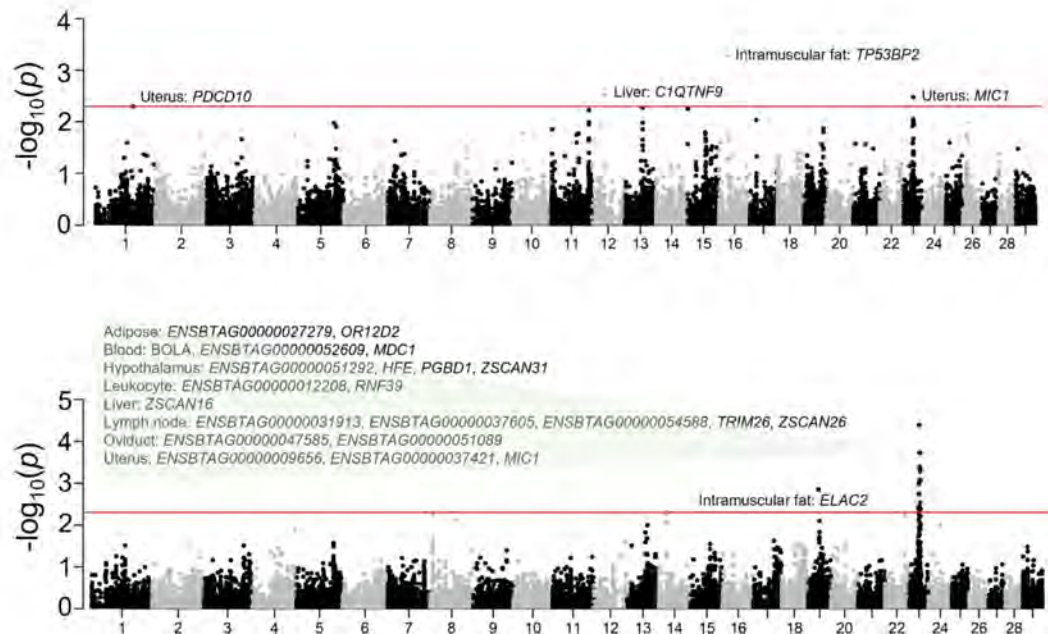


Fig. 4 Manhattan plots of transcriptome-wide association study (TWAS) for heifer livability and early calving

Discussion

In this study, we identified genomic regions and candidate genes associated with heifer livability and early first calving using large-scale GWAS, functional enrichment analysis, and TWAS. We reported a major QTL in the bovine MHC region to be related to early first calving, suggesting a potential connection between the immune system and reproduction. TWAS using the CattleGTEx data confirmed the association and revealed several candidate genes in the bovine MHC locus.

Generally, health and fertility traits are expected to be complex traits with low heritability due to their multifactorial nature [32]. The two traits analyzed in this study, heifer livability and early first calving, also have low estimated heritability, 0.0223 and 0.0328, respectively. These two heifer traits can be more complex due to the unique characteristics of heifers. For instance, heifers are not yet fully developed and may not be ready for reproduction. With the large sample sizes used in this study, we only found one major QTL for early first calving but none for heifer livability. Still, when we explored the genome-wide enrichment of GWAS signals with functional genomic regions, we reported significant enrichment of the association signals in promoter and enhancer regions, indicating an exciting connection between the regulation of gene expression and these two complex traits.

The immune system plays a central role in protecting the body from pathogens and infections. Moreover, it has a role in fertility and reproduction. In females, the immune system is involved in the whole reproductive process, from the development and maturation of the egg to implantation and maintenance of the pregnancy [33]. The immune system must tolerate the developing fetus, which is genetically different from the mother, while still protecting against infections. If the immune system is overactive, it can cause infertility or miscarriage, while an underactive immune system can lead to increased susceptibility to infections and complications during pregnancy. These potential connections between the immune system and reproduction further support the MHC QTL and candidate genes with early first calving in heifers.

This study showcased the usefulness of functional genomics data in post-GWAS and fine-mapping studies in cattle. When the associated variants are located in non-coding or intergenic regions, functional genomics data like those from the FAANG [27] and CattleGTEx projects would be useful to provide information about the biological mechanisms underlying the associations. Integration of functional genomics data with GWAS may also boost the power of detection when the power of the original GWAS is limited. For instance, the TWAS results in this research provided additional evidence for the MHC QTL with early first calving.

Table 3 Gene-trait association pairs detected by TWAS based on CattleGTEx database

Trait	Gene	Chr	Start	End	P value	#SNP	Tissue
Heifer Livability	<i>TP53BP2</i>	16	27,139,848	27,207,478	0.000527	18	Intramuscular fat
	<i>C1QTNF9</i>	12	34,293,438	34,304,505	0.003071	52	Liver
	<i>MIC1</i>	23	27,841,095	27,913,198	0.003337	4	Uterus
	<i>PDCD10</i>	1	99,744,376	99,804,996	0.004992	1	Uterus
Early Calving		23	28,925,617	28,926,246	4.00E-05	2	Lymph node
		23	27,796,195	27,797,556	4.09E-05	1	Lymph node
		23	29,612,534	29,613,169	0.000185	1	Adipose
	<i>ZSCAN26</i>	23	30,416,140	30,427,406	0.000185	1	Lymph node
	<i>PGBD1</i>	23	30,390,348	30,412,787	0.000185	1	Hypothalamus
	<i>TRIM26</i>	23	28,777,770	28,787,696	0.000185	1	Lymph node
		23	29,930,410	29,931,333	0.000187	1	Oviduct
	<i>BOLA</i>	23	28,720,501	28,724,399	0.000413	5	Blood
	<i>ZSCAN16</i>	23	30,561,557	30,568,915	0.000448	1	Liver
	<i>ZSCAN31</i>	23	30,377,190	30,379,817	0.000526	2	Hypothalamus
	<i>H4C3</i>	23	31,847,243	31,847,554	0.000836	2	Blood
		23	28,677,524	28,686,666	0.000877	1	Uterus
	<i>OR12D2</i>	23	29,305,933	29,309,785	0.000878	2	Adipose
	<i>MIC1</i>	23	27,841,095	27,913,198	0.001024	4	Uterus
	<i>ELAC2</i>	19	31,362,746	31,378,583	0.001419	19	Intramuscular fat
		23	27,871,206	27,875,056	0.001801	1	Leukocyte
		23	30,510,284	30,513,892	0.002936	1	Hypothalamus
	<i>MDC1</i>	23	28,304,399	28,316,822	0.003641	3	Blood
		23	25,691,259	25,695,296	0.003865	1	Lymph node
		23	25,583,083	25,589,209	0.003865	1	Uterus
	<i>HFE</i>	23	31,855,234	31,864,562	0.004076	1	Hypothalamus
	<i>RNF39</i>	23	28,904,289	28,908,861	0.00474	1	Leukocyte
		23	28,741,064	28,750,116	0.00474	1	Oviduct

Conclusion

Due to the complex genetic architecture of health and fertility traits, our large-scale GWAS analyses only detected a few major QTL for heifer livability and early first calving. Interestingly, the major QTL for early first calving is located in the bovine MHC region. This association was further supported by post-GWAS analyses and TWAS, indicating a connection between the immune system and early reproduction. Despite the low power for major QTL, we evaluated the distribution of GWAS signals across different functional genomic regions. We found significant enrichment in promoter and enhancer-related regions, which supports the contribution of gene regulation to the genetics of complex traits.

Methods

Data description

In this study, we conducted GWAS analyses with two datasets, a discovery dataset including 3,649,734 Holstein cattle (336,386 bulls and 3,313,348 cows) genotyped by various SNP chips and imputed to 79,060 SNPs and a fine-mapping dataset including 27,235 bulls genotyped by 50 K SNP chips and imputed to 3,148,506 sequence variants. The original SNP data of the discovery dataset were from multiple SNP chips with densities ranging

from 3 to 50 K [34]. The CDCB and USDA AGIL laboratory routinely process the original genotype data and impute to 79 K common SNPs specifically selected for official evaluations using FindHap program [35]. For the discovery dataset, we applied PLINK 1.9 [36] to remove SNPs with call rates < 95%, minor allele frequencies (MAF) < 0.01, Hardy-Weinberg equilibrium (HWE) $P < 10^{-6}$, and to remove animals with > 5% missing genotypes. After this filtering, 73,554 SNPs and 3,520,002 animals (325,905 bulls and 3,194,097 cows) were retained for downstream analyses.

The phenotype data were part of the December 2021 genomic evaluations from the U.S. Council on Dairy Cattle Breeding (CDCB), which routinely calculates predicted transmitting ability (PTA) values for dairy cattle of multiple breeds. We only included Holstein data for this study. We used deregressed PTA values as phenotype in the GWAS of two traits, heifer livability and early first calving [37]. To ensure robustness and accuracy, we excluded animals with low reliability. The majority of filtered animals were young cows without any phenotypic records. Finally, the total number of animals used was 510,318 and 768,645 for heifer livability and early calving, respectively.

For the fine-mapping dataset, we obtained imputed sequence data of 27,235 bulls from previous studies [8]. Briefly, the imputation was conducted with FindHap v3 [35] and 444 Holstein bulls from the Run5 of 1000 Bull Genomes Project as reference. Stringent filtering and removal of intergenic SNPs resulted in an enriched set of 3,148,506 sequence variants. The imputation was highly accurate with an average percentage of consistent genotypes 96.7%. Similarly, we excluded animals with low reliability for deregressed PTA values, retaining 11,562 and 10,700 bulls for heifer livability and early calving, respectively. In this study, we only considered autosomal chromosomes BTA 1–29 from the *Bos taurus* ARS-UCD1.2 assembly [38].

GWAS analysis

We analyzed the discovery and fine-mapping datasets separately in the GWAS analysis. We performed the GWAS using a linear mixed model approach implemented in the SLEMM program [20]. SLEMM can handle large-scale (up to millions) genome-wide association studies while accounting for genomic relationships. In addition, SLEMM can model differences in the reliability between individual phenotypes using an error weight parameter to account for the variation of deregressed PTAs, which is calculated by $1/r^2-1$, where r^2 is the reliability of deregressed PTAs.

After GWAS analysis, we retrieved genes within 1 Mb of the significant SNPs using BioMart in the Ensembl database (Ensembl Genes 106). We carried out Gene Ontology (GO) and Pathway analysis using KOBAS [24]. GO terms with a False Discovery Rate (FDR) less than 5% were considered statistically significant. Furthermore, we compared the regions within 1 Mb of the significant SNPs with our previous GWAS results [8] and the cattle QTLs in the Animal QTL database [25] to check if any associated genomic regions were previously reported.

Functional enrichment analysis with genome annotations

To evaluate the potential functions of the associated genomic regions, we explored the enrichment of GWAS results in different functional regions using the 27,235 bulls and imputed sequence variants. We performed enrichment analyses via MPH (MINQUE for Partitioning Heritability, <https://github.com/jiang18/ MPH>) with the annotations inferred by SnpEff [28] and 14 chromatin states from eight tissues reported by Kern et al. [27]. MPH is designed to partition SNP heritability with genotypes of related individuals or with long-spanning LDs. MPH is comparable to GREML in terms of accuracy, while being much faster and more memory efficient. It can do weighted analyses if residual variances are unequal and use many overlapping functional annotations. This approach included two steps: building

genomic relationship matrices (GRMs) based on the different SNP annotation datasets, and partitioning SNP heritability accordingly. We set `--min_maf` and `--min_hwe_pval` as 0 and 1e-8 respectively. We calculated standard errors using the Delta method.

Transcriptome-wide association study (TWAS)

We performed TWAS analyses based on the 27 K bull data using S-PrediXcan [39] to link GWAS results with transcriptome data that we assembled in a previous study [30]. For the TWAS analyses, we used the CattleGTEx v.1 eQTL models [36]. For each trait, we imputed and harmonized GWAS summary statistics and then performed TWAS across 24 cattle tissues separately. We considered genes with $P < 0.005$ as suggestive significant.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-023-09736-0>.

Supplementary Material 1

Acknowledgements

We thank the Council on Dairy Cattle Breeding (CDCB; Bowie, MD), Cooperative Dairy DNA Repository (Verona, WI), and dairy industry contributors for providing data access.

Authors' contributions

LM conceived the study. YG, AM, JJ, VI, JATV analyzed and interpreted data. YG and LM wrote the manuscript. JJ, MN, and GEL contributed tools and materials. All authors read and approved the final manuscript.

Funding

This work was supported in part by the USDA National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) grant 2020-67015-31398 and 2021-67015-33409. This work was also supported in part by USDA ARS appropriated projects 8042-31000-001-00-D, 8042-31000-002-00-D, and 8042-31310-078-00-D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

The original genotype and phenotype data are owned by third parties and maintained by the Council on Dairy Cattle Breeding (CDCB). A request to CDCB is necessary for getting data access on research, which may be sent to: João Dürr, CDCB Chief Executive Officer (joao.durr@cdcb.us). All other data have been included in the manuscript and supplementary data.

Declarations

Ethics approval and consent to participate

No live animals were used in this study, and Institutional Animal Care and Use Committee approval was not required.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 19 May 2023 / Accepted: 12 October 2023

Published online: 21 October 2023

References

- Moorey SE, Biase FH. Beef heifer fertility: importance of management practices and technological advancements. *J Anim Sci Biotechnol*. 2020;11(1):1–12.
- Wathes D, Pollett G, Johnson K, Richardson H, Cooke J. Heifer fertility and carry over consequences for life time production in dairy and beef cattle. *Animal*. 2014;8(s1):91–104.
- Krpálková L, Cabrera V, Kvapilík J, Burdych J, Crump P. Associations between age at first calving, rearing average daily weight gain, herd milk yield and dairy herd production, reproduction, and profitability. *J Dairy Sci*. 2014;97(10):6573–82.
- Wathes D, Brickell J, Bourne N, Swali A, Cheng Z. Factors influencing heifer survival and fertility on commercial dairy farms. *animal* 2008, 2(8):1135–1143.
- Kuhn M, Hutchison J, Wiggans G. Characterization of Holstein heifer fertility in the United States. *J Dairy Sci*. 2006;89(12):4907–20.
- Zhang H, Wang Y, Chang Y, Luo H, Brito LF, Dong Y, Shi R, Wang Y, Dong G, Liu L. Mortality-culling rates of dairy calves and replacement heifers and its risk factors in Holstein cattle. *Animals*. 2019;9(10):730.
- Freebern E, Santos DJ, Fang L, Jiang J, Parker Gaddis KL, Liu GE, VanRaden PM, Maltecca C, Cole JB, Ma L. GWAS and fine-mapping of livability and six Disease traits in Holstein cattle. *BMC Genomics*. 2020;21:1–11.
- Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and bayesian fine-mapping reveals candidate genes for important agromomic traits in Holstein bulls. *Commun Biology*. 2019;2(1):212.
- Nayeri S, Sargolzaei M, Abo-Ismael MK, May N, Miller SP, Schenkel F, Moore SS, Stothard P. Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet*. 2016;17(1):1–11.
- Tenghe A, Bouwman A, Berglund B, Strandberg E, de Koning D, Veerkamp R. Genome-wide association study for endocrine fertility traits using single nucleotide polymorphism arrays and sequence variants in dairy cattle. *J Dairy Sci*. 2016;99(7):5470–85.
- Johnston D, Mukiibi R, Waters SM, McGee M, Surlis C, McClure JC, McClure MC, Todd CG, Earley B. Genome wide association study of passive immunity and Disease traits in beef-suckler and dairy calves on Irish farms. *Sci Rep*. 2020;10(1):1–10.
- Narayana SG, de Jong E, Schenkel FS, Fonseca PA, Chud TC, Powel D, Wachoski-Dark G, Ronksley PE, Miglior F, Orsel K. Underlying genetic architecture of resistance to mastitis in dairy cattle: a systematic review and gene prioritization analysis of genome-wide association studies. *J Dairy Sci* 2022.
- Neupane M, Hutchison J, Van Tassell C, VanRaden P. Genomic evaluation of dairy heifer livability. *J Dairy Sci*. 2021;104(8):8959–65.
- Hutchison J, VanRaden P, Null D, Cole J, Bickhart D. Genomic evaluation of age at first calving. *J Dairy Sci*. 2017;100(8):6853–61.
- Gulliksen S, Lie K, Løken T, Østerås O. Calf mortality in Norwegian dairy herds. *J Dairy Sci*. 2009;92(6):2782–95.
- Fuerst-Waltl B, Sørensen M. Genetic analysis of calf and heifer losses in Danish holstein. *J Dairy Sci*. 2010;93(11):5436–42.
- Weller JL, Gershoni M, Ezra E. Genetic and environmental analysis of female calf survival in the Israel Holstein cattle population. *J Dairy Sci*. 2021;104(3):3278–91.
- Vergara O, Elzo M, Cerón-Muñoz M. Genetic parameters and genetic trends for age at first calving and calving interval in an Angus-Blanco Orejinegro-Zebu multibreed cattle population in Colombia. *Livest Sci*. 2009;126(1–3):318–22.
- Grossi D, Venturini G, Paz C, Bezerra L, Lôbo RB, Oliveira J, Munari D. Genetic associations between age at first calving and heifer body weight and scrotal circumference in Nelore cattle. *J Anim Breed Genet*. 2009;126(5):387–93.
- Cheng J, Maltecca C, VanRaden PM, O'Connell JR, Ma L, Jiang J. SLEMM: million-scale genomic predictions with window-based SNP weighting. *Bioinformatics*. 2023;39(3):btad127.
- Ellis SA, Ballingall KT. Cattle MHC: evolution in action? *Immunol Rev*. 1999;167(1):159–68.
- Nishimura S, Watanabe T, Mizoshita K, Tatsuda K, Fujita T, Watanabe N, Sugimoto Y, Takasuga A. Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese black cattle. *BMC Genet*. 2012;13:1–11.
- Xu Y, Harder KW, Huntington ND, Hibbs ML, Tarlinton DM. Lyn tyrosine kinase: accentuating the positive and the negative. *Immunity*. 2005;22(1):9–18.
- Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C-Y, Wei L. KOBAS 2.0: a web server for annotation and identification of enriched pathways and Diseases. *Nucleic Acids Res*. 2011;39(suppl2):W316–22.
- Hu Z-L, Park CA, Reecy JM. Bringing the animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res*. 2022;50(D1):D956–61.
- Abrams ET, Miller EM. The roles of the immune system in women's reproduction: evolutionary constraints and life history trade-offs. *Am J Phys Anthropol*. 2011;146(S53):134–54.
- Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, Saelao P, Waters S, Xiang R, Chamberlain A. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun*. 2021;12(1):1821.
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92.
- Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, De Geus EJ, Boomsma DI, Wright FA. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245–52.
- Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, Li B, Xiang R, Chamberlain AJ, Pairo-Castineira E. A multi-tissue atlas of regulatory variants in cattle. *Nat Genet*. 2022;54(9):1438–47.
- Younger RM, Amadou C, Bethel G, Ehlers A, Lindahl KF, Forbes S, Horton R, Milne S, Mungall AJ, Trowsdale J. Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome Res*. 2001;11(4):519–30.
- Ma L, Cole J, Da Y, VanRaden P. Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. *Journal of dairy science* 2019, 102(4):3735–3743.
- Lee SK, Kim CJ, Kim D-J, Kang J-h. Immune cells in the female reproductive tract. *Immune Netw*. 2015;15(1):16–26.
- Wiggans G, Cooper T, VanRaden P, Van Tassell C, Bickhart D, Sonstegard T. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. *J Dairy Sci*. 2016;99(6):4504–11.
- VanRaden PM, Sun C, O'Connell JR. Fast imputation using medium or low-coverage sequence data. *BMC Genet*. 2015;16:1–12.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015, 4(1):s13742–13015–10047–13748.
- Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Selection Evol*. 2009;41:1–8.
- Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elisk CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*. 2020;9(3):giaa021.
- Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9(1):1825.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RESEARCH

Open Access



Comparative transcriptome in large-scale human and cattle populations

Yuelin Yao^{1,2†}, Shuli Liu^{3,4†}, Charley Xia^{5,6†}, Yahui Gao^{3,7†}, Zhangyuan Pan^{8,9†}, Oriol Canela-Xandri¹, Ava Khamseh^{1,2}, Konrad Rawlik⁵, Sheng Wang¹⁰, Bingjie Li¹¹, Yi Zhang⁴, Erola Pairo-Castineira^{1,5}, Kenton D'Mellow¹, Xiujin Li¹², Ze Yan⁴, Cong-jun Li³, Ying Yu⁴, Shengli Zhang⁴, Li Ma⁷, John B. Cole³, Pablo J. Ross⁸, Huaijun Zhou⁸, Chris Haley^{1,5}, George E. Liu^{3*}, Lingzhao Fang^{1,13*}  and Albert Tenesa^{1,5*}

[†]Yuelin Yao, Shuli Liu, Charley Xia, Yahui Gao and Zhangyuan Pan contributed equally to this work.

*Correspondence: George.Liu@usda.gov; lingzhao.fang@qgg.au.dk; Albert.Tenesa@ed.ac.uk

¹ MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, EH4 2XU Edinburgh, UK

³ Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA

⁵ The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, UK
Full list of author information is available at the end of the article

Abstract

Background: Cross-species comparison of transcriptomes is important for elucidating evolutionary molecular mechanisms underpinning phenotypic variation between and within species, yet to date it has been essentially limited to model organisms with relatively small sample sizes.

Results: Here, we systematically analyze and compare 10,830 and 4866 publicly available RNA-seq samples in humans and cattle, respectively, representing 20 common tissues. Focusing on 17,315 orthologous genes, we demonstrate that mean/median gene expression, inter-individual variation of expression, expression quantitative trait loci, and gene co-expression networks are generally conserved between humans and cattle. By examining large-scale genome-wide association studies for 46 human traits (average $n = 327,973$) and 45 cattle traits (average $n = 24,635$), we reveal that the heritability of complex traits in both species is significantly more enriched in transcriptionally conserved than diverged genes across tissues.

Conclusions: In summary, our study provides a comprehensive comparison of transcriptomes between humans and cattle, which might help decipher the genetic and evolutionary basis of complex traits in both species.

Keywords: Comparative transcriptome, Gene co-expression, Heritability enrichment, Inter-individual variability, RNA-seq

Background

Cross-species comparison of the transcriptome enables a better interpretation of how natural selection shapes gene expression and is crucial for exploring the evolutionary basis of phenotypic variation between and within species. Comparison of the transcriptome between human and mouse has enhanced the use of mouse as models for a wide variety of diseases including neurological and muscular disorders, as well as cancer [1].



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Additionally, the comparison of the transcriptome across primates has provided molecular insights into human evolution, particularly in the brain [2].

Previous studies on comparative transcriptomics were essentially restricted to model organisms and human data from a few individuals, hindering the comparison of inter-individual variation of gene expression and associated genetic regulatory effects (e.g., expression quantitative trait loci, eQTLs) across species. Moreover, although it has been suggested that the genetic architecture underlying complex traits is conserved at a certain degree between humans and livestock [3–5], the molecular mechanisms underpinning such conservation are largely unknown. Until now, no study has systematically explored the conservation of transcriptome across a wide range of tissues in large populations of humans and any livestock species.

Cattle is one of the most economically important livestock species, supplying humans with a substantial fraction of animal protein. Driven by the high selection intensity of economically important traits, compared to humans, cattle has a different population structure, such as smaller effective population size ($N_e \sim 100$), higher linkage disequilibrium (LD) among genomic variants, and higher inbreeding rate (i.e., resulting in the accumulation of deleterious mutations) [6]. Furthermore, millions of highly accurate phenotypic records, including fertility, health, and growth traits, have been collected for cattle [7, 8]. As such, a better understanding of transcriptome conservation between humans and cattle may not only contribute to establishing cattle as a potential biomedical model for certain human diseases, but also enhance the cattle genetic improvement program by leveraging prior information from humans [5, 9]. Here, we select 10,830 and 4866 high-quality RNA-seq profiles from the human GTEx project (v8) [10] and the CattleGTEx project [11], respectively. We group human samples from similar tissues (e.g., different brain regions as brain) into bigger tissue classes, resulting in 20 matched tissues in humans and cattle (Additional file 1: Table S1). The large and tissue-diverse dataset analyzed allowed us to systematically compare the transcriptome of humans and a livestock species to gauge the conservation of gene expression in two outbred mammalian populations. We compare mean gene expression, inter-individual variation of gene expression, *cis*-eQTLs, and co-expression networks between humans and cattle, and then integrate results with large-scale genome-wide association studies (GWAS) from 46 human traits and 45 cattle traits to understand the genetic and evolutionary basis of complex traits.

Results

Global conservation of gene expression

We focused on the expression of 17,315 one-to-one orthologous genes, including 72% and 76% of all annotated protein-coding genes in humans and cattle, respectively. These orthologous genes, representing 16,510 protein-coding genes with 664 on sex chromosome, contributed to the majority of transcriptional outputs among all 20 tissues being studied in both humans and cattle (Additional file 2: Fig. S1). We analyzed an average of 243 and 541 RNA-seq samples across these 20 tissues in cattle and humans, respectively (Fig. 1a, Additional file 1: Table S1). We observed a significant correlation (Spearman's $r = 0.59$, $p = 6.7 \times 10^{-3}$) between the number of expressed (median Transcripts per Kilo-base Million, TPM > 0.1) genes in each tissue in humans and cattle (Fig. 1b). Testis has

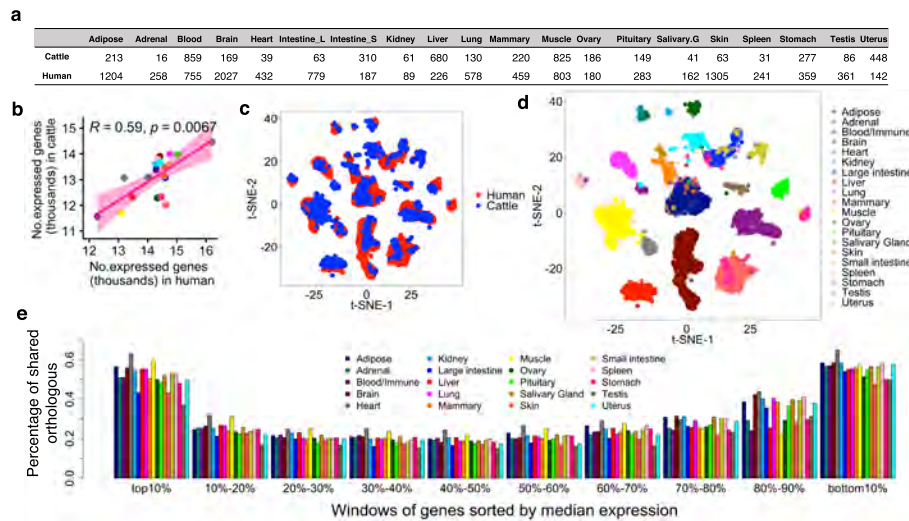


Fig. 1 Data summary and conservation of transcriptomes of 20 common tissues in humans and cattle. **a** Sample size per tissue in humans and cattle. **b** Spearman's correlation of number of expressed genes (median TPM > 0.1) across tissues between humans and cattle. Each dot represents a tissue. **c** Plot of t-SNE of samples based on batch-corrected gene expression (Methods). Each dot represents a sample, colored by species types. **d** Same as in **c**, but colored by tissue types. **e** Percentage of orthologous genes shared in each window between humans and cattle. Genes were ranked (from largest to smallest) by median expression in each tissue each species, and then divided into ten windows evenly (1731 genes per window)

the largest number of expressed genes in both species ($n_{\text{Human}} = 16,204$; $n_{\text{Cattle}} = 14,457$), while muscle ($n_{\text{Human}} = 13,081$; $n_{\text{Cattle}} = 11,707$) and blood ($n_{\text{Human}} = 12,283$; $n_{\text{Cattle}} = 11,573$) have the smallest in cattle and humans, respectively.

The t-SNE-based visualization of expression variation among samples clearly recapitulated tissue types (Fig. 1c, d). The hierarchical clustering of tissues based on mean or median gene expression in each tissue also showed that tissues rather than species clustered together (Additional file 2: Fig. S2a-b). These results demonstrate that gene expression profiles of orthologous genes are generally conserved within corresponding tissues between cattle and humans (Additional file 2: Fig. S3a). Tissues with the highest similarity of gene expression between humans and cattle included brain, pituitary, muscle, and adipose, while tissues with the lowest included stomach (the majority were rumen in cattle), skin, testis, and mammary gland (Additional file 2: Fig. S3b). In addition, we sorted all orthologous genes according to their median level of expression in each tissue, and observed that humans and cattle share most genes in the top (highest expression) and bottom (lowest expression) 10% of genes (Fig. 1e).

Conservation of tissue specificity of gene expression

We found that the distribution of median gene expression across tissues was U-shaped (tending towards either tissue-specific or ubiquitously expressed) in both humans and cattle, with the majority of genes (69% and 66% in humans and cattle, respectively) expressed in all 20 tissues (Fig. 2a). The number of tissues in which each gene was expressed was significantly correlated between the two species (Spearman's $r = 0.75$, $p < 2.2 \times 10^{-16}$), indicating that among orthologous genes there is global conservation of tissue-specific expression between humans and cattle. We found that 639 and 337 genes, with a significant (Hypergeometric test, $p < 2.2 \times 10^{-16}$) overlap of 165, were not

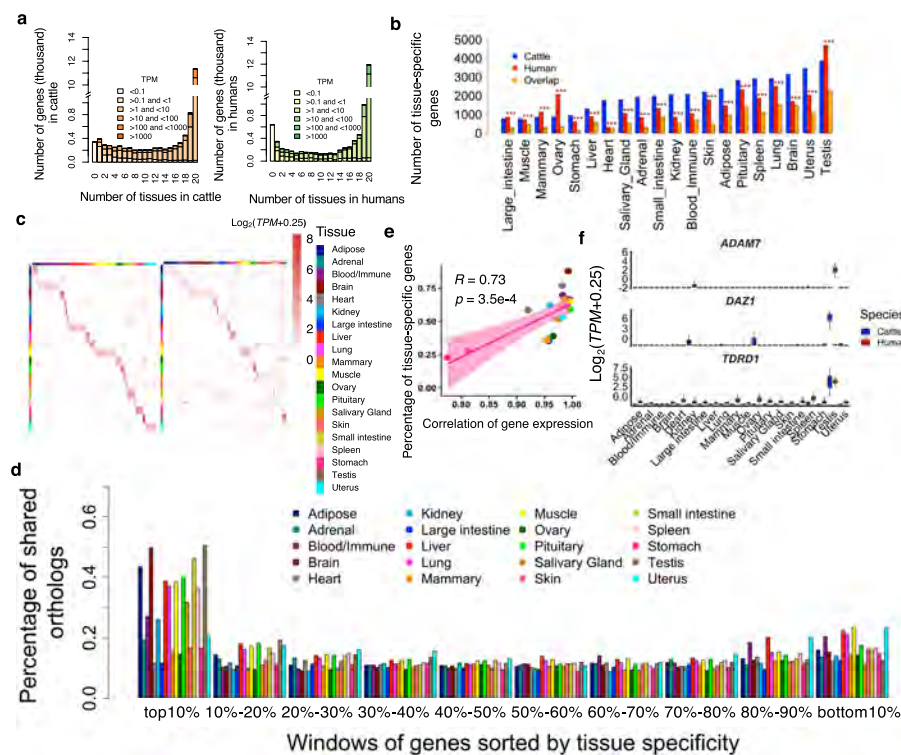


Fig. 2 Comparison of tissue specificity of gene expression. **a** Gene expression levels and number of tissues in which genes were expressed (median TPM > 0.1) in cattle (left) and humans (right). **b** Number of tissue-specific genes ($\log_2(\text{fold-change}) > 1.5$ and $\text{FDR} < 0.05$) and their overlap across 20 tissues in humans and cattle. The overlap was tested using hypergeometric test. **** represents FDR (Benjamini-Hochberg method corrected P -value) less than 1.0×10^{-3} . **c** Expression profiles of top 10 tissue-specific genes that are detected in cattle among both cattle (left) and humans samples (right). Each row represents a gene and each column represents a sample from the corresponding tissue. The color represents \log_2 -transformed expression value, i.e., $\log_2(\text{TPM}+0.25)$. **d** Percentage of orthologous genes shared in each bin between humans and cattle. Genes were ranked (from largest to smallest) by degree (measured by $-\log_{10}p$) of tissue specificity, and then divided into ten bins (1731 genes per bin). **e** Spearman's correlation between the percentage (%) of overlapping tissue-specific genes and gene expression correlation between humans and cattle across 20 tissues. Each dot represents a tissue. **f** Expression profiles of *ADAM7* (human-specific testis gene), *DAZ1* (cattle-specific testis gene), and *TDRD1* (conserved testis gene)

measurably expressed (TPM < 0.1) at the time of measurement in any of 20 tissues in humans and cattle, respectively. These non-expressed genes were significantly enriched in embryonic development processes, such as embryonic morphogenesis, angiogenesis, and regulation of stem cell division (Additional file 2: Fig. S4a). This might be due to the underrepresentation of embryonic samples in the current study.

We found that the number of tissue-specific genes across tissues was significantly correlated (Spearman's $r = 0.68$, $p = 1.2 \times 10^{-3}$) between humans and cattle (Additional file 2: Fig. S4b). The testis had the largest number of tissue-specific genes, while the large intestine and heart had the smallest in cattle and humans, respectively. In general, tissue-specific genes of the same tissues overlapped significantly (Hypergeometric test, $\text{FDR} < 1.0 \times 10^{-3}$) between humans and cattle (Fig. 2b). In each tissue, the top 10 tissue-specific genes with the largest expression values detected in cattle tissues also exhibited a strong pattern of tissue-specific expression in human tissues (Fig. 2c), and vice versa for the top 10 tissue-specific genes detected in human tissues (Additional file 2: Fig. S4c).

We observed that tissue specificity in gene expression was linked to the chances of genes being transcriptionally conserved between humans and cattle (Fig. 2d). The more similar the expression of two tissues was between species the larger the number of shared tissue-specific genes the tissues had (Spearman's $r = 0.73$; $p = 3.5 \times 10^{-4}$) (Fig. 2e). This finding indicates that tissues with more tissue-specific genes shared between humans and cattle tend to be more transcriptionally conserved between these two species.

We found that the tissue-specific genes shared by species (conserved) accurately reflected the known biology of tissues, while tissue-specific genes that were not shared by species (diverged) showed distinct biological functions in humans and cattle (Additional file 3: Table S2). For instance, the conserved testis-specific genes were significantly engaged in germ cell development, while human-specific and cattle-specific ones were significantly engaged in cilium organization and synapse assembly, respectively (Additional file 2: Fig. S4d). Of note, the difference in gene annotation databases between humans and cattle might bias the biological interpretation of human- and cattle-specific genes. We took *ADAM7*, *DAZ1*, and *TDRD1* as examples of human-specific, cattle-specific, and conserved genes in testis (Fig. 2f). *ADAM7* plays roles in sperm maturation and sperm-egg fusion [12]. *DAZ1* and *TDRD1* are essential for spermatogenesis [13, 14]. These species-specific genes in testis might be linked to the difference in fertility between humans and cattle, e.g., the difference in embryo implantation [15].

Comparison of mean gene expression level

We identified differentially expressed genes (DEGs) in each tissue between humans and cattle (Additional file 2: Fig. S5), and found that brain and pituitary showed the lowest number of DEGs (Fig. 3a), consistent with previous report that the central neural system evolves slowly across mammals [16]. In contrast, skin and stomach had the greatest number of DEGs, which was in line with the distinct physiological and anatomical characteristics of skin and stomach between humans and cattle. Using independent epigenetic data (i.e., ATAC-seq, and ChIP-seq for H3K4me3, H3K4me1, H3K27ac, and H3K27me3) in six common tissues in humans and cattle, we predicted 15 distinct chromatin states (Additional file 2: Fig. S6). We furthermore confirmed that TSS \pm 2kb of human upregulated DEGs showed an increased enrichment of active promoter-related states (e.g., TssA and TxFlnk) and decreased enrichment of repression-related states (e.g., TssBiv, TssAHet, Repr, and ReprWk) in humans when compared to their orthologous genes in cattle, and vice versa for cattle upregulated DEGs (Fig. 3b,c, Additional file 2: Fig. S7a). Furthermore, the upregulated DEGs in either humans or cattle exhibited distinct biological functions (Additional file 2: Fig. S7b, Additional file 4: Table S3). For instance, genes that were upregulated in cattle mammary gland were significantly engaged in protein secretion regulation, while genes that were upregulated in the human mammary gland were significantly engaged in responses to oxygen level (Additional file 4: Table S3). The oxygen level is important for supporting the increased metabolic rate during pregnancy and lactation in mammary gland. The downregulation of these genes in cattle mammary gland compared to humans might be partially due to the intensive selection of milk production and mammary gland health traits (e.g., mastitis) in cattle. We detected 511 and 461 genes were up- and downregulated in cattle rumen compared to human stomach. The upregulated genes in cattle rumen were mainly enriched

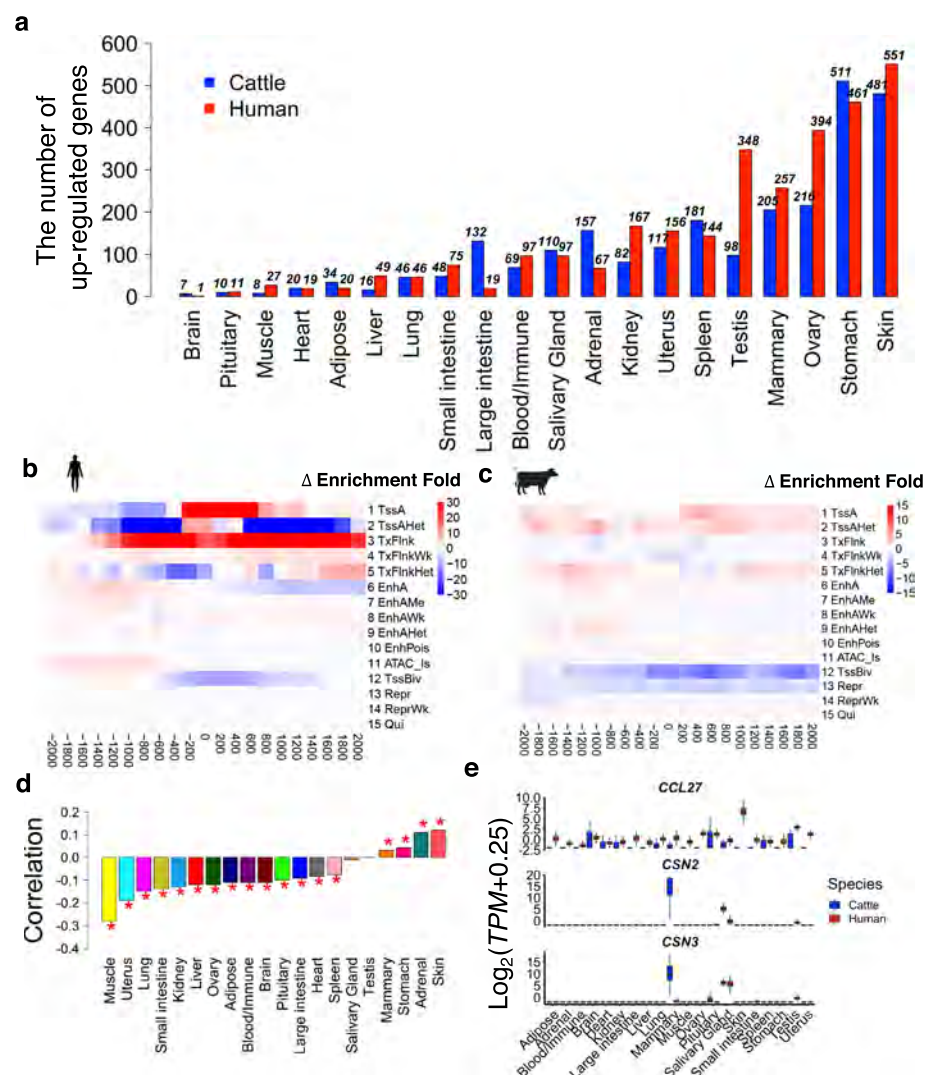


Fig. 3 Comparison of average gene expression across 20 tissues between humans and cattle. **a** Number of significantly upregulated genes across tissues in humans (red) and cattle (blue) using the cutoff of fold-change (FC) > 1.2 and FDR < 0.05. **b, c** Changes of enrichment folds of 15 chromatin states around (\pm 2kb) transcriptional start sites (TSS) of top 500 upregulated genes in human and cattle adipose when compared with each other, respectively. The 15 chromatin states are predicted based on six epigenetic marks (i.e., ATAC, CTCF, H3K27ac, H3K27me3, H3K4me1 and H3K4me3). **d** Spearman's correlation of genes between their tissue specificity (measured by $-\log_{10}p$ from tissue specificity expression analysis) of expression and degrees ($-\log_{10}p$) of differential expression between species. "*" represents the correlation coefficient is significant (FDR < 0.01). **e** Expression profiles of *CNS2*, *CNS3*, and *CCL27* across human (red) and cattle (blue) tissues

in multicellular organismal water homeostasis, cell-cell adhesion, and tissue development, while the downregulated genes were significantly enriched in digestion, response to topologically incorrect protein, response to endoplasmic reticulum stress, and muscle contraction. In addition, we detected 481 and 551 genes were up- and downregulated in cattle skin compared to human skin. The upregulated genes in cattle skin were mainly enriched in anatomical structure morphogenesis, vasculature development, blood vessel development, and inflammatory response, while the downregulated genes were significantly enriched in skin development, epidermis development, regulation of water loss

via skin, establishment of skin barrier, and keratinocyte differentiation. However, further experimental follow-ups are required to understand how the differential expression of these genes reflects biological differences in corresponding tissue functions between humans and cattle.

To further explore whether the findings were consistent between humans and mice, we integrated 113 RNA-seq samples from 14 tissues in mice [17, 18]. We found that gene expression profiles of most of tissues were generally conserved among the three mammals (Additional file 2: Fig. S8a), and the differential expression of genes (measured by *t*-statistics) were significantly but moderately correlated between humans *vs.* cattle and humans *vs.* mice (Additional file 2: Fig. S8b-c). We then detected genes that showed conservation ($|FC| < 1.2$ and $FDR > 0.05$) in humans *vs.* cattle, but divergence ($|FC| > 1.2$ and $FDR < 0.05$) in humans *vs.* mice (Additional file 2: Fig. S8d). For instance, those genes in adipose, spleen, lung, and mammary gland were significantly enriched for immune systems, such as T cell activation and regulation of lymphocyte proliferation (Additional file 2: Fig. S8e, Additional file 5: Table S4). This might suggest that cattle show a greater similarity to humans than mice in terms of several aspects of immunophysiology, which was in agreement with previous studies that cattle is a preferred model for human immunology [19, 20]. We also noticed that those genes in heart and liver were significantly involved in muscle contraction, ATP processing, and glucose metabolism, which might be in line with that cattle has been proposed as a model for some muscular disorders, e.g., brody disease [21].

Furthermore, we found that the degree (measured by $-\log_{10}p$) of differential expression of genes between humans and cattle was significantly and negatively correlated with their tissue specificity of expression in most of the tissues within humans (Fig. 3d), suggesting that genes with higher tissue-specific expression are more likely to be transcriptionally conserved (i.e., less differentially expressed) between humans and cattle. However, this was not universal as the opposite trend was found in skin, adrenal, and stomach, suggesting that certain functions of such tissues might be under positive selection in humans and cattle [22]. In addition, we found that dN/dS ratios (measuring DNA sequence conservation) of orthologous genes were weakly but significantly with their Tau values (measuring tissue-specific expression) in humans and cattle (Additional file 2: Fig. S9). We then investigated 30 genes with dN/dS ratio > 1 , considered as positively selected between humans and cattle. Among them, 26 showed tissue-specific expression, and 14 were also significantly differentially expressed in at least one tissue between humans and cattle (Additional file 2: Fig. S10). For instance, *CSN2* and *CSN3*, which are associated with milk production traits in cattle [8], were significantly upregulated in the cattle mammary gland compared to human mammary gland (Fig. 3e). *CCL27*, which participates in T cell-mediated skin inflammation [23], was highly expressed in human skin, but not in cattle skin (Fig. 3e).

Comparison of inter-individual variation of gene expression and their cis-genetic regulatory effects

Like mean gene expression levels, we found that the inter-individual variation of gene expression (measured by median absolute deviation, MAD) was generally conserved in humans and cattle (Fig. 4a). We then sorted all orthologous genes according to their

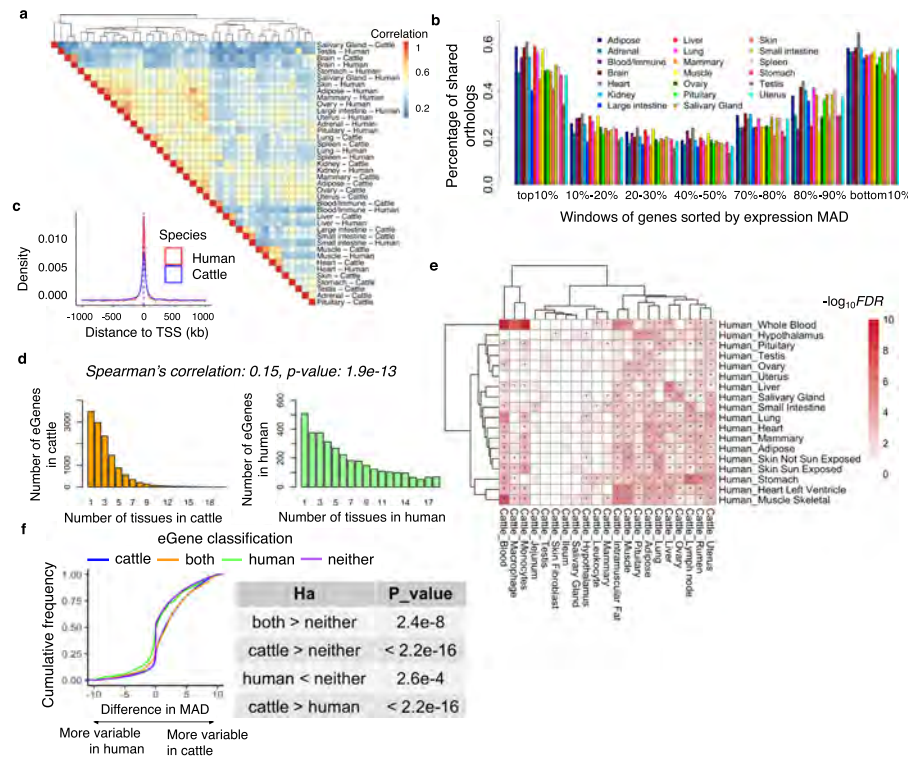


Fig. 4 Comparison of inter-individual variability of gene expression and their *cis*-genetic regulatory effects. **a** Hierarchical clustering of tissues in humans and cattle based on Pearson's correlation of median absolute deviation (MAD) of expression. **b** Percentage of orthologous genes shared in each bin between humans and cattle. Genes were ranked (from largest to smallest) by MAD, and then divided into ten bins (1731 genes per bin). **c** Distribution of top *cis*-eQTLs around transcriptional start sites (TSS) in human and cattle liver. **d** Number of eGenes (genes with significant *cis*-expression quantitative trait loci, *cis*-eQTLs) in what number of tissues in cattle (left) and humans (right). There is a weak but significant correlation (Spearman's $r = 0.15$; $p = 1.91 \times 10^{-13}$) between the number of tissues an eGene was detected on across both species. **e** Enrichment of eGenes between human and cattle tissues. Color represents $-\log_{10}FDR$. P -values are computed using the hypergeometric test for the overlaps of eGenes between human and cattle tissues, and then are adjusted for multiple testing with FDR method. "*" represents $FDR < 0.05$. **f** Distribution of difference in median absolute deviation (MAD) between humans and cattle among four groups of genes in blood, i.e., cattle-specific eGenes (cattle), human-specific eGenes (human), species-shared eGenes (both), and non-eGenes in neither species (neither)

level of variability and found that humans and cattle share most (around 55%, on average) in the top (most variable) and bottom (most consistent) 10% of genes (Fig. 4b). This result was consistent after adjusting for the mean of expression (i.e., the coefficient of variation, CV, which is the ratio of the standard deviation to the mean) (Additional file 2: Fig. S11a, b). The variable genes were significantly engaged in tissue-relevant functions, while consistent genes were significantly involved in essential biological functions, such as system processes and stimulus detection (Additional file 6: Table S5).

Since inter-individual variation of gene expression is partially due to genetic factors, we then compared *cis*-eQTLs of genes across tissues between humans and cattle. We found that compared to all tested SNPs that were evenly distributed around transcription start sites (TSS), top *cis*-eQTLs of eGenes centered around TSS in both humans and cattle (Fig. 4c). However, there was a higher enrichment of *cis*-eQTLs around TSS in humans than in cattle (Additional file 2: Fig. S12), which might be due to the difference

in LD patterns between the two species [24]. For instance, 95% of top *cis*-eQTLs were within 873 kb and 698 kb around TSS in cattle and humans, respectively (Additional file 2: Fig. S12). We found that the majority of eGenes (i.e., genes with *cis*-eQTLs) were tissue-specific (shared with less than five tissues) in humans and cattle (Fig. 4d). We observed a weak but significant correlation (Spearman's $r = 0.15$; $p = 1.91 \times 10^{-13}$) between the number of tissues, in which an eGene was detected on across two species (Fig. 4d). We further observed a significant overlap of eGenes within similar tissues between humans and cattle (Fig. 4e). For instance, eGenes in human blood had the highest enrichment with those in cattle blood, monocytes, and macrophage, and the same was observed for liver, muscle, and heart (Fig. 4e).

Furthermore, we observed that species-specific eGenes had a significantly (one-side Wilcoxon rank-sum test, $p < 2.20 \times 10^{-16}$) higher variability than other genes in the corresponding species (Fig. 4f). Additionally, we found that eGenes showed significantly higher differential expression between humans and cattle than non-eGenes (one-side Wilcoxon rank-sum test, $p < 2.2 \times 10^{-16}$), and conserved eGenes showed significantly higher differential expression than species-specific ones (Additional file 2: Fig. S11c). Overall, this suggests that *cis*-genetic variants may contribute to the inter-species differences in inter-individual variation of gene expression.

Comparison of gene co-expression network

We estimated the conservation of gene co-expression profiles by calculating the correlation of the correlation coefficient (corCor, Methods) of genes between tissues within cattle, between tissues within humans, and within tissues between humans and cattle (Fig. 5a). We found that the overall corCors of genes among tissues within a species were significantly (one-side Student's t test, $p < 1.00 \times 10^{-4}$) higher than those within tissues between species (Fig. 5b). This suggests that gene co-expression networks are less conserved than mean gene expression across species. However, we observed that tissues exhibited distinct conservation levels of gene co-expression between humans and cattle. For instance, muscle and brain showed the highest conservation levels, while ovary, skin, and spleen showed the lowest (Fig. 5c). In addition, we compared the conservation between gene expression and co-expression and found that expression-conserved genes showed significantly (Wilcoxon test, $p < 2.20 \times 10^{-16}$) higher co-expression conservation (i.e., corCors) than expression-diverged genes across tissues (Additional file 2: Fig. S13).

We here took muscle as an example, due to its highest conservation based on corCors, to show the conservation of individual gene co-expression modules between humans and cattle. We first conducted the weighted gene co-expression network analysis (WGCNA) in humans and cattle muscle samples to detect gene co-expression modules, separately (Methods). In general, we found that multiple gene co-expression modules were conserved between species (Fig. 5d–f). Genes in the most conserved module were significantly engaged in fundamental biological processes, such as histone modifications and covalent chromatin modifications. In contrast, genes in the least conserved gene module were significantly involved in skin development and keratinocyte differentiation (Fig. 5g). We repeated the analysis in all the 20 tissues and detected the most conserved and divergent gene co-expression modules, as well as found that these genes in different tissues were significantly enriched in distinct biological functions

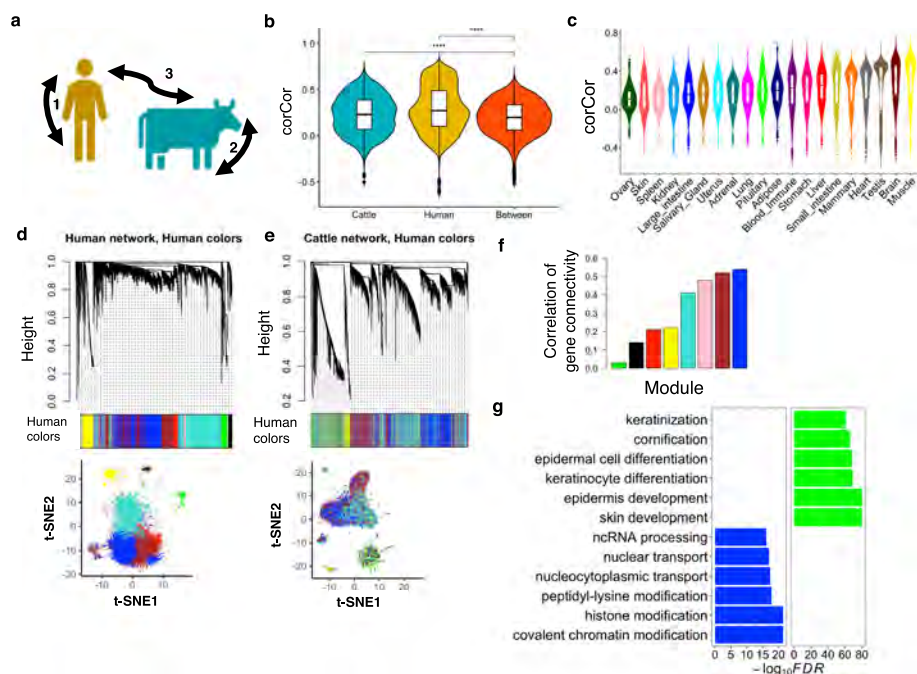


Fig. 5 Comparison of gene co-expression network. **a** The diagram shows three comparisons, i.e., (1) between tissues within humans, (2) between tissues within cattle, and (3) within tissues between species. **b** Comparisons of corCor (measurement of gene co-expression conservation, details in “Methods”) among three groups. “****” represents the $P < 0.0001$ from one-side Student’s t test. **c** Comparisons of corCor in (3) across tissues. **d** The weighted gene co-expression network is constructed in human muscle using WGCNA package (“Methods”). Color represents gene co-expression module. Gene clustering is also visualized through t -SNE method. Each dot in the t -SNE plot represents a gene. **e** Similar with **d**, but the weighted gene co-expression network is constructed in cattle muscle. Genes in the cattle network are assigned same color as they in human modules to reflect the extent of module conservation between species. **f** Bar plot shows correlation of gene connectivity (measuring the conservation of gene co-expression module) between humans and cattle across human co-expression modules. **g** The top significantly ($FDR < 0.05$) enriched Gene Ontology terms for genes in most conserved module (left) and most diverged module (right)

(Additional file 2: Fig. S14-15). For instance, genes of the most diverged module in blood were significantly enriched in neutrophil-mediated immunity, while genes of the most diverged module in brain were significantly enriched in mitochondrial ATP functions (Additional file 2: Fig. S15).

Heritability of complex traits enriched in transcriptionally conserved genes

To better understand the genetic architecture underlying complex traits from an evolutionary point of view, we tested whether transcriptionally conserved genes were more enriched for genetic variants of complex traits than diverged genes (Methods). We analyzed GWAS summary statistics for 46 human complex traits with an average sample size of 327,973, and 45 cattle complex traits with a sample size of 27,214 (Additional file 7: Table S6). After ranking (from the largest to smallest) genes in each tissue according to their degree of differential expression (measured by $-\log_{10}p$) between humans and cattle, we considered the top and bottom 10% as diverged and conserved genes ($n = 1731$), respectively. The distributions of conserved and diverged genes across tissues are shown in Figure S16, and the majority of them were tissue-specific (shared with less than five tissues). In addition, the MAF and LD of SNPs were comparable between conserved and

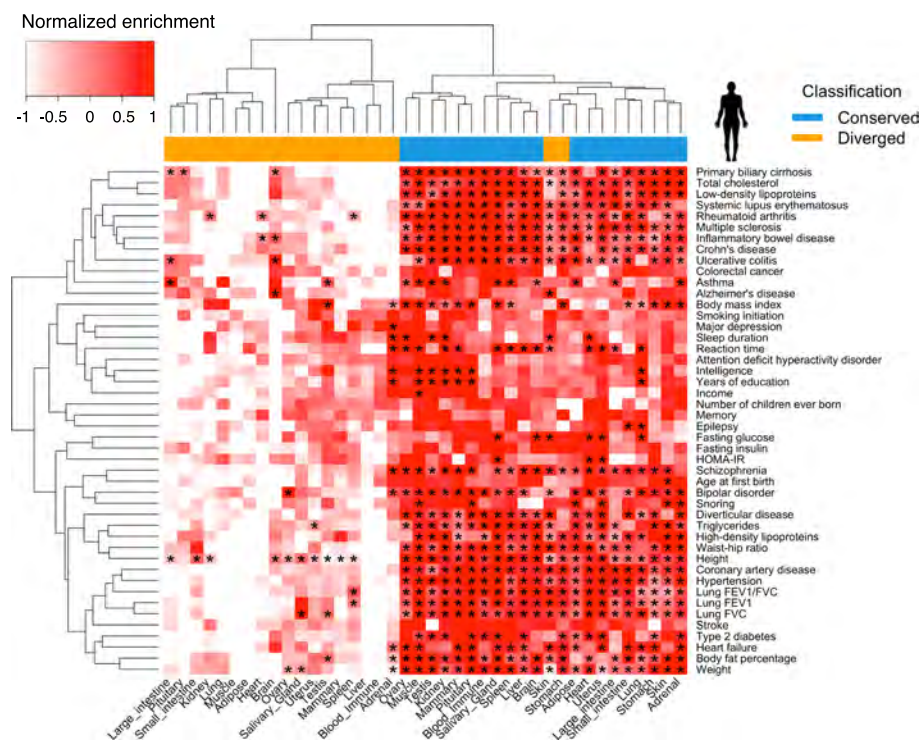


Fig. 6 Heatmap of heritability enrichments of 46 human complex traits in transcriptionally conserved and diverged genes. Heritability enrichments obtained from LDSC for 46 human complex traits in transcriptionally diverged and conserved genes between humans and cattle (“Methods”). All orthologous genes are ranked (from largest to smallest) based on $-\log_{10} p$ obtained from the differential gene expression analysis in each of 20 tissues between humans and cattle. The top and last 10% of genes are considered as transcriptionally diverged and conserved genes in each tissue, respectively. The enrichment is scaled to have mean of zero and variance of one by traits. “*” represents the adjusted P -value (FDR) < 0.05. Traits and tissues are clustered using the Hierarchical clustering method

diverged genes (Additional file 2: Fig. S17). We found that genes with conserved mean expression explained more heritability or enriched more GWAS signals of complex traits than diverged ones (one-side Student’s t test, $p < 2.20 \times 10^{-16}$), and this was consistent across tissues and traits in both humans and cattle (Figs. 6 and 7, Additional files 8, 9 and 10: Table S7-9). We observed similar results for conserved and diverged genes that were detected from inter-individual variation and gene co-expression analyses (Additional file 2: Fig. S18). By further examining GWAS-discovered genes of 4756 complex traits (at least 10 genes per trait) using FUMA [25], we confirmed that conserved genes were significantly enriched for more complex traits GWAS signals than diverged ones, which was consistent across tissues except for skin, adrenal, and stomach (Additional file 2: Fig. S19). In addition to using the sum-based permutation method in cattle, we also employed the three-component GREML-LDMS model to estimate the per-SNP heritability of converged and diverged genes in three milk production traits (i.e., milk, fat and protein yield) (Additional file 10: Table S9), which had the largest sample size and the highest reliability of phenotypes [8, 26]. We found that the expression-conserved genes showed higher per-SNP heritability than DNA sequence-conserved genes and expression-diverged genes across most of the tissues (Additional file 2: Fig. S20a). We also found that the enrichment degrees based on the sum-based permutation test were significantly correlated with

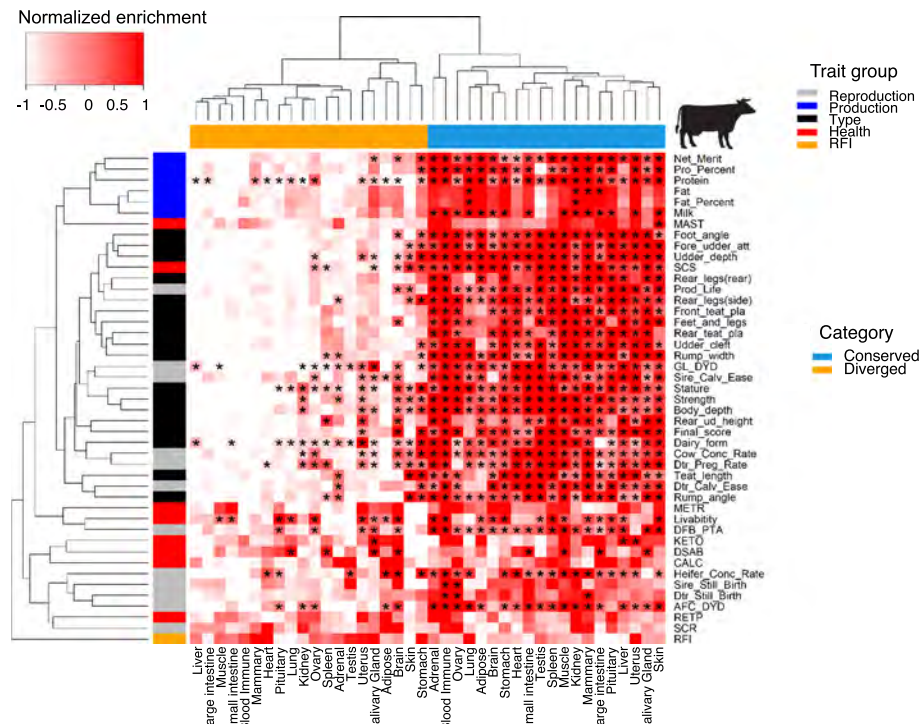


Fig. 7 Heatmap of GWAS signal enrichments of 45 cattle complex traits in transcriptionally conserved and diverged genes. GWAS signal enrichments (i.e., $-\log_{10}p$ from 10,000 times permutation, “Methods”) of cattle complex traits for transcriptionally diverged and conserved genes. All orthologous genes are ranked (from largest to smallest) based on $-\log_{10}p$ obtained from the differential gene expression analysis in each of 20 tissues between humans and cattle. The top and last 10% of genes are considered as transcriptionally diverged and conserved genes in each tissue, respectively. The enrichment is scaled to have mean of zero and variance of one by traits. “*” represents $FDR < 0.05$

per-SNP heritability across tissues for milk and fat yield but not protein yield (Additional file 2: Fig. S20b). For other complex traits in cattle, the GREML-LDMS model could not converge properly across many tissues, mainly due to the variance components being estimated were close to zero. Compared to the GREML-LDMS or LDSC models, the sum-based permutation test only does the GWAS signal enrichment analysis rather than estimate proportions of genetic variance explained [27].

To test if the human-cattle conservation at the transcriptomic level could provide extra information than the conservation at the DNA level, we conducted the same heritability enrichment analysis for sequence-conserved genes (top 10% of genes with the highest sequence conservation between humans and cattle, measured by both Dn/Ds and PhastCons scores) together with expression-conserved genes. As shown in Fig. 8a, although sequence-conserved genes showed the highest enrichment for several traits (e.g., weight and years of education), expression-conserved genes in relevant tissues showed higher enrichments for certain traits. For instance, expression-conserved genes in blood showed the highest enrichment for immune/health traits (e.g., ulcerative colitis, systemic lupus erythematosus, rheumatoid arthritis, and inflammatory bowel disease). Similar findings were observed for genes showing conserved co-expression patterns (Additional file 2: Fig. S21a). For instance, we found that genes

with conserved co-expression in brain showed the highest enrichment for schizophrenia, while genes in small intestine for immune-relevant traits (e.g., rheumatoid arthritis and inflammatory bowel disease), might be due to its immune function (Additional file 2: Fig. S21a). Although the interpretations of some trait-tissue associations were not straightforward due to the complexity in both complex traits and tissues, these results indicate that the transcriptome conservation in relevant tissues could provide additional information for interpreting complex trait genetics. We further compared the heritability enrichment of these 46 human traits for four groups of genes, i.e., the top 10% (most diverged), 40–50%, 50–60%, and bottom 10% (most conserved), ordered by $-\log_{10}\text{FDR}$ (from largest to smallest) from the differential expression analysis between humans and cattle. We found that genes with higher conserved expression showed higher enrichments for the heritability of complex traits, and similar results were observed for gene co-expression (Additional file 2: Fig. S21b).

To investigate whether the transcriptional conservation between humans and cattle could help us identify new causal genes for complex traits, we took the well-studied human height as an example to perform the functionally informed (using conserved genes as functional priors) fine-mapping analysis using PolyFun + SuSiE [28, 29]. Comparing to results from the fine-mapping analysis without conserved genes (i.e., SuSiE [29] only), we fine-mapped more variants/genes for human height (Fig. 8b). Considering $\text{PIP}(\text{PolyFun} + \text{SuSiE}) > 0.95$ but $\text{PIP}(\text{SuSiE}) < 0.95$, we detected 53 variants for human height, out of which 10 were not genome-wide significant ($p > 5 \times 10^{-8}$) in the

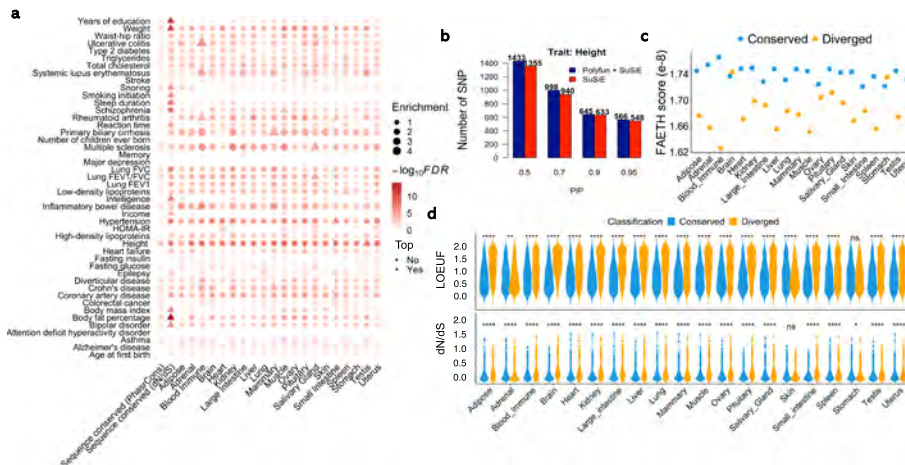


Fig. 8 Transcriptionally conserved genes provide insights into the genetics of complex traits. **a** Heatmap of heritability enrichments obtained from LDSC for 46 human complex traits in transcriptionally and DNA sequence-conserved genes (measured by both dN/dS and PhastCons scores). All orthologous genes are ranked (from largest to smallest) based on $-\log_{10}p$ obtained from the differential gene expression analysis in each of 20 tissues between humans and cattle. The bottom 10% of genes are considered as transcriptionally conserved genes in each tissue, respectively. The 10% of genes with the smallest dN/dS ratios are considered as DNA sequence conserved, whereas the 10% of genes with the largest PhastCons scores are also considered as DNA sequence conserved. In each trait, tissues or sequence-conserved genes with the top heritability enrichment are denoted as triangle, others as dots. **b** Bar plot shows the number of SNPs identified by PolyFun + SuSiE (blue) and SuSiE (red) at different PIP (posterior inclusion probability) cutoffs, respectively. **c** Comparison of FAETH scores of SNPs within transcriptionally conserved and diverged genes. SNPs located within 1000 bp up- and downstream of a gene are included. **d** Violin plot compares the LOEUF scores (up) and the dN/dS ratios (bottom) of transcriptionally conserved and diverged genes across 20 tissues. **, ***, and **** represents the $P < 0.05$, 0.01, 0.0001, respectively, from one-side Student's t test

original GWAS. Out of these 10 variants, six could be mapped to protein-coding genes (Additional file 11: Table S10). By conducting phenome-wide association analysis for these genes using PheWAS (<https://atlas.ctglab.nl/>) [30], we found all these genes were associated with human height or relevant traits (Additional file 11: Table S10). We took *PFKP* and *CYP27B1* as examples in Figure S21c. To explore whether conserved genes could provide useful information in the cattle genomic prediction, we compared the FAETH scores of SNPs within conserved and diverged genes [26], which measures the predictive ability of SNPs for complex traits in dairy cattle. We found that SNPs in conserved genes had higher FAETH scores than those in diverged genes, consistent across all tissues except for stomach and brain (Fig. 8c).

We further explored the properties of transcriptionally conserved and diverged genes as a function of their tolerance to Loss-of-Function (LoF) variants (measured by Loss-of-Function observed/expected upper bound fraction, LOEUF) [31]. We observed that conserved genes had significantly smaller LOEUF scores (i.e., more depleted for LoF variation) compared to diverged genes across tissues, consistent for results from mean gene expression, inter-individual variation of gene expression, and co-expression networks (Fig. 8d, Additional file 2: Fig. S22a). Moreover, compared to diverged genes, we found that conserved genes had significantly smaller dN/dS ratios, indicating that transcriptionally conserved genes also exhibit more constrained protein-coding sequences (Fig. 8d, Additional file 2: Fig. S22b).

Discussion

We comprehensively compared the transcriptomes of 20 tissues in humans and cattle. Despite the differences in experimental conditions and sample characteristics, we found that the mean expression of orthologous genes was, to a certain degree, conserved between humans and cattle. This is consistent with previous findings that the global gene expression pattern of orthologous genes between humans and mice is conserved, particularly for the central nervous system, liver, and heart/muscle [32]. We found that the brain had the highest correlation of median gene expression between humans and cattle, while testis and stomach had the lowest. This is in line with previous findings that suggested that the transcriptome evolves rapidly in testis but slowly in the central nervous system, based on a comparison of the gene expression profiles of six organs across ten mammals [33]. In addition, we investigated whether the gene expression of cattle-specific tissues (e.g., horn and rumen) were significantly correlated with those of human tissues, and found that cattle rumen showed the highest similarity with vagina, esophagus, and skin in humans compared to other tissues, which was due to the high enrichment of epithelial cells in these tissues. Meanwhile, cattle horns showed a low correlation of gene expression across all human tissues, while among them fallopian tube was the most similar one (Additional file 2: Fig. S23a-b).

Additionally, we found that inter-individual variability of gene expression was generally conserved in humans and cattle, which agrees with a previous comparison of gene expression between mice and humans [32]. However, we have taken this further and have shown that *cis*-genetic regulatory effects of gene expression (eGenes) were also conserved between humans and cattle, reflecting that the genetic regulation of gene expression evolves under similar evolutionary pressures among mammals [34]. In

contrast, we found that gene co-expression networks were more conserved among tissues within a species than within corresponding tissues between species, suggesting that changes of gene co-expression networks play important roles in the adaptive evolution of species [2]. Of note, apart from the gene expression, many other functional elements (e.g., enhancers, ncRNAs, TFBS, and translation) and cell type composition might contribute to the difference in phenotypes between species.

The interpretation of the molecular mechanisms underlying complex traits has always been the research focus of genetics. GWASs provide strong evidence that most complex traits are extremely polygenic, yet the distribution of causal variants across the genome remains elusive. Finucane et al. reported that the heritability of complex traits was enriched in genomic regions with constrained DNA sequence across species [35]. We demonstrate that among orthologous genes, transcriptionally conserved genes had significantly higher enrichment for the heritability of complex traits than diverged genes in humans and cattle. We still noted that although on the relative scale, conserved genes seem to be more enriched with heritability than divergent genes, the total amount of heritability explained by conserved genes is not great in either humans or cattle on average across tissues. However, the top tissue for a complex trait could explain a relatively high proportion of heritability. For instance, 8% of SNPs in blood expression-conserved genes could explain 31% and 33% of heritability for inflammatory bowel disease and systemic lupus erythematosus, respectively (Additional file 9: Table S8). This finding suggested that expression-conserved genes contribute to the heritability of complex traits at a tissue-specific manner. Compared to previous studies [26, 35], we found a relatively lower enrichment of heritability in expression-conserved genes than sequence-conserved regions. This may be due to the previous studies considered the sequence-conserved regions in the entire genome, including both genic and intergenic regions, whereas we here only focused on orthologous genes between humans and cattle. Future research, with the increasing availability of functional annotation of animal genomes from the FAANG project [36], will allow examining the conservation of functionally regulatory elements (e.g., enhancer, promoter, and topologically associating domain) and non-coding RNAs in a wide range of tissues/cell types and species, as over 90% of GWAS hits are in non-coding regions [37].

Conclusions

In summary, we showed the conservation of transcriptome among 20 common tissues between humans and cattle. We observed that transcriptionally conserved genes exhibited significantly higher enrichments for the heritability or GWAS signals of complex traits than diverged genes in both species. Our findings provided novel insights into the evolutionary basis of complex traits in humans and cattle.

Methods

RNA-seq samples in humans and cattle

All human RNA-seq samples were analyzed uniformly by human GTEx (v8) consortium previously [10], and the normalized gene expression (TPM) data were obtained in <https://gtexportal.org/home/datasets>. For cattle, we analyzed 11,642 publicly available RNA-seq runs from 8536 samples (by July 2019) using a similar pipeline as human

GTEx [10, 11]. Briefly, we filtered out low-quality reads using Trimmomatic (v0.39) and mapped clean reads to cattle ARS-UCD1.2 reference genome using STAR (v2.7.0). We obtained TPM of all annotated genes ($n = 27,608$) in Ensembl (v96) using Stringtie (v2.1.1). We kept cattle samples with unique mapping reads $> 70\%$ and the number of clean reads $> 800,000$ for subsequent analysis. All gene expression data and the meta-data of samples in cattle were available in <https://cgtext.roslin.ed.ac.uk/>. Ultimately, we obtained normalized gene expression values (TPM) for 10,830 and 4866 RNA-seq samples from 20 common tissues in humans and cattle, respectively. We obtained 17,315 one-to-one orthologous genes and their annotation information from Ensembl (v96).

Sample clustering and differential gene expression analysis

We used the function *IntegrateData(anchorset = expression, dims = 1:30)* in R Seurat package [38] to combine expression values of orthologous genes in humans and cattle by removing hidden confounding factors. Afterward, we performed *t*-distributed stochastic neighbor embedding (t-SNE), implemented in Rtsne [39]: *Rtsne(expression, dims = 2, perplexity=150, theta=0.5, verbose=TRUE, max_iter = 1000, check_duplicates = FALSE, partial_pca = T, num_threads=50)* to project samples to a two-dimensional space based on corrected expression values of orthologous genes. We calculated the median gene expression in each tissue in cattle and humans separately, to represent the “true” expression of the particular tissue in each species. We then performed hierarchical clustering using R package *pheatmap* [40]: *pheatmap(corr_mat, cluster_rows = T, cluster_cols = T, clustering_distance_rows = "correlation", clustering_distance_cols = "correlation")*, to explore the relationship of tissues in humans and cattle based on the median gene expression.

We detected genes with tissue-specific expression using R *Limma* package [41] with function *model.matrix*, *lmFit*, *contrasts.fit*, *eBayes*, and *topTable* by comparing gene expression of samples in a given tissue to those in the remaining tissues. We also employed *Limma* package to detect species-specific genes in each tissue between humans and cattle. *Limma* returned adjusted *P*-values for multiple testing using Benjamini and Hochberg methods (FDR). Here, we used $\log_2(\text{FC}) > 1.5$ and $\text{FDR} < 0.05$ to detect tissue-specific genes. In contrast, we used $\text{FC} > 1.2$ and $\text{FDR} < 0.05$ to identify genes differentially expressed between species, as the differences in gene expression are much bigger between tissues within species than within tissues between species. We also ranked genes according to their degrees of differential expression ($-\log_{10}p$) from DEG analysis between humans and cattle. We then considered the top and last 10% of all orthologous genes as the most diverged and conserved genes for partitioning the heritability of complex traits.

We obtained and analyzed 113 RNA-seq samples from 14 tissues in mice from recount3 (<http://rna.recount.bio/>) [17, 18]. We used *Limma* package [41] to identify species-specific genes for human vs. cattle, and human vs. mouse, similarly as described above.

Detection and comparison of chromatin states between humans and cattle

We analyzed genome-wide sequence data of five epigenetic marks (i.e., ATAC-seq and ChIP-Seq for H3K27ac, H3K27m3, H3K4m1, and H3K4m3) and their corresponding

background inputs in six common tissues (two biological replicates per tissue) in humans and cattle. The tissues included liver, lung, spleen, muscle, brain, and adipose. We downloaded the human data from ENCODE (<https://www.encodeproject.org/>), and cattle data from FAANG (<https://www.faang.org/>). Using BWA algorithm with default settings [42], we mapped human and cattle data to GRCh38 and ARS-UCD1.2 reference genomes, respectively. We then employed a multivariate Hidden Markov Model (HMM), implemented in ChromHMM v1.18 [43], to define 15 chromatin states using 200-bp sliding windows through combining these epigenomic marks across samples in humans and cattle, separately. We calculated the enrichment fold of each chromatin state in TSS ± 2 kb of diverged genes as $(C/A)/(B/D)$, where A is the number of bases in the state, B is the number of bases in TSS ± 2 kb, C is the number of bases overlapped between the state and TSS ± 2 kb, and D is the number of bases in the entire genome.

Detection of differentially variable genes between species

We used the following F-test to conduct differential variability analysis of gene expression in each of 20 tissues between humans and cattle [44]. In a given tissue, $f = \frac{s_1^2}{s_2^2}$, where s_1^2 and s_2^2 are variances of gene expression values (i.e., \log_2 TPM) in humans and cattle, respectively, with the null hypothesis: $s_1^2 = s_2^2$. Under the assumption that the expression of a gene follows a normal distribution, f follows an $F_{(n-1, m-1)}$ distribution (where n and m is the number of human samples and cattle samples, respectively), from which we obtained P -values. We adjusted P -values for multiple testing using Benjamini and Hochberg methods (FDR) with R function `p.adjust(variance_diff$p_value, method = "BH")`. According to their $-\log_{10}$ FDR, we then ranked genes (from largest to smallest) and considered the top and last 10% genes as diverged and conserved genes.

Furthermore, we obtained fine-mapped results of *cis*-eQTLs for similar tissues in humans and cattle from the Human GTEx project [10] (<https://gtexportal.org/home/datasets>) and Cattle GTEx project (<http://cgtex.roslin.ed.ac.uk/>), respectively. We considered genes with significant *cis*-eQTLs ($P < 10^{-5}$) as eGene. We used the hypergeometric test, implemented in *phyper* function in R: `phyper(Overlap-1, human, 17315-human, cattle, lower.tail=FALSE)`, to test the significance of overlaps of eGenes across tissues between species. We adjusted P -values for multiple testing using the Benjamini-Hochberg method (FDR).

Gene co-expression analysis

We employed an R package, *MergeMaid* with function `intCor(merged, method="pearson", exact=F)` [45], to calculate corCors for all orthologous genes in three scenarios, (1) between tissues within cattle, (2) between tissues within humans, (3) within tissues between humans and cattle. For a gene A in an expression matrix of a tissue in a species containing n genes, we computed the Spearman's correlation of expression value between gene A and any other genes, resulting in a vector of length $n-1$ (vector A). Given gene A' is the ortholog of gene A on the other expression matrix (a different tissue or species), we obtained a vector of length $n-1$ (vector A') similarly by calculating Spearman's correlation of A' with any other genes in the same order

as in vector A. We then computed the correlation between vector A and vector A' (corCor), to represent the conservation level of gene A in terms of the co-expression network between two groups. We also applied another R package WGCNA with function *cutreeDynamic(dendro = hierTOM, distM = distTOM, deepSplit = 2, pamRespectsDendro = FALSE, minClusterSize = minModuleSize)* [46], to detect the weighted gene co-expression networks within each tissue in humans and cattle separately. We assigned colors to genes in each co-expression module using function *labels2colors(dynamicMods)*.

Stratified LD score regression (S-LDSC) and POLYgenic FUNctionally informed fine-mapping (PolyFun) analysis for human complex traits

To determine whether transcriptionally conserved genes explain the more genetic variance of complex traits than diverged genes, we employed the commonly used stratified LD score regression to partition the heritability of human complex traits into distinct functional categories [35]. The stratified LD scores were calculated in 500 kb window using 1000G Phase 3 European human samples. Only HapMap3 SNPs with $\text{INFO} \geq 0.9$ and $\text{MAF} > 0.05$ in 1000G European samples were included for LD score calculation. We obtained 1000G samples and default SNP weights from (<https://github.com/bulik/ldsc>).

We collected GWAS summary statistics for 46 human complex traits from a public database (Additional file 6: Table S5). These GWAS are mainly European-ancestry based, with an average sample size of 327,973, a good overlap with HapMap3 panel, a mean χ^2 statistics of > 1.02 and a heritability Z-score of > 4 [47]. For each GWAS summary, default quality control was performed by LDSC to remove GWAS SNPs that are with $\text{MAF} \leq 0.01$, $\text{INFO} \leq 0.9$, genotype call rate ≤ 0.75 , duplicated rsid, out-of-bounds P -value, extreme large χ^2 statistics, strand ambiguous variants, and in discordance with those used in previous LD score calculation³². After filtering, the average number of markers for LDSC regression was over one million. A summary of GWAS used in this study and the LDSC regression results of base model (without partitioning heritability) are available in Tables S6 and S7, respectively.

We tested 41 functional categories for each trait, including 20 groups of the most conserved genes (a group per tissue), 20 groups of the most diverged genes and a group of all SNPs to capture the total heritability. We extended $-/+50$ kb of gene regions to include their *cis*-regulatory regions. We detected the most conserved/diverged genes within each of 20 tissues between humans and cattle in three scenarios below:

- (1) The top 10% (diverged) and last 10% (conserved) of all orthologous genes based on $-\log_{10}P$ (ranked from largest to smallest) from differentially expression analysis between humans and cattle;
- (2) The top 10% (diverged) and last 10% (conserved) of all orthologous genes based on $-\log_{10}P$ (ranked from largest to smallest) from differential variability analysis between humans and cattle.
- (3) The top 10% (conserved) and last 10% (diverged) of all orthologous genes based on corCor scores (ranked from largest to smallest).

PolyFun [28] is an extension of S-LDSC [35] that computes SNP prior causal probabilities via the same statistical framework (Step 1). These prior causal probabilities were then used priors in SuSiE [29] for the fine-mapping (Step 2) analysis. Settings in Step 1 were the same as S-LDSC [35] analysis with two exceptions. First, we only annotated 21 functional categories, including a group of all SNPs to capture the total heritability and 20 groups of the conserved genes between humans and cattle. Second, to gain more power, we used the UK Biobank data as the reference panel and the LD scores were computed using pre-computed UK Biobank LD matrices composed of ~19M SNPs from [28]. In Step 2, we performed fine-mapping analysis using two models in SuSiE [29]. The first model only took into account LD information (i.e., pre-computed UK Biobank LD matrices), whereas the second model considered both LD information and SNP prior causal probabilities estimated from Step 1. We compared how many loci were detected at difference posterior causal probability (PIP) thresholds between these two models.

GWAS signal enrichment analysis for cattle complex traits

We collected GWAS summary statistics from 45 agronomic traits of economic importance in cattle, including reproduction ($n = 12$), production (milk-relevant; $n = 6$), body conformation ($n = 18$), health (immune/metabolic-relevant; $n = 8$) and one feed efficiency trait (i.e., residual feed intake, RFI). For body type, reproduction and production traits, we conducted a single-marker GWAS by fitting a linear mixed model in 27,214 U.S. Holstein bulls as described previously [8]. For health traits, we conducted GWAS using the same method in a subset (ranging from 11,880 for hypocalcemia to 24,699 for livability) of the 27,214 available bulls [48]. GWAS of feed efficiency (i.e., residual feed intake, RFI) was conducted based on 3947 Holstein cows [49].

As linkage disequilibrium (LD) pattern is extremely complicated in the cattle population, we applied a commonly used genotype cyclical permutation method, implemented in QGG package [50], to test the enrichment of cattle GWAS signals in each of the functional categories defined above. Previous studies showed that results from this method were highly correlated with those from LDSC and other GWAS signal enrichment methods [5, 51, 52].

$$T_{sum} = \sum_{i=1}^{m_f} b^2,$$

where m_f is the total number of genomic markers linked to a list of genes (e.g., transcriptionally conserved genes in liver), and b is the marker effect from single-marker GWAS. The markers linked to different genes were often not in LD. We controlled marker-set sizes and LD patterns among markers through applying a genotype cyclical permutation strategy [53]. To obtain an empirical P -value for a gene list, we repeated this permutation procedure 10,000 times and employed a one-tailed test of the proportion of random summary statistics greater than that observed.

In order to explore the patterns of MAF and LD between conserved and diverged groups, we calculated the MAF and LD using PLINK (v.1.9) (--freq and --r2) of 20 gene groups' SNPs.

GREML-LDMS

For cattle, we applied the 3-component GREML-LDMS model below [54] to estimate how much genetic variance in three milk production traits (i.e., milk, fat, and protein yield) could be attributed to common genetic variants within distinct gene groups (e.g., expression-conserved and divergent genes). This analysis included 27,235 individuals and 3,085,572 autosomal variants with MAF > 5% [8].

$$\mathbf{y} = \mu + \mathbf{g}_{con} + \mathbf{g}_{div} + \mathbf{g}_{rest} + \mathbf{e};$$

where \mathbf{y} was the vector of phenotypes of individuals being analyzed. The phenotypes were deregressed transmitting ability, i.e., the additive genetic values of cattle after correcting for all the known fixed effects. μ is global mean, \mathbf{g}_{con} was the vector of polygenic effects for SNPs within conserved genes, where $\mathbf{g}_{con} \sim N(0, \mathbf{G}_{con}\sigma^2_g)$, \mathbf{G}_{con} was the genomic relationship matrix (GRM) calculated by SNPs within conserved genes; \mathbf{g}_{div} was the vector of polygenic effects for SNPs within diverged variants, where $\mathbf{g}_{div} \sim N(0, \mathbf{G}_{div}\sigma^2_g)$, \mathbf{G}_{div} was the GRM calculated by SNPs within diverged genes; \mathbf{g}_{rest} was the vector of polygenic effects for the rest of SNPs, where $\mathbf{g}_{rest} \sim N(0, \mathbf{G}_{rest}\sigma^2_g)$, \mathbf{G}_{rest} was the GRM calculated by the rest variants; and \mathbf{e} was the vector of residual. We applied GREML in GCTA [55] to calculate the heritability of each trait, h^2_{con} and h^2_{div} respectively. For each group, the per-variant h^2 was calculated as the h^2 divided by the number of SNPs in the corresponding group.

Other downstream bioinformatics analysis

We used the hypergeometric test, implemented in *clusterProfiler* R package [56], to explore the function of a list of genes based on Gene Ontology (GO) database. We applied function `bitr(gene_list, fromType="ENSEMBL", toType = c("SYMBOL", "ENTREZID"), OrgDb=org.Hs.eg.db, drop = T)` to translate Ensembl ID to gene symbols, and `enrichGO(gene = gene_cattle$ENTREZID, OrgDb = org.Hs.eg.db, ont = "BP", pAdjustMethod = "BH", minGSSize = 1, pvalueCutoff = 0.05, qvalueCutoff = 0.05, readable = TRUE)` to detect the enriched GO terms. We considered GO terms with FDR < 0.05 as significant.

We utilized `tspx` [57] to calculate the tau score (τ) (ranging from 0 to 1, with 1 for highly tissue-specific genes and 0 for ubiquitously transcribed genes) for each orthologous gene to measure its tissue-specific expression in humans and cattle. In each tissue, we used the median gene expression across all samples to calculate τ scores.

To explore whether transcriptionally conserved/diverged genes were significantly enriched for GWAS signals of complex traits in humans, we performed gene-set enrichment analysis for our conserved/diverged genes on reported gene-sets for a large number of human complex traits and diseases from GWAS-catalog using GENE2FUNC in FUMA (<https://fuma.ctglab.nl/>) [25]. To investigate the association of a gene/variant with a variety of complex traits, we performed genome-wide association analysis using PheWAS (<https://atlas.ctglab.nl>) (<https://atlas.ctglab.nl>) [30], which includes totally 4756 GWAS. Only GWAS traits with Bonferroni-corrected P -value < 0.05 were displayed in the PheWAS plots.

Abbreviations

<i>cis</i> -eQTLs	<i>cis</i> -expression Quantitative Trait Loci
corCor	Correlation of the Correlation coefficient
DEG	Differentially expressed gene
FAANG	Functional Annotation of Animal Genomes
GO	Gene Ontology
GTEX	Genotype-Tissue Expression
GWAS	Genome-Wide Association Studies
LD	Linkage disequilibrium
LDSC	Linkage Disequilibrium Score Regression
LOEUF	Loss-of-function Observed/Expected Upper bound Fraction
MAD	Median absolute deviation

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-022-02745-4>.

Additional file 1: Table S1. Summary of RNA-seq samples in humans and cattle.

Additional file 2: Supplementary Figs. S1–23. Comparative transcriptome in large-scale human and cattle populations.

Additional file 3: Table S2. Significantly enriched Gene Ontology terms for three groups of tissue-specific genes.

Additional file 4: Table S3. Significantly enriched Gene Ontology terms for up-regulated genes in cattle and humans.

Additional file 5: Table S4. Significantly enriched Gene Ontology terms for genes with more conserved expression between human and cattle than between human and mouse.

Additional file 6: Table S5. Significantly enriched Gene Ontology terms for genes with variable and consistent expression across tissues in humans and cattle.

Additional file 7: Table S6. Summary of 46 GWAS in humans.

Additional file 8: Table S7. Summary of LDSC results of base model (without partitioning heritability) for 46 human complex traits.

Additional file 9: Table S8. Heritability enrichment analysis of expression-conserved and divergent genes in human complex traits using LDSC.

Additional file 10: Table S9. Partitioning heritability with expression-conserved and divergent genes in milk production traits using GREML-LDMS.

Additional file 11: Table S10. Summary of novel variants detected by PolyFun + SuSiE in human height.

Additional file 12. Peer review history.

Acknowledgements

We thank Professor Chris Ponting (MRC Human Genetic Unit, The University of Edinburgh) and Dr. Paul M. Vanraden (ARS, USDA) for the valuable comments and suggestions. We thank US dairy producers for providing phenotypic, genomic, and pedigree data through the Council on Dairy Cattle Breeding under ARS-USDA Material Transfer Research Agreement 58-8042-8-007. Access to 1000 Bull Genomes Project data was provided under ARS-USDA Data Transfer Agreement 15443. International genetic evaluations were calculated by the International Bull Evaluation Service (Interbull; Uppsala, Sweden).

Review history

The review history is available as Additional file 12.

Peer review information

Tim Sands was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

L.F., A.T., and G.E.L. conceived and designed the project. Y.Yao, S.L., C. X., Y.G., Z.P., O.C.-X., S.W., B.L., L.F., and X.L. performed bioinformatics analyses. O.C.-X., A.K., K.R., L.F., Y.Z., E.P.-C., K. D., Z.Y., C.-J.L., Y.Yu, S.Z., L.M., J.B.C., P.J.R., H.Z., C.H., and G.E.L. contributed to the resource generation. Y.Yao and L.F. drafted the manuscript. All authors read, edited, and approved the final manuscript.

Funding

A. Khamseh was supported by the XDF program from the University of Edinburgh and Medical Research Council (MC_UU_00009/2). A. Tenesa acknowledged funding from the BBSRC through program grants BBS/E/D/10002070 and BBS/E/D/30002275, MRC research grant MR/P015514/1, and HDR-UK award HDR-9004. O. Canela-Xandri was supported by MR/R025851/1. L. Fang. was partially funded through HDR-UK award HDR-9004 and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 801215. This project was partially supported by Agriculture and Food Research Initiative Competitive Grant no. 2020-67015-31398 and 2021-67015-33409 from the USDA National Institute of Food and Agriculture. Y. Zhang. was supported by the earmarked fund for CARS36.

This work was supported in part by AFRI grant numbers 2013-67015-20951, 2016-67015-24886, and 2019-67015-29321, 2020-67015-31398, and 2021-67015-33409 from the USDA National Institute of Food and Agriculture (NIFA) and BARD

grant number US-4997-17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund. G.E.L. and C.P.V.T. were supported by appropriated project 8042-31000-001-00-D, "Enhancing Genetic Merit of Ruminants Through Improved Genome Assembly, Annotation, and Selection" of the Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA). C.-J.L. was supported by appropriated project 8042-31310-078-00-D, "Improving Feed Efficiency and Environmental Sustainability of Dairy Cattle through Genomics and Novel Technologies" of ARS-USDA. J.B.C. was supported by appropriated project 8042-31000-002-00-D, "Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals" of ARS-USDA. This research used resources provided by the SCINet project of the USDA ARS project number 0500-00093-001-00-D. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Availability of data and materials

All gene expression data analyzed in this study are publicly available in <https://gtexportal.org/home/datasets> [58] for humans and <https://cgtex.roslin.ed.ac.uk/> [59] for cattle. All scripts codes used in this study can be found in <https://github.com/B160389-2019/Comparative-Project> [60].

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, EH4 2XU Edinburgh, UK. ²School of Informatics, The University of Edinburgh, Edinburgh EH8 9AB, UK. ³Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, Maryland 20705, USA. ⁴College of Animal Science and Technology, China Agricultural University, Beijing 100193, China. ⁵The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, UK. ⁶Department of Psychology, 7 George Square, The University of Edinburgh, Edinburgh EH8 9JZ, UK. ⁷Department of Animal and Avian Sciences, University of Maryland, College Park, MA 20742, USA. ⁸Department of Animal Science, University of California, Davis, CA 95616, USA. ⁹Present address: Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China. ¹⁰State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, Yunnan, China. ¹¹Scotland's Rural College (SRUC), Roslin Institute Building, Midlothian EH25 9RG, UK. ¹²Guangdong Provincial Key Laboratory of Waterfowl Healthy Breeding, College of Animal Science & Technology, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, Guangdong, China. ¹³Present address: Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark.

Received: 17 December 2020 Accepted: 9 August 2022

Published online: 22 August 2022

References

- Breschi A, Gingeras TR, Guigo R. Comparative transcriptomics in human and mouse. *Nat Rev Genet.* 2017;18:425–40.
- Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc Natl Acad Sci U S A.* 2006;103:17973–8.
- Raymond B, Yengo L, Costilla R, Schrooten C, Bouwman AC, Hayes BJ, et al. Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock. *PLoS Genet.* 2020;16:e1008780.
- Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat Genet.* 2018;50:362–7.
- Liu S, Yu Y, Zhang S, Cole JB, Tenesa A, Wang T, et al. Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of complex traits in cattle and human. *BMC Biol.* 2020;18:80.
- Subramanian S. Deleterious protein-coding variants in diverse cattle breeds of the world. *Genet Sel Evol.* 2021;53:80.
- Fang L, Cai W, Liu S, Canela-Xandri O, Gao Y, Jiang J, et al. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res.* 2020;30:790–801.
- Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun Biol.* 2019;2:212.
- Fang L, Liu S, Liu M, Kang X, Lin S, Li B, et al. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol.* 2019;17:1–16.
- Consortium GTEx. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369:1318–30.
- Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, et al. A multi-tissue atlas of regulatory variants in cattle. *Nat Genet.* 2022.

12. Cho C. Testicular and epididymal ADAMs: expression and function during fertilization. *Nat Rev Urol.* 2012;9:550–60.
13. Chuma S, Hosokawa M, Kitamura K, Kasai S, Fujioka M, Hiyoshi M, et al. Tdrd1/Mtr-1, a tudor-related gene, is essential for male germ-cell differentiation and nuage/germinal granule formation in mice. *Proc Natl Acad Sci U S A.* 2006;103:15894–9.
14. Li Q, Qiao D, Song NH, Ding Y, Wang ZJ, Yang J, et al. Association of DAZ1/DAZ2 deletion with spermatogenic impairment and male infertility in the South Chinese population. *World J Urol.* 2013;31:1403–9.
15. Menezo YJ, Herubel F. Mouse and bovine models for human IVF. *Reprod BioMed Online.* 2002;4:170–5.
16. Gu X, Su Z. Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc Natl Acad Sci U S A.* 2007;104:2779–84.
17. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol.* 2021;22:323.
18. Lin S, Lin Y, Nery JR, Urlich MA, Breschi A, Davis CA, et al. Comparison of the transcriptional landscapes between human and mouse tissues. *Proc Natl Acad Sci U S A.* 2014;111:17224–9.
19. Baldwin CL, Telfer JC. The bovine model for elucidating the role of gammadelta T cells in controlling infectious diseases of importance to cattle and humans. *Mol Immunol.* 2015;66:35–47.
20. Hein WR, Griebel PJ. A road less travelled: large animal models in immunological research. *Nat Rev Immunol.* 2003;3:79–84.
21. Mascarello F, Sacchetto R. Structural study of skeletal muscle fibres in healthy and pseudomyotonia affected cattle. *Ann Anat.* 2016;207:21–6.
22. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, et al. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science.* 2005;309:1850–4.
23. Homey B, Alenius H, Muller A, Soto H, Bowman EP, Yuan W, et al. CCL27-CCR10 interactions regulate T cell-mediated skin inflammation. *Nat Med.* 2002;8:157–65.
24. Qanbari S. On the extent of linkage disequilibrium in the genome of farm animals. *Front Genet.* 2019;10:1304.
25. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2018;9:1826.
26. Xiang R, Berg IVD, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc Natl Acad Sci U S A.* 2019;116:19398–408.
27. Fang L, Sahana G, Ma P, Su G, Yu Y, Zhang S, et al. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol.* 2017;49:44.
28. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat Genet.* 2020;52:1355–63.
29. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Ser B (Stat Methodol).* 2020;82:1273–300.
30. Watanabe K, Stringer S, Frei O, Umicovic Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet.* 2019;51:1339–48.
31. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581:434–43.
32. Zheng-Bradley X, Rung J, Parkinson H, Brazma A. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.* 2010;11:R124.
33. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, et al. The evolution of gene expression levels in mammalian organs. *Nature.* 2011;478:343–8.
34. Fair BJ, Blake LE, Sarkar A, Pavlovic BJ, Cuevas C, Gilad Y. Gene expression variability in human and chimpanzee populations share common determinants. *Elife.* 2020;9:e59929.
35. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet.* 2015;47:1228–35.
36. Giuffra E, Tuggle CK, Consortium F. Functional Annotation of Animal Genomes (FAANG): current achievements and roadmap. *Annu Rev Anim Biosci.* 2019;7:65–88.
37. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 2016;44:D877–81.
38. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. *Cell.* 2019;177(1888–1902):e1821.
39. Van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res.* 2008;9:11.
40. Kolde R, Kolde MR. Package 'pheatmap'. *R package.* 2015;1:790.
41. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43:e47.
42. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25:1754–60.
43. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods.* 2012;9:215–6.
44. Ho JW, Stefani M, dos Remedios CG, Charleston MA. Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics.* 2008;24:i390–8.
45. Cope L, Zhong X, Garrett E, Parmigiani G. MergeMaid: R tools for merging and cross-study validation of gene expression data. *Stat Appl Genet Mol Biol.* 2004;3:Article29.
46. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9:559.
47. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47:291–5.
48. Freebern E, Santos DJA, Fang L, Jiang J, Parker Gaddis KL, Liu GE, et al. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. *BMC Genomics.* 2020;21:41.

49. Li B, Fang L, Null DJ, Hutchison JL, Connor EE, VanRaden PM, et al. High-density genome-wide association study for residual feed intake in Holstein dairy cattle. *J Dairy Sci.* 2019;102:11067–80.
50. Rohde PD, Fourie Sorensen I, Sorensen P. qgg: an R package for large-scale quantitative genetic analyses. *Bioinformatics.* 2020;36:2614–5.
51. Sorensen IF, Edwards SM, Rohde PD, Sorensen P. Multiple trait covariance association test identifies gene ontology categories associated with chill coma recovery time in *Drosophila melanogaster*. *Sci Rep.* 2017;7:2413.
52. Fang L, Sahana G, Su G, Yu Y, Zhang S, Lund MS, et al. Integrating sequence-based GWAS and RNA-seq provides novel insights into the genetic basis of mastitis and milk production in dairy cattle. *Sci Rep.* 2017;7:45560.
53. Rohde PD, Demontis D, Cuyabano BC, Genomic Medicine for Schizophrenia G, Borglum AD, Sorensen P. Covariance Association Test (CVAT) identifies genetic markers associated with schizophrenia in functionally associated biological processes. *Genetics.* 2016;203:1901–13.
54. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet.* 2015;47:1114–20.
55. Yang J, Lee SH, Wray NR, Goddard ME, Visscher PM. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proc Natl Acad Sci.* 2016;113:E4579–80.
56. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16:284–7.
57. Antonio PC, Adrielle AV, Mateus BF, Gonalo AGP, Marcelo FC. tspex: a tissue-specificity calculator for gene expression data. *Research Square.* 2020.
58. GTEx Analysis V9. <https://gtexportal.org/home/datasets>. Accessed 9 Aug 2022.
59. The cattle Genotype-Tissue Expression atlas. <https://cgtex.roslin.ed.ac.uk/>. Accessed 9 Aug 2022.
60. Yao, Y., Liu, S., Xia, C., Gao, Y., Pan, Z., Canela-Xandri, O. et al. Comparative transcriptome between human and cattle. *GitHub.* 2022. <https://github.com/B160389-2019/Comparative-Project>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions





A multi-tissue atlas of regulatory variants in cattle

Shuli Liu^{1,2,3,15}, Yahui Gao^{1,4,15}, Oriol Canela-Xandri^{5,15}, Sheng Wang^{6,15}, Ying Yu^{2,15}, Wentao Cai⁷, Bingjie Li⁸, Ruidong Xiang^{9,10}, Amanda J. Chamberlain¹⁰, Erola Pairo-Castineira^{10,11}, Kenton D'Mellow⁵, Konrad Rawlik^{10,11}, Charley Xia¹¹, Yuelin Yao⁵, Pau Navarro^{10,11}, Dominique Rocha¹², Xiujin Li¹³, Ze Yan², Congjun Li^{10,11}, Benjamin D. Rosen^{10,11}, Curtis P. Van Tassell^{10,11}, Paul M. Vanraden¹, Shengli Zhang^{2,16}, Li Ma^{10,11}, John B. Cole^{10,11}, George E. Liu¹✉, Albert Tenesa^{10,11}✉ and Lingzhao Fang^{10,11,14}✉

Characterization of genetic regulatory variants acting on livestock gene expression is essential for interpreting the molecular mechanisms underlying traits of economic value and for increasing the rate of genetic gain through artificial selection. Here we build a Cattle Genotype-Tissue Expression atlas (CattleGTEx) as part of the pilot phase of the Farm animal GTEx (FarmGTEx) project for the research community based on 7,180 publicly available RNA-sequencing (RNA-seq) samples. We describe the transcriptomic landscape of more than 100 tissues/cell types and report hundreds of thousands of genetic associations with gene expression and alternative splicing for 23 distinct tissues. We evaluate the tissue-sharing patterns of these genetic regulatory effects, and functionally annotate them using multiomics data. Finally, we link gene expression in different tissues to 43 economically important traits using both transcriptome-wide association and colocalization analyses to decipher the molecular regulatory mechanisms underpinning such agronomic traits in cattle.

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits in human and livestock populations^{1,2}. Because the majority of these variants are noncoding, characterization of the molecular mechanisms by which such variants affect complex traits has been extremely challenging. Indeed, in human genetics, undertakings such as the Genotype-Tissue Expression (GTEx) project have characterized genetic effects on the human transcriptome and paved the way to understanding the molecular mechanisms of human variation³.

However, livestock genomic resources lag behind human genomic resources, and to date, no study has systematically explored regulatory variants of the transcriptome across a wide range of tissues. GWAS signals of agronomic traits are significantly enriched in the regulatory regions of genes expressed in trait-relevant tissues in cattle^{4–6}, but studies of genetic variation in gene expression have generally been small, in terms of both the number of individuals and the number of tissues. For instance, previous studies have explored expression/splicing quantitative trait loci (e/sQTL) in blood⁷, milk cells⁸, muscle⁹ and mammary gland in cattle⁹.

There has been a recent polynomial growth in the number of RNA-seq samples made publicly available in cattle (Extended Data Fig. 1a), but these data have not been uniformly processed and jointly analyzed before. Here, we present a pipeline to uniformly integrate 7,180 public RNA-seq samples, representing more than 100 different

tissues and cell types, and we identify eQTLs and sQTLs for 23 distinct cattle tissues with sufficient sample sizes ($n \geq 40$). The latter is facilitated by calling variants directly from the RNA-seq reads and imputing to the sequence level using a large multibreed reference panel¹⁰, in a process similar to that used with human data¹¹. Next, we conducted *in silico* analyses to annotate eQTLs and sQTLs with a variety of omics data in cattle, including DNA methylation, chromatin states and chromatin conformation characteristics. Finally, we integrated gene expression with a large GWAS of 27,214 dairy bulls and 43 cattle traits *via* both transcriptome-wide association study (TWAS) and colocalization analyses to detect genes and variants associated with these economically important traits. We make the results freely and easily accessible to the research community through a web portal (<http://cgtex.roslin.ed.ac.uk/>). This CattleGTEx atlas, as part of the Farm animal GTEx (FarmGTEx) project, will serve as a primary reference for cattle genomics, breeding, adaptive evolution, veterinary medicine and comparative genomics.

Results

Data summary. We analyzed 8,653 public RNA-seq samples, yielding ~200 billion clean reads. Data summary details are shown in Extended Data Fig. 1b–i and Supplementary Table 1. We retained 7,180 samples with sufficient quality (Methods) for subsequent analyses, representing 114 tissues from 46 breeds and breed combinations.

¹Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, MD, USA. ²National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China.

³School of Life Sciences, Westlake University, Hangzhou, China. ⁴Department of Animal and Avian Sciences, University of Maryland, College Park, MD, USA. ⁵MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh, UK. ⁶State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China. ⁷Institute of Animal Science, Chinese Academy of Agricultural Science, Beijing, China. ⁸Scotland's Rural College (SRUC), Roslin Institute Building, Midlothian, UK. ⁹Faculty of Veterinary & Agricultural Science, The University of Melbourne, Parkville, Victoria, Australia. ¹⁰Agriculture Victoria, AgriBio, Centre for AgriBiosciences, Bundoora, Victoria, Australia. ¹¹The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian, UK. ¹²INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France. ¹³Guangdong Provincial Key Laboratory of Waterfowl Healthy Breeding, College of Animal Science & Technology, Zhongkai University of Agriculture and Engineering, Guangzhou, China. ¹⁴Present address: Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark. ¹⁵These authors contributed equally: Shuli Liu, Yahui Gao, Oriol Canela-Xandri, Sheng Wang, Ying Yu. ¹⁶Deceased: Shengli Zhang. ✉e-mail: George.Liu@usda.gov; Albert.Tenesa@ed.ac.uk; lingzhao.fang@qgg.au.dk

The most represented breed was Holstein (35.5% of all samples), reflecting its global economic value. A total of 1,831 samples (21%) had no breed records, but this information could be inferred from the genotypes called from RNA-seq data. We grouped the 114 tissues into 13 categories based on known biology and the 46 breeds into six ancestry groups, with *Bos taurus* representing 87% of all samples (Supplementary Table 1). To investigate the tissue specificity of DNA methylation for functionally annotating QTLs, we also uniformly analyzed 144 whole-genome bisulfite sequence (WGBS) samples from 21 cattle tissues, producing ~73 billion clean reads with an average mapping rate of 71% (Supplementary Table 2).

General characteristics of the transcriptome across samples.

As expected, the number of expressed genes (transcripts per million (TPM) >0.1) increased with the number of clean reads across samples. However, we observed a plateau at 50 million clean reads (Extended Data Fig. 2a) where we detected only ~60% of 27,607 Ensembl annotated genes. Only 61 genes were not expressed in any of the samples, and 33 of them (54.10%) were located in unplaced scaffolds, with significantly shorter gene length, fewer exons, higher CG density and lower sequence constraints than expressed genes (Extended Data Fig. 2b–f). Similarly, we detected more alternative splicing events with increasing numbers of clean reads across samples (Extended Data Fig. 2g). However, we did not detect splicing events for 874 genes in any sample, which also exhibited shorter gene length, fewer exons, lower expression and lower sequence constraints than spliced genes (Extended Data Fig. 2h–k). Furthermore, 27% of genes without splicing events were small nuclear RNAs, small nucleolar RNAs and ribosomal RNAs that have important roles in RNA splicing¹² (Extended Data Fig. 2l). Genes without splicing events were significantly enriched in the integral component of the membrane and G-protein coupled receptor signaling pathways (Extended Data Fig. 2m). We found that ~25% of CpG sites in the entire genome were not covered at 5× in any of the WGBS samples, even if these had more than 300 million clean reads, partially because of bisulfite treatment and polymerase chain reaction amplification bias (Extended Data Fig. 3a). These CpG sites were enriched in gene deserts (for example, telomeres) with significantly higher CG density than the CpG sites captured by the WGBS (Extended Data Fig. 3b,c).

We called a median of 21,623 single nucleotide polymorphisms (SNPs) from all RNA-seq samples (Extended Data Fig. 4a), and then imputed each sample up to 3,824,444 SNPs using a multibreed reference population of 3,310 animals¹⁰. We validated the imputation accuracy by comparing SNPs derived from RNA-seq with those called from whole-genome sequences (WGS) in the same individuals, including Holstein, Limousin and Angus breeds, and the concordance rates were over 99% (Extended Data Fig. 4b, and Supplementary Table 3). We also compared the imputed genotypes from RNA-seq data with those imputed using 50K SNP array genotypes in 109 Holstein animals. Although there was a depletion of high-quality (dosage R -squared (DR^2) > 0.80) imputed intergenic variants among SNPs imputed from RNA-seq only (Extended Data Fig. 4c), the imputation accuracies of SNPs from RNA-seq were similar to those from the SNP array 1 Mb up-/downstream of the gene body (Extended Data Fig. 4d). In addition, the correlation of genotype counts between imputed SNPs from RNA-seq data and those from the SNP array was around 0.80 (Extended Data Fig. 4e). For subsequent *cis*-QTL mapping, we focused on 23 distinct tissues with more than 40 individuals after removing duplicate samples within each tissue (Extended Data Fig. 4f), and this encompassed 4,889 samples.

We found that clusters of samples derived from both gene expression and alternative splicing could accurately recapitulate tissue types (Fig. 1a,b), reinforcing the quality and therefore their utility for our follow-up analysis. For instance, all the muscle samples

from more than 40 projects clustered together. Similar to expression and splicing, DNA methylation profiles also recapitulated tissue types (Fig. 1c). When clustering was based on imputed genotypes, as expected, samples clustered by ancestry (Fig. 1d).

Tissue specificity of transcriptome and methylome. The tissue specificity of gene expression was conserved between cattle and humans (Fig. 2a), and the function of genes with tissue-specific expression accurately reflected known tissue biology. For instance, brain-specific genes were significantly enriched for synapse and neuron function, and testis-specific genes for spermatogenesis and reproduction (Extended Data Fig. 5a). We also calculated the tissue specificity of promoter DNA methylation and gene alternative splicing. Similarly, the function of genes with tissue-specific promoter hypomethylation and splicing reflected the known tissue biology (Extended Data Fig. 5b,c). We found that, based on tissue specificity, gene expression level was significantly and negatively correlated with DNA methylation level in promoters (Fig. 2b), and positively correlated with the splicing ratios of introns (Fig. 2c). For example, *CEL2F2*, a brain-related gene, had significantly higher expression, lower promoter DNA methylation and a higher splicing ratio of the first intron in brain than in other tissues considered (Fig. 2d). Tissue-specific genes exhibited distinct patterns of sequence constraints (Extended Data Fig. 5d), supporting the hypothesis of tissue-driven genome evolution⁴. We found that whereas brain-specific genes evolve slowly, blood- or testis-specific genes evolve rapidly. This trend was also observed within tissue-specific hypomethylated regions (Extended Data Fig. 5e,f).

Discovery of expression and splicing QTLs. We identified *cis*-e/sQTLs for 23 distinct tissues with 40 or more individuals, while accounting for relevant confounding factors and multiple testing (Extended Data Fig. 6a,b). The number of genes with significant *cis*-eQTLs (eGenes) discovered ranged from 172 in ileum to 10,157 in blood, with 19,559 (83% of all 23,523 tested genes) classed as eGenes in at least one tissue (Supplementary Table 4). The number of genes with significant *cis*-sQTLs (sGenes) discovered ranged from 4 in the salivary gland to 7,913 in macrophages, with 15,376 (70.8%) classed as sGenes in at least one tissue. Genes with no *cis*-eQTL or *cis*-sQTL in any of the tissues were significantly enriched in hormone activity, regulation of receptor activity, neuropeptide signaling pathway and reproduction (Supplementary Tables 5–7). In general, the greater the number of samples for the tissue, the larger the number of *cis*-e/sGenes detected (Fig. 3a,b). As expected, with a larger sample size, we had more power to detect *cis*-eQTLs with smaller effect sizes (Extended Data Fig. 6c,d). Consistent with findings in humans¹³, significant variants (eVariants) centered around the transcript start sites (TSS) of measured genes (Extended Data Fig. 6e,f). Across 23 distinct tissues, an average of 46% (range 25.5%–76.6%) of eVariants were found within 100 kb around the TSS of target genes. In non-eGenes, there was also an enrichment of SNPs with the smallest P values (although not statistically significant at a false discovery rate (FDR) of 0.05) around the TSS, suggesting a lack of power to detect such associations for those genes (Extended Data Fig. 6e). Furthermore, we fine-mapped eGenes to assess whether the identified signals could be attributed to one or more causal SNPs. We found that an average of 46% (range 14.5%–73.9%) of eGenes across 23 tissues had more than one independent *cis*-eQTLs (Fig. 3c), indicating complex genetic control of gene expression. SNPs with larger effects within a locus tended to be closer to the TSS (Fig. 3d). To complement and validate the *cis*-eQTL analysis within individuals, we conducted allele-specific expression (ASE) analysis, and found that *cis*-eQTLs were significantly overrepresented in loci with significant (FDR < 0.05) ASE (Fig. 3e); the effect sizes of *cis*-eQTLs were significantly correlated with those of ASEs (Fig. 3f and Extended Data Fig. 6g).

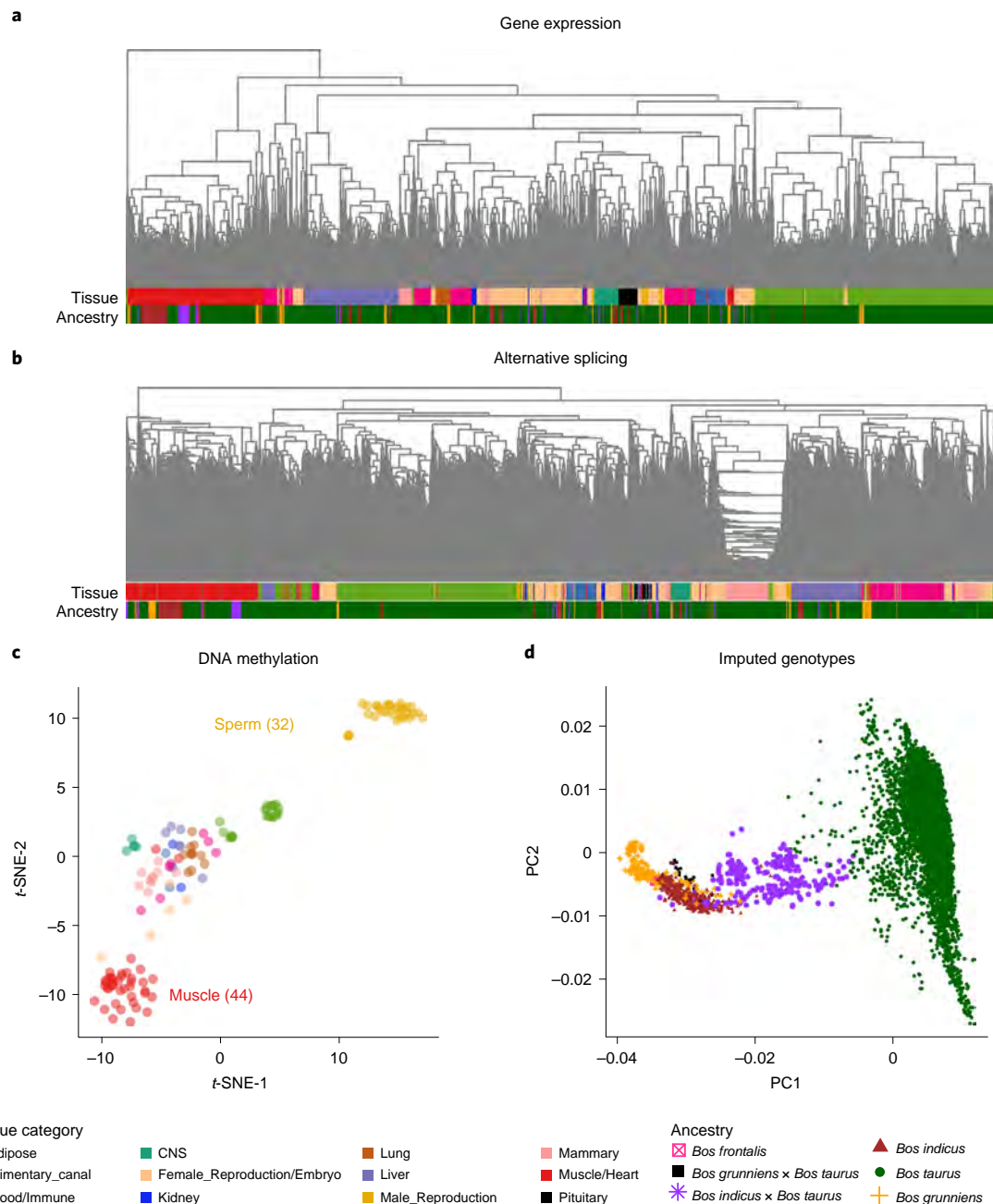


Fig. 1 | Hierarchical clustering and PCA of samples. a, Sample ($n=7,180$) hierarchical clustering based on the expression levels of all transcribed genes (TPM > 0.1). **b**, Sample ($n=7,180$) hierarchical clustering based on the alternative splicing value (percent spliced in(PSI)) of spliced introns. **c**, Sample ($n=144$) clustering using t -distributed stochastic neighbor embedding coordinates based on DNA methylation levels of CpG sites (coverage $\geq 5\times$). **d**, Principal component analysis (PCA) of samples ($n=7,180$) based on imputed genotypes. CNS, central nervous system.

To investigate whether *cis*-eQTLs are conserved among breeds, we conducted *cis*-eQTL mapping for muscle samples from *Bos indicus*, *B. taurus* and their hybrids separately, yielding 86, 2,766 and 800 eGenes, respectively. We observed that *cis*-eQTLs were more conserved across breeds than across tissues (Fig. 3g). For example, the expression of *NMRAL1* in muscle was consistently and significantly regulated by a *cis*-eQTL (rs208377990) among *B. indicus*, *B. taurus* and their hybrids (Fig. 3h). Combining the summary statistics of each breed in a meta-analysis showed that eGene–eVariant associations identified in one breed are potentially transferable to other breeds, particularly for SNPs with a larger effect size (Extended Data Fig. 6h,i). Combining samples from

different breeds will increase the statistical power to detect shared *cis*-eQTLs, and enable more accurate mapping of the causal variants by reducing the linkage disequilibrium (LD) patterns. In total, 131 of 437 eGene–eVariant pairs that were specifically discovered in *B. indicus* showed significant (FDR < 0.05) genotype × breed interactions (Supplementary Table 8). For instance, the expression of an immune-related gene, *SSNA1*, was regulated by a *cis*-eQTL (rs110492559) in *B. indicus*, but not in *B. taurus* or the hybrids, showing a significant genotype × breed interaction (Fig. 3i). In addition, we found that breed-specific *cis*-eQTLs had a lower minor allele frequency (MAF) than breed-common *cis*-eQTLs, in both *B. indicus* and *B. taurus* (Extended Data Fig. 7a,b). This may indicate

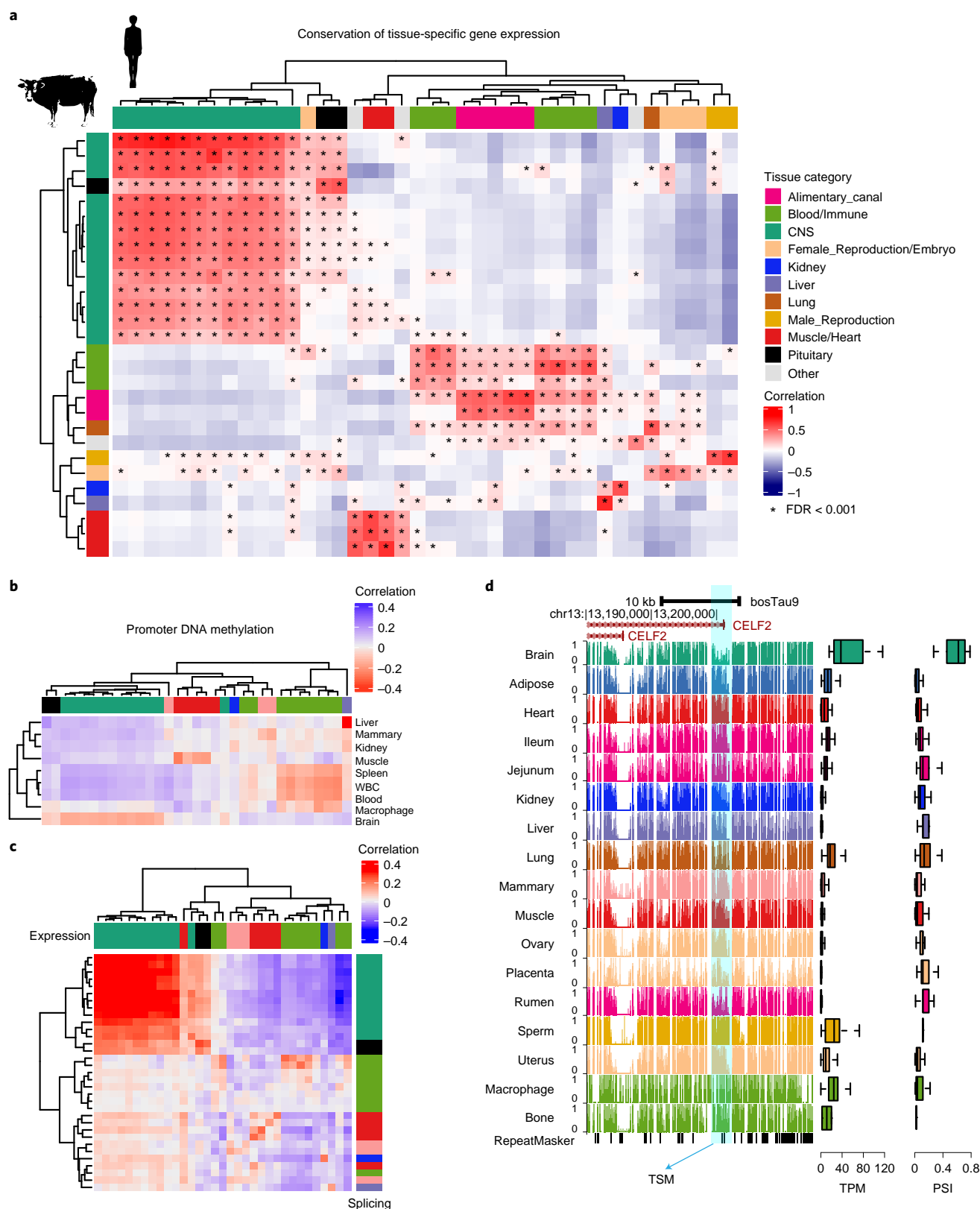


Fig. 2 | Tissue specificity of gene expression, alternative splicing and DNA methylation. **a**, Pearson correlation of the tissue specificity (measured as t -statistics) of 22,752 orthologous genes between cattle and human tissues (GTEx v.8)³. Multiple testing is corrected for use of the Benjamini-Hochberg method. * indicates the false discovery rate (FDR) < 0.001. **b**, Pearson correlation of tissue specificity between gene expression (x axis) and promoter DNA methylation levels (y axis). WBC, white blood cells. The color code for tissues is the same as in **a**. **c**, Pearson correlation of tissue specificity between gene expression (TPM, x axis) and alternative splicing (PSI, y axis). The color code for tissues is the same as in **a**. **d**, *CELF2* shows lower DNA methylation levels in splice sites (right), higher gene expression (middle) and a higher PSI value of spliced introns (left) in brain tissue ($n=15$) compared with the rest of the tissues. chr, chromosome; TSM, tissue-specific methylation.

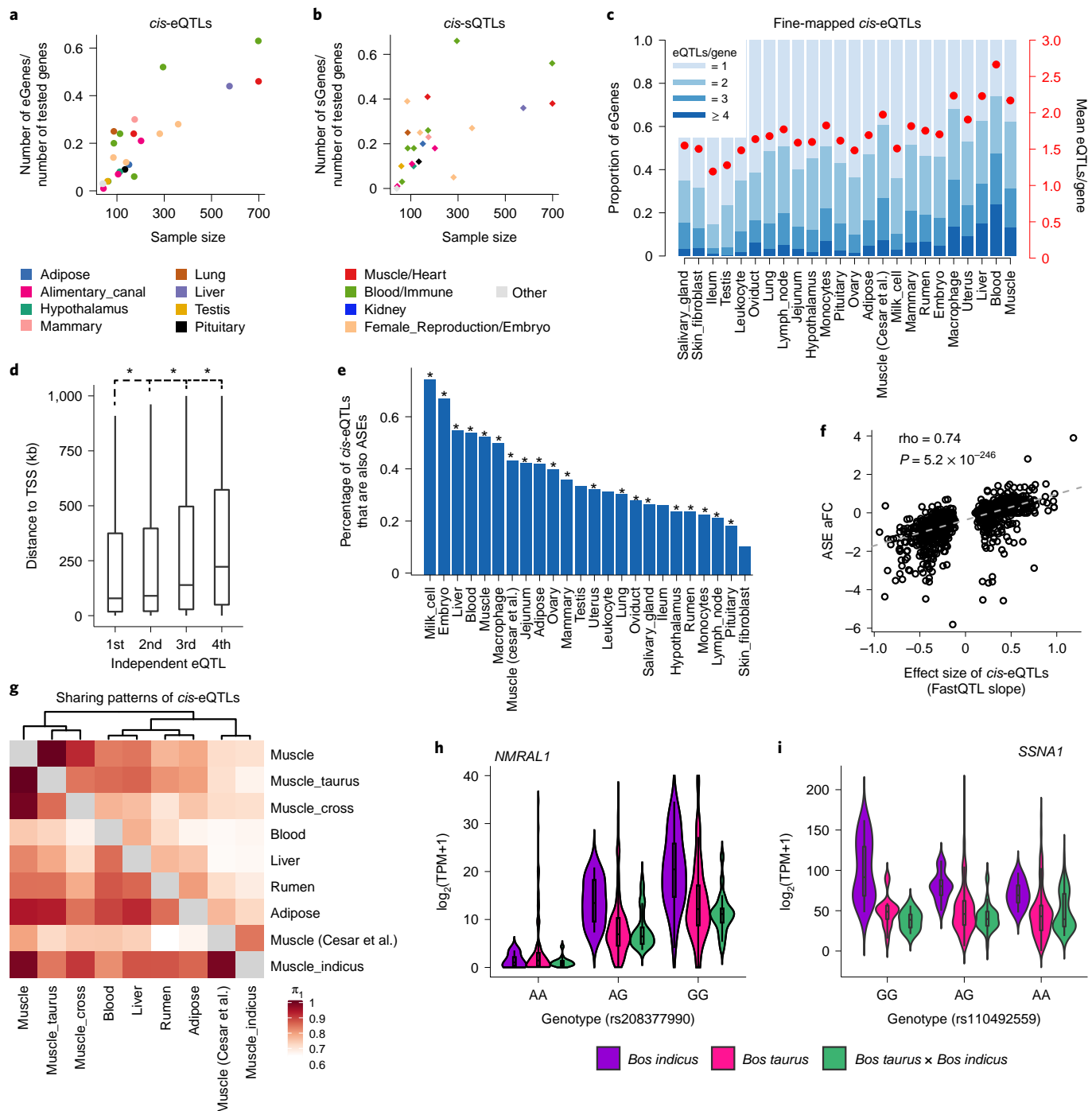


Fig. 3 | Discovery and characterization of *cis*-eQTLs and *cis*-sQTLs. **a**, Relationship between the percentages of eGenes over all tested genes and sample sizes (Pearson $r = 0.85$; two-sided Student's t -test: $P = 1.30 \times 10^{-7}$) across 23 distinct tissues. **b**, Relationship between the percentage of sGenes over all tested genes and sample sizes (Pearson $r = 0.63$; two-sided Student's t -test: $P = 1.06 \times 10^{-3}$) across 23 distinct tissues. Tissues are colored according to their tissue categories. **c**, Distribution and average number of conditionally independent *cis*-eQTLs per gene across tissues. Tissues are ordered by sample size. **d**, Distance to the TSS increases from the first to the fourth independent *cis*-eQTL. Only 7,276 gene-tissue pairs with at least four independent *cis*-eQTLs were chosen. Significant differences (denoted by an asterisk) were observed between the first versus second ($P = 2.4 \times 10^{-3}$), second versus third ($P = 3.0 \times 10^{-26}$) and third versus fourth ($P = 1.9 \times 10^{-27}$) independent *cis*-eQTLs based on the two-sided paired sample t -test. **e**, *cis*-eQTLs are significantly ($P < 1.0 \times 10^{-14}$, denoted by an asterisk, Fisher's exact test) overrepresented in the loci with ASE. The y axis indicates the percentage of *cis*-eQTLs that are also ASEs over all tested SNPs in the ASE analysis. **f**, Correlation of effect sizes (FastQTL slope) of *cis*-eQTLs and aFC of ASEs (Spearman's $\rho = 0.74$, two-sided Student's t -test: $P = 5.2 \times 10^{-246}$) in liver. **g**, Pairwise *cis*-eQTL sharing patterns (π_1 value) of muscle tissue across three breed groups (*B. indicus*, *B. taurus* and their crosses) and other tissues. Rows are discovery population and columns are validation population. "Muscle (Cesar et al.)" is for 160 skeletal muscle samples of *B. indicus* obtained from Cesar et al.⁸ **h**, A *cis*-eQTL (rs208377990) of *NMRAL1* in muscle is shared across *B. indicus* ($n = 51$), *B. taurus* ($n = 505$) and their crosses ($n = 108$). **i**, A *cis*-eQTL (rs110492559) of *SSNA1* in muscle is specific in *B. indicus* (MAF = 0.25 and 0.37 in *B. taurus* and *B. indicus*, respectively), and has a significant (two-sided t -test, $P = 5.61 \times 10^{-3}$) genotype × breed interaction. The samples are the same as in **h**.

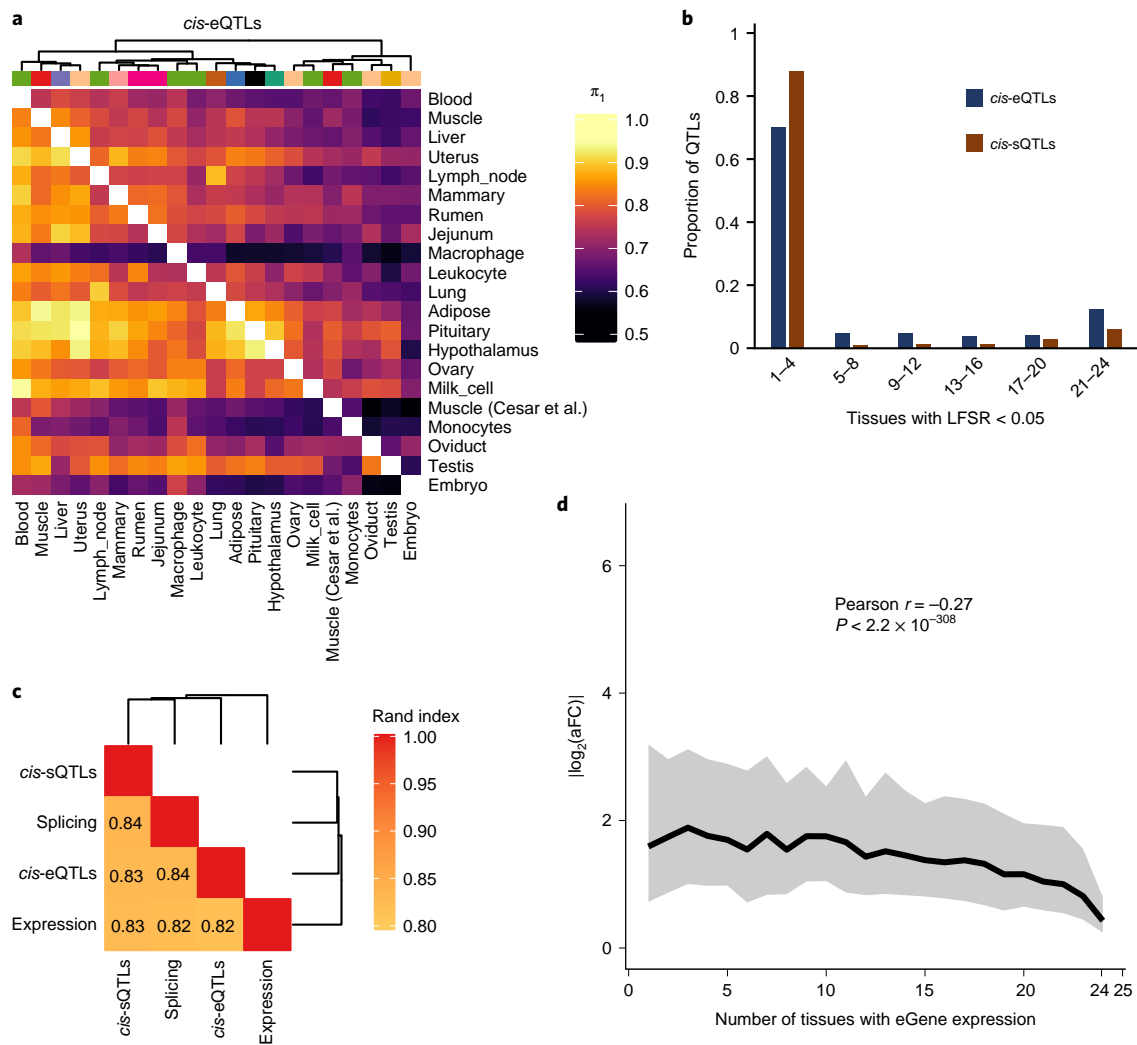


Fig. 4 | Tissue-sharing patterns of *cis*-QTLs. **a**, Pairwise *cis*-eQTL sharing patterns (π_1 value) across 23 distinct tissues. **b**, Tissue activity of *cis*-eQTLs and *cis*-sQTLs, where a *cis*-QTL is considered active in a tissue if it has a *mashr* local false sign rate (LFSR, equivalent to FDR) of < 5%. **c**, The similarity of tissue clustering across four data types (*cis*-eQTL, *cis*-sQTL, gene expression and splicing) based on the π_1 values³¹³. The *k*-means clustering is performed based on 2–22 clusters with 100,000 iterations. For each pairwise data type, we report the median pairwise Rand index across all clusters. **d**, Median (line) and 95% confidence interval (shading) of *cis*-eQTL effect size (y axis, measured as the absolute \log_2 transformed allelic fold change ($|\log_2(aFC)|$)), as a function of the number of tissues in which the eGene is expressed (x axis; TPM > 0.1). Pearson correlation between $|\log_2(aFC)|$ and the number of tissues with eGene expression is -0.27 (two-sided Student's *t*-test: d.f. = 43,721; $P < 2.2 \times 10^{-308}$).

that the difference in *cis*-eQTLs between breeds could be due in part to differences in the frequency of segregating variants, provided that there are no epistatic/environmental/developmental effects.

The tissue-sharing patterns of *cis*-QTLs could provide novel insights into molecular regulatory mechanisms underlying complex phenotypes³. We applied the π_1 statistics to measure the sharing patterns of *cis*-eQTLs between tissues (Fig. 4a and Extended Data Fig. 7c). In general, we observed that both *cis*-eQTLs and *cis*-sQTLs tended to be tissue-specific or ubiquitous across tissues (Fig. 4b). We also calculated the tissue-sharing patterns of gene expression and alternative splicing (Extended Data Fig. 7d,e), and found that the tissue-sharing patterns of the four core data types (gene expression, alternative splicing and *cis*-eQTLs) were similar (Fig. 4c and Extended Data Fig. 7f). This suggests that tissues with similar transcriptional profiles shared the genetic regulatory mechanisms of transcription. Further analysis on the expression of eGenes across tissues revealed that effect sizes of eVariants decreased with the increasing number of tissues in which target eGenes were

expressed, indicating that, on average, tissue-specific genes might be regulated by SNPs with larger genetic regulatory effects than widely expressed genes (Fig. 4d). Because of the limitations and challenges of *trans*-eQTLs analysis in this study, which include insufficient statistical power, the relatively lower imputation accuracy of distant intergenic SNPs and complex inter-chromosomal LD in cattle (which could lead to increased type I error rates)¹⁴, we only conducted exploratory *trans*-eQTL mapping for 15 tissues with over 100 individuals. We detected an average of 1,058 and 84 *trans*-eGenes and *trans*-sGenes (FDR < 0.05) across tissues, respectively (Supplementary Table 9). Details of *trans*-eQTL mapping, including LD patterns of *trans*-eQTLs and *cis*-eQTLs, tissue-sharing patterns of *trans*-eQTLs and their validations, are summarized in Extended Data Fig. 8.

Functional annotation of QTLs. We employed multiple layers of biological data to better define the molecular mechanisms of genetic regulatory effects. As expected, *cis*-eQTLs were significantly

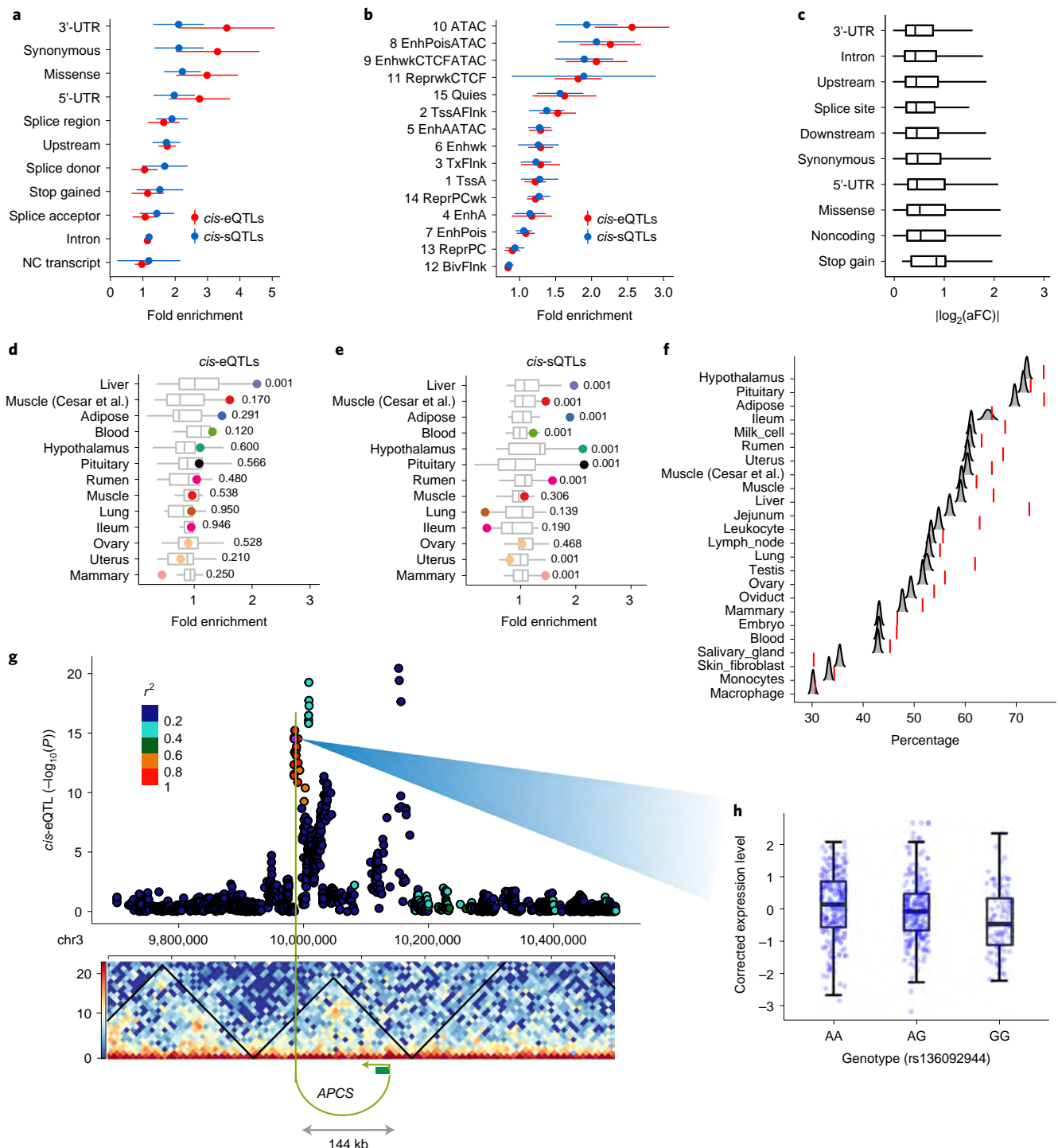


Fig. 5 | Functional annotation of *cis*-QTLs. **a**, Enrichment (fold change, two-sided 1,000 times permutation test) of *cis*-eQTLs and *cis*-sQTLs of 23 distinct tissues in sequence ontology. Data are presented as mean \pm s.d. NC, non-coding; UTR, untranslated region. **b**, Enrichment (fold change, two-sided 1,000 times permutation test) of *cis*-eQTLs and *cis*-sQTLs of 23 distinct tissues in 15 chromatin states predicted from cattle rumen epithelial primary cells in Holstein animals¹⁵. Data are presented as mean \pm s.d. **c**, Effect sizes (measured as $|\log_2(\text{aFC})|$) of *cis*-eQTLs of 23 distinct tissues across sequence ontology. **d, e**, Enrichment of *cis*-eQTLs (**d**) and *cis*-sQTLs (**e**) of 13 tissues in tissue-specific hypomethylated regions. These 13 tissues have both DNA methylation and *cis*-QTL data. The numbers are *P* values for enrichments of matched tissues (highlighted dots) based on the permutation test (two-sided, 1,000 times). **f**, Percentages of eGene-eVariant pairs that are located within TADs are significantly (*FDR* < 0.01, one-sided) higher than those of random eGene-SNP pairs with matched distances, except for ileum, macrophage and skin fibroblast. The null distributions of percentages of eGene-SNP pairs within TADs are obtained by performing 5,000 bootstraps. TADs are obtained from the lung Hi-C data. **g**, An eGene (*APCS*) and its eVariant (rs136092944) are located within a TAD, and linked by a significant Hi-C contact (10 kb bins, position 9,985,000 is linked to 10,135,000 in chromosome 3 with Benjamini-Hochberg corrected *P* = 1.4×10^{-6}). The *P* value is obtained based on the binominal distribution model. The Manhattan plot shows the *P* values of all tested SNPs in the *cis*-eQTL mapping analysis of *APCS*. The LD (r^2) values between eVariant (rs136092944) and surrounding SNPs are shown in colors. **h**, The boxplot shows the PEER-corrected expression levels of *APCS* across the three genotypes of eVariant (rs136092944): AA (*n* = 237), AG (*n* = 245) and GG (*n* = 94).

($P < 0.05$, 1,000 times permutation test) enriched in functional elements, such as the 3' untranslated region and open chromatin regions, by assay for transposase-accessible chromatin using sequencing data in cattle rumen epithelial primary cells¹⁵ (Fig. 5a,b). The *cis*-sQTLs had a higher enrichment in splice donors/acceptors than *cis*-eQTLs. The *cis*-eQTLs associated with stop gains had larger effect sizes than other *cis*-eQTLs (Fig. 5c). The *cis*-e/sQTLs were enriched in hypomethylated regions of the matching tissues across 13 tissues (Fig. 5d,e). For instance, the liver exhibited the highest enrichment of *cis*-e/sQTL in liver-specific hypomethylated regions. Consistent with the brain having a distinct abundance of alternative splicing, related to the development of the nervous system¹³, *cis*-sQTLs in the hypothalamus and pituitary had the highest enrichments in their specific hypomethylated regions (Fig. 5e).

Topologically associated domains (TADs) enable chromatin interactions between distant regulatory regions and target promoters¹⁶. By examining Hi-C data of lung tissue in cattle¹⁷, we obtained TADs and significant Hi-C contacts, which were likely to be conserved across tissues¹⁶. By comparing with random eGene-SNP pairs with matched distances, we observed significantly (FDR < 0.01 , 5,000 bootstrapping test) higher percentages of eGene-eVariant pairs within TADs across the majority of tissues, except for ileum and skin fibroblast (Fig. 5f). For instance, *APCS* and its *cis*-eQTL peak (144 kb upstream of its TSS) were encompassed by a TAD and linked by a significant Hi-C contact, which allowed regulation of its expression by a distant eVariant (rs136092944) (Fig. 5g,h).

***cis*-QTLs and complex trait associations.** The primary goal of this study is to provide a resource for elucidating the genetic and biological mechanisms involved in cattle complex traits. We thus evaluated *cis*-e/sQTLs detected in each tissue for associations with four distinct agronomic traits as examples: ketosis, milk yield, age at first calving and somatic cell score. The top SNPs associated with ketosis from GWAS were significantly ($P < 0.05$, the 1,000 times permutation test) enriched for liver *cis*-e/sQTLs (Fig. 6a). Similarly, SNPs associated with milk yield were significantly overrepresented in *cis*-e/sQTLs from mammary gland (Fig. 6b). Compared with other tissues, mammary gland, milk cells and liver were the tissues with the highest enrichment of milk-yield associated SNPs among *cis*-eQTLs (Fig. 6c). In addition, SNPs associated with age at first calving were significantly enriched for monocyte *cis*-eQTLs, and somatic cell score for mammary gland (Extended Data Fig. 9a). We observed that a larger sample size of *cis*-eQTL tissue resulted in a higher enrichment of GWAS loci and *cis*-eQTLs, potentially explaining the associations of complex traits with nonmatching tissues (Extended Data Fig. 9b).

We detected 854 significant gene-trait pairs for 43 agronomic traits (Supplementary Table 10) in cattle *via* single-tissue TWAS (S-PrediXcan), representing 337 unique genes (Supplementary Table 11). Of 319 fine-mapped genes^{18,19}, we validated 54, including linking expression of *DGAT1* in liver and mammary gland, and expression of *MGST1* in milk cells, as well as expression of *CLN3* in

liver to milk yield (Fig. 6d). Expression of *ZNF613* in the hypothalamus was the most significant association for many reproduction and body conformation traits, including daughter-still-birth and stature (Supplementary Table 11), supporting our previous finding that *ZNF613* is significantly associated with gestation length, possibly through its influence on embryonic development²⁰. Furthermore, we conducted a colocalization analysis of *cis*-eQTLs and GWAS loci using fastENLOC, and detected 115 unique eGenes that were colocalized (regional colocalization probability, $rcp > 0.5$) with 260 GWAS loci associated with 25 of the 43 complex traits analyzed. These represented 235 significant gene-trait pairs (Fig. 6e and Supplementary Table 12). For instance, *TIGAR*, a muscle *cis*-eGene, with roles in phosphatase activity, energy storage and consumption, was colocalized ($rcp = 0.529$) with one of the independent GWAS signals of strength on chromosome 5 (Extended Data Fig. 9c). GWAS loci of milk yield were colocalized with *ARHGAP39* in hypothalamus, *TEF* in embryo, *SYT11* in blood, *CCDC166* in oviduct and *ASPHD1* in jejunum (Supplementary Table 12). We also took sire calving ease, for which GWAS loci were colocalized with 21 eGenes in at least one tissue, as an example in Extended Data Fig. 9d. In addition, we further employed Coloc and S-MultiXcan to conduct colocalization and multi-tissue TWAS analysis, and detected 110 and 590 significant gene-trait pairs, respectively (Supplementary Tables 13 and 14). By comparing results from TWAS and colocalization, we found an overlap of seven gene-trait pairs (Fig. 6f and Extended Data Fig. 10). For instance, we found that *cis*-eQTLs of *DGAT1* in liver were colocalized ($rcp = 0.78$) with GWAS signals of protein yield, and the P values from GWAS were highly ($r = 0.91$) correlated with those from *cis*-eQTL (Fig. 6g,h).

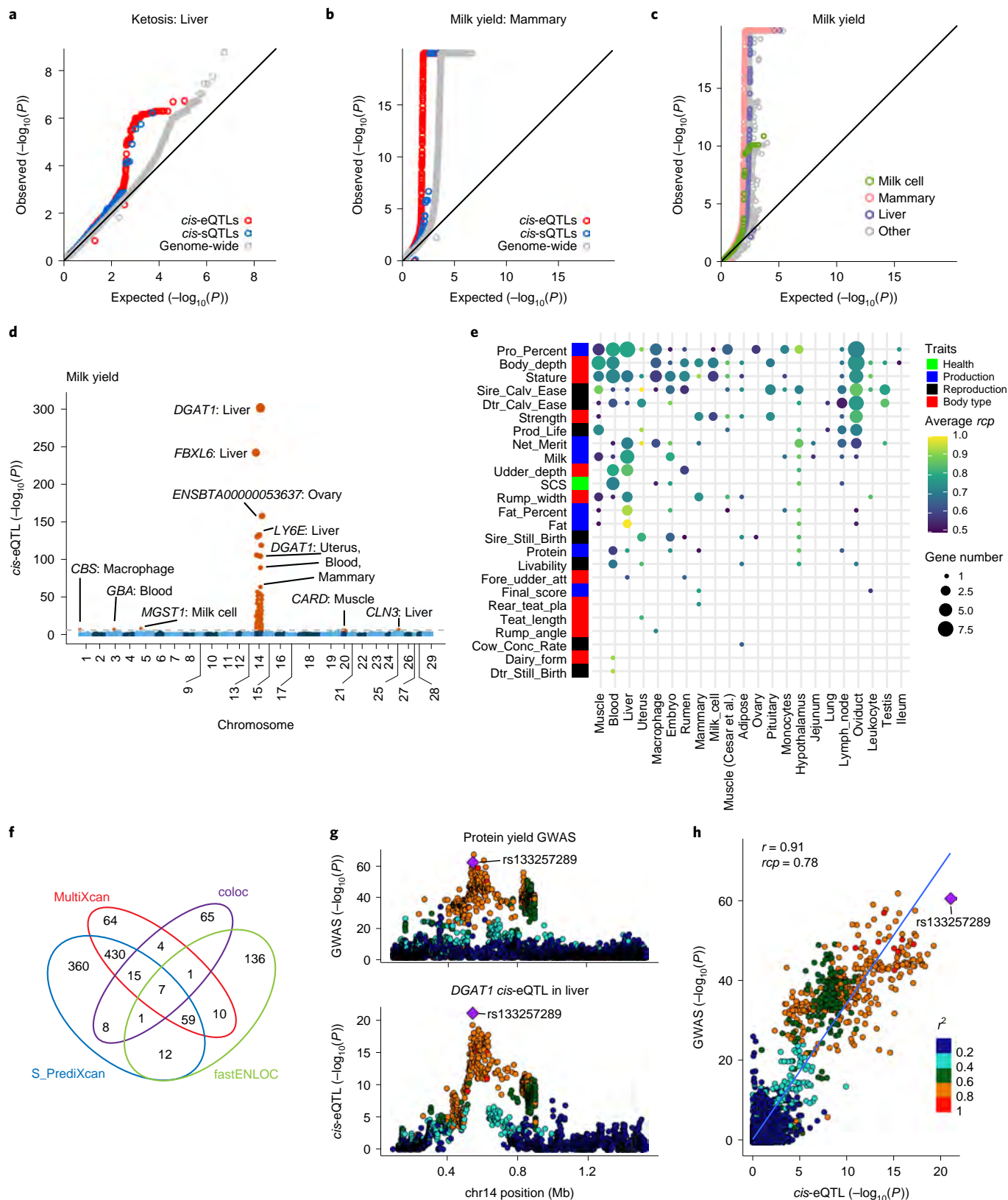
Discussion

The CattleGTEx atlas represents one of the most comprehensive reference resources of the cattle transcriptome to date. It provides a detailed characterization of genetic control of gene expression and splicing across 23 distinct tissues in cattle. This study demonstrates that it is possible to discover gene expression regulatory variants by deriving and imputing genetic variants from livestock RNA-seq data alone. We established an *in silico* protocol to generate a FarmGTEx atlas in a timely manner and show the value of reanalyzing published data to find novel biology, avoiding the notable costs of data generation in livestock. Although we have provided a comprehensive view of the genetic regulatory variants in cattle, we are also mindful that this resource can be further improved with the inclusion of more individuals/breeds and further data types. The imputation accuracy for breeds that are very underrepresented in the reference panel might be relatively low. In addition, generating SNP genotypes or WGS for individuals with RNA-seq data can provide additional information for distal intergenic variants compared with RNA-seq data alone. The FarmGTEx consortium is currently extending the bioinformatics pipeline developed here to other livestock species (for example, pig, sheep, goat and chicken).

Fig. 6 | Relationship between complex traits and *cis*-QTLs. **a**, *cis*-eQTLs ($P = 0.001$) and *cis*-sQTLs ($P = 0.02$) in liver show significantly higher enrichment for top 10% of SNPs associated with ketosis compared with genome-wide SNPs (shown in gray). **b**, *cis*-eQTLs ($P = 0.001$) and *cis*-sQTLs ($P = 0.03$) in mammary gland show higher enrichment for top SNPs associated with milk yield compared with genome-wide SNPs (shown in gray). All the P values above are obtained by the two-sided 1,000 times permutation test. **c**, Enrichment of *cis*-eQTLs for genetic associations with milk yield is tissue-dependent. *cis*-eQTLs in mammary gland, milk cells and liver exhibit higher enrichment for genetic associations with milk yield compared with those in other tissues. **d**, Manhattan plots of TWAS for milk yield across all 23 distinct tissues. **e**, Number of genes that were colocalized ($rcp > 0.5$ in fastENLOC) between GWAS significant loci of complex traits and *cis*-eQTLs across tissues. Point size indicates the number of genes, whereas point color indicates the average rcp of each trait-tissue pair. The abbreviations for GWAS traits are explained in Supplementary Table 10. **f**, Overlap of significant gene-trait pairs from TWAS with S-PrediXcan (Bonferroni corrected $P < 0.05$) and S-MultiXcan (Bonferroni corrected $P < 0.05$) and colocalization with fastENLOC ($rcp > 0.5$) and Coloc (PP, $H4 > 0.8$). **g**, Example of colocalization ($rcp = 0.78$) of *cis*-eQTLs of *DGAT1* gene in liver and GWAS loci of protein yield in cattle on chromosome 14. The top colocalized SNP (rs133257289) is the top *cis*-eQTL of *DGAT1* and the second top GWAS signal of protein yield. **h**, Pearson correlation ($r = 0.91$, two-sided Student's t -test: d.f. = 2,933; $P < 2.2 \times 10^{-308}$) between P values from *cis*-eQTLs of *DGAT1* in liver and GWAS of protein yield.

The CattleGTEx also provides a resource to explore tissue-sharing patterns of the transcriptome and its genetic regulation in cattle. By contrast to the human GTEx³, in which RNA-seq samples across tissues were collected from the same individuals, the CattleGTEx used public data, in which individuals or even breeds differed from tissue to tissue. This might explain why there

is a lower proportion of shared *cis*-eQTLs across tissues compared with the human GTEx. In addition, the difference in the cell-type composition of tissues can also affect the tissue-sharing patterns of *cis*-QTLs³. When single-cell RNA-seq data are available for multiple tissues in farm animals in the near future²¹, it will be of interest to computationally estimate the cell-type proportions



in the bulk-tissue samples to uncover the cellular specificity of genetic regulatory effects²².

This CattleGTEx atlas provides an important tool for studying the mechanisms underlying complex traits by systematically linking SNPs, genes, tissues and complex traits. The *cis-e*/sQTLs detected here provide a rich set of functional variants for agronomic traits in cattle, because we found that top GWAS associations of traits were significantly enriched for regulatory QTLs in their relevant tissues. Our TWAS and colocalization analyses further provide a list of promising candidate genes/variants for functional follow-ups. We noted the relatively small overlap of results from TWAS and colocalization. This might be because these methods assume the genetic architecture of both the trait of interest and tissue gene expression differently. In addition, we observed a discrepancy between high *rcp* values and the lack of correlation of raw *P* values for GWAS and eQTL in the entire region of each colocalized locus. This may be because of: (1) allelic heterogeneity and complex LD in each locus; (2) an imperfect LD match between GWAS (only Holstein population) and *cis-e*QTLs populations (multiple breeds); and (3) current commonly used colocalization methods based on GWAS summary statistics might not work well in highly related individuals in livestock. We therefore suggest focusing analyses on loci where colocalization and TWAS methods agree.

Further integration of these regulatory QTLs with functional annotations from the Functional Annotation of Animal Genomes (FAANG) project will provide opportunities to understand transcriptional/post-transcriptional regulatory mechanisms underpinning GWAS hits for agronomic traits²³. The multi-tissue *cis-e*/sQTLs generated here will also enable the exploration of molecular mechanisms underlying the extensive pleiotropic effects identified in livestock²⁴. This information will allow an understanding of the response mechanisms to intended selection as well as disentangling negative correlated responses to this same selection (for example, increasing mastitis or deteriorating fertility when selecting for increased milk production). Furthermore, this resource will assist in the development of genomic selection methods and tools to improve animal health and wellbeing. For instance, a better understanding of the genetic architecture underpinning agronomic traits will benefit genetic improvement programs by incorporating biological knowledge into genomic prediction models, which has been shown to improve prediction accuracy across populations and breeds^{10,24}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-022-01153-5>.

Received: 23 November 2020; Accepted: 7 July 2022;

Published online: 11 August 2022

References

- Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
- Hu, Z. L., Park, C. A. & Reecy, J. M. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res.* **47**, D701–D710 (2019).
- GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Fang, L. et al. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res.* **30**, 790–801 (2020).
- Xiang, R. et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proc. Natl Acad. Sci. USA* **116**, 19398–19408 (2019).
- Prowse-Wilkins, C. P. et al. Putative causal variants are enriched in annotated functional regions from six bovine tissues. *Front. Genet.* **12**, 664379 (2021).
- Xiang, R. et al. Genome variants associated with RNA splicing variations in bovine are extensively shared between tissues. *BMC Genomics* **19**, 521 (2018).
- Cesar, A. S. M. et al. Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. *BMC Genomics* **19**, 499 (2018).
- Littlejohn, M. D. et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. *Sci. Rep.* **6**, 25376 (2016).
- Hayes, B. J. & Daetwyler, H. D. 1000 Bull Genomes Project to map simple and complex genetic traits in cattle: applications and outcomes. *Annu. Rev. Anim. Biosci.* **7**, 89–102 (2019).
- Deelen, P. et al. Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7**, 30 (2015).
- Hombach, S. & Kretz, M. Non-coding RNAs: classification, biology and functioning. *Adv. Exp. Med. Biol.* **937**, 3–17 (2016).
- GTEx Consortium et al. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
- Tenesa, A. et al. Estimation of linkage disequilibrium in a sample of the United Kingdom dairy cattle population using unphased genotypes. *J. Anim. Sci.* **81**, 617–623 (2003).
- Fang, L. et al. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol.* **17**, 68 (2019).
- Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
- Rosen, B. D. et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**, gaa021 (2020).
- Jiang, J. et al. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun. Biol.* **2**, 212 (2019).
- Freebern, E. et al. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. *BMC Genomics* **21**, 41 (2020).
- Fang, L. et al. Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. *Commun. Biol.* **2**, 100 (2019).
- Gao, Y. et al. Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning. *Genomics* **113**, 2045–2055 (2021).
- Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
- Clark, E. L. et al. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. *Genome Biol.* **21**, 285 (2020).
- Xiang, R. D. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. *Nat. Commun.* **12**, 860 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Methods

Quantification of gene expression. We downloaded 11,642 RNA-seq datasets (by 24 June 2019) from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA; $n = 11,513$, <https://www.ncbi.nlm.nih.gov/sra/>) and BIGD databases ($n = 129$, <https://bigd.big.ac.cn/bioproject/>) by searching the 'Organism' for 'Cattle' and the 'Strategy' for 'RNA seq'. We merged multiple datasets from single samples, yielding 8,536 unique RNA-seq samples. We applied a stringent and uniform pipeline to filter and analyze all the data. Briefly, we first removed adapters and low-quality reads using Trimmomatic (v.0.39)²⁵ with parameters: adapters/TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. We filtered out samples with clean read counts $\leq 500K$, resulting in 7,680 samples, and mapped clean reads to the ARS-UCD1.2 cattle reference genome¹⁷ using single or paired mapping modules of STAR (v.2.7.0) with parameters of outFilterMismatchNmax 3, outFilterMultimapNmax 10 and outFilterScoreMinOverRead 0.66. We kept 7,264 samples with uniquely mapping rates $\geq 60\%$ (mean, 91.07%; range, 60.44–100%; mapping details in Supplementary Table 1). We then obtained normalized expression (TPM) of 27,608 Ensembl (v.96) annotated genes using Stringtie (v.2.1.1)³⁶ and extracted raw read counts of them with featureCounts (v.1.5.2)³⁷. We finally clustered 7,264 samples based on $\log_2(\text{TPM} + 1)$ using a hierarchical clustering method, implemented in R (v.3.4.1) package *dendextend*, with distance = $(1 - r)$, where r is the Pearson correlation coefficient. We excluded samples with obvious clustering errors (e.g., samples labeled as liver that were not clustered with other liver samples), resulting in 7,180 samples for subsequent analysis.

Quantification of alternative splicing. We used LeafCutter (v.0.2.9)²⁸ to identify and quantify variable alternative splicing events of genes by leveraging information of junction reads (that is, reads spanning introns) that were obtained from the STAR alignment. LeafCutter enables the identification of splicing events without relying on existing annotations that are typically incomplete, especially in the setting of large genes or individual- and/or population-specific isoforms²⁸. We first converted bam files from STAR alignment into junction files using the script 'bam2junc.sh', and then performed intron clustering using the script 'leafcutter_cluster.py' with default settings of 50 reads per cluster and a maximum intron length of 500 kb. We employed the 'prepare_genotype_table.py' script in LeafCutter to calculate intron excision ratios and to remove introns used in fewer than 40% of individuals or with no variation. Ultimately, we standardized and quantile normalized intron excision ratios as percent spliced-in (PSI) values across samples. We clustered 7,180 samples based on PSI using the same method as used in gene expression.

Genotyping and imputation. We called genotypes of known genomic variants in the 1000 Bull Genomes Projects¹⁰ for 7,180 high-quality RNA-seq samples individually, following the recommended best practices pipeline in Genome Analysis Toolkit (v.4.0.8.1)²⁹ with default settings. We filtered out low-quality SNPs using --filter-expression 'FS > 30.0 & QD < 2.0'. We then imputed the filtered SNPs on autosomes to sequence level using Beagle (v.5.1)³⁰ based on a multiple-breed reference population that consisted of 3,103 individuals from run7 of the 1000 Bull Genomes Project¹⁰ and 207 public individuals from *B. taurus* ($n = 101$), *B. indicus* (zebu, $n = 20$) and *Bos grunniens* (yak, $n = 86$) (Supplementary Table 15). Finally, we obtained 6,123 samples that were genotyped and imputed successfully. We filtered out variants with MAF < 0.05 and DR² < 0.8, resulting in 3,824,444 SNPs used for QTL mapping. To evaluate the accuracy of imputation, we called genotypes (~6 million SNPs) from WGS (average read depth $\times 10\times$) of Holstein ($n = 4$), Limousin ($n = 3$) and Angus ($n = 5$) animals, which had RNA-seq data as well. We then measured the genotype concordance rates between WGS-SNPs and RNA-seq/imputed SNPs. We extracted 153,913 LD-independent SNPs using plink (v.1.90)³¹ (--indep-pairwise 1000 5 0.2), and conducted PCA for all 6,123 samples using these SNPs in EIGENSOFT (v.7.2.1)³². We calculated the identity-by-state (IBS) distance among samples by using these independent SNPs to remove duplicate individuals. IBS distance = $(\text{IBS2} + 0.5 \times \text{IBS1}) / (\text{IBS0} + \text{IBS1} + \text{IBS2})$, where IBS0 is the number of IBS 0 nonmissing variants, IBS1 is the number of IBS 1 nonmissing variants and IBS2 is the number of IBS 2 nonmissing variants. We set an IBS distance cutoff of 0.85 to deem two samples as duplicates and kept one of them. When conducting QTL mapping, we removed an average of 43 duplicate samples within each tested tissue (ranging from one in salivary gland and leukocyte to 132 in muscle), resulting in 4,889 samples.

Allele-specific expression. We conducted ASE analysis using the Genome Analysis Toolkit ASEReadCounter tool (v.4.0.8.1) with the following settings: --U ALLOW_N_CIGAR_READS --minDepth 10 --minMappingQuality 255 --minBaseQuality 10. SNPs for ASE detection fulfilled the following criteria: heterozygous in at least five samples, at least ten reads per allele and at least 2% of all reads supporting the minor allele. We then calculated a binomial P value by comparison with the expected ratio under the null hypothesis, followed by multiple-test correction with the Benjamini–Hochberg approach (FDR). SNPs with FDR < 0.05 were considered as significant ASE. We estimated the effect size (allelic fold change, aFC) of regulatory variants at ASE loci using a haplotype-based approach implemented in phASER (v.1.1.1)³³.

Bioinformatics analysis of WGBS data. For WGBS data analysis, we first used FastQC (v.0.11.2) to determine read quality and Trim Galore v.0.4.0 (--max_n

15 --quality 20 --length 20 -e 0.1) to filter reads of low quality. We then mapped clean reads to the same reference genome (ARS-UCD1.2) using Bismark software (v.0.14.5)³⁴ with default parameters. After deduplication of reads, we extracted methylation levels of cytosines using the *bismark_methylation_extractor* (--ignore_r2 6) function. The coverages of all WGBS data were calculated using clean reads with an average of 27.6-fold coverage (range: 5–47 \times). Ultimately, we kept CpG sites that were represented by at least five reads for subsequent analyses. We visualized sample clusters based on DNA methylation levels of shared CpGs using t -distributed stochastic neighbor embedding approaches.

Identification of TAD and significant Hi-C contacts. To find potential chromatin interactions between distant eVariants and target eGenes, we identified TADs and Hi-C contacts from Hi-C data from lung tissue in cattle that were retrieved from the NCBI SRA under accessions SRR5753600, SRR5753603 and SRR5753606. We used Trim Galore (v.0.4.0) to trim adapter sequences and low-quality reads (--max_n 15 --quality 20 --length 20 -e 0.1), resulting in ~820 million clean reads. We then mapped clean reads to the reference genome (ARS-UCD1.2) using BWA (v.0.7.17)³⁵. We applied HiCExplorer v.3.4.1³⁶ to build a Hi-C contact matrix with 10 kb resolution and identified TAD with hicFindTAD. We kept TADs with FDR < 0.01 to link eQTLs to eGenes. We further employed HiC-Pro (v.2.11.4)³⁷ to call Hi-C contacts with 10 kb resolution from Hi-C data. Briefly, HiC-Pro aligned clean reads to the reference genome with Bowtie2 (v.2.3.5)^{35,38}. After building a contact matrix, HiC-Pro generated intra- and inter-chromosomal maps and normalized them using the iterative correction and eigenvector decomposition normalization algorithm. We converted Hi-C contact matrix in HiC-Pro format to FitHiC format using HiCPro2FitHiC.py in FitHiC (v.2.0.7) and applied statistical confidence estimates to determine the significant intra-chromosome contacts (Benjamini–Hochberg corrected $P < 0.05$).

Tissue-specificity analysis of gene expression, alternative splicing and DNA methylation.

To quantify tissue-specific expression of genes, we computed t -statistics for each gene in each of the 114 tissues. We grouped 114 tissues into 13 categories (Supplementary Table 1). We scaled the \log_2 (transformed expression) ($\log_2(\text{TPM})$) of genes to have a mean of zero and variance of one within each tissue. We then fitted a linear model as described previously¹⁵ for each gene in each tissue using the least squares method. When constructing the matrix of dummy variables (design matrix) for tissues, we denoted samples of the target tissue/cell type (for example, CD4 cells) as '1', and samples outside the target category (for example, nonblood/immune tissues) as '–1'. We excluded samples within the same category (for example, CD8 cells and lymphocytes) to detect genes with specific expression in each particular category, even if they were not specific to the target tissue within this category. We obtained t -statistics for each gene to measure its expression specificity in a given tissue. We considered the top 5% of genes ranked by the largest t -statistics as genes with high tissue-specific expression. To explore the conservation of tissue-specific expression between cattle and humans, we employed the same method to quantify the tissue-specific expression of all orthologous genes in each of 55 human tissues using GTEx (v.8) data³.

To detect tissue-specific alternative splicing, we used LeafCutter to analyze the differential intron excision by comparing the samples from the target tissue with the remaining tissues²⁸, while excluding samples from tissues of the same category as the target tissue. We used the Benjamini–Hochberg method (FDR) to control multiple testing.

For DNA methylation, we focused on gene promoters (from upstream 1,500 bp to downstream 500 bp of TSS based on the ARS-UCD1.2 from Ensembl v.99), the methylation levels of which were calculated with a weighted methylation method using the roimethstat function in MethPipe (v.3.4.3)³⁹. We computed a t -statistic for the promoter of each gene using the same method as in tissue-specific expression analysis. We considered the bottom 5% of genes ranked by t -statistics as genes with tissue-specific promoter hypomethylation. We also detected tissue-specific methylation regions in a genome-wide mode using SMART2 (v.2.2.8)⁴⁰ with parameters of -t DeNovoDMR -MR 0.5 -AG 1.0 -MS 0.5 -ED 0.2 -SM 0.6 -CD 500 -CN 5 -SL 20 -PD 0.05 -PM 0.05.

Covariate analysis for QTL discovery. To account for hidden batch effects and other technical/biological sources of transcriptome-wide variation in gene expression, we estimated latent covariates in each tissue using the Probabilistic Estimation of Expression Residuals (PEER v.1.3) method⁴¹. In each tissue, we estimated 75 PEER factors first. The posterior variances of factor weights dramatically decreased and reached or nearly reached plateaus when ten PEER factors were included (Extended Data Fig. 6a). Therefore, we used ten PEER covariates to account for the effects of confounding variables on gene expression in all following QTL analyses. For instance, the variance of gene expression among samples in adipose captured by nine of ten PEER factors were significantly (FDR < 0.05) correlated with known technical and biological covariates like clean data size, mapping rate, project, breeds, subspecies, sex and age (Extended Data Fig. 6b). To further control the effect of population structure on the discovery of QTLs, we included genotype principal components (PCs) based on sample size bins: three PCs for tissues with <150 samples, five PCs for tissues with ≥ 150 and <250 samples, and ten PCs for tissues with ≥ 250 samples.

cis-eQTL mapping. We conducted *cis*-eQTL mapping for 23 distinct tissues with at least 40 individuals each, while adjusting for corresponding PEER factors and genotype PCs. Detailed information about these 23 distinct tissues is given in Supplementary Table 4. Because the majority of *cis*-eQTLs are shared across subspecies/breeds (Fig. 3g), we combined, adjusting for species/breed, all of the datasets from the same tissue to perform *cis*-eQTL mapping in order to increase the statistical power. We kept genes with TPM > 0.1 in ≥20% samples in each tissue. Gene expression values of all samples in a given tissue were quantile normalized to the average empirical distribution and expression values for each gene then inverse normal transformed across samples. The *cis*-eQTL mapping was done using a linear regression model, implemented in FastQTL (v.2.184)⁴², to test associations of the normalized expression level of genes with genetic variants in 1 Mb of TSS of target genes. We only considered imputed variants with MAF > 0.05 and at least four minor alleles across samples within the target tissue. We first conducted *cis*-eQTL mapping in a permutation mode with the setting --permute 1000 10000, to identify genes with at least one significant *cis*-eQTL (eGene). We considered FDR ≤ 0.05 as significant, calculated using the Benjamini–Hochberg method based on the beta distribution-extrapolated empirical *P* values from FastQTL. To identify a list of significant eGene–eVariant pairs, we applied the nominal mode in FastQTL. A genome-wide empirical *P* value threshold p_i for each gene was defined as the empirical *P* value of the gene closest to the 0.05 FDR threshold³. We then calculated the nominal threshold as $F^{-1}(p_i)$ for each gene using the permutation mode of FastQTL (v.2.184), where F^{-1} is the binomial inverse cumulative distribution. We considered variants with nominal *P* values below the nominal threshold as significant, and included them in the list of eGene–eVariant pairs. We calculated the aFC, defined as the ratio of the expression level of the haplotype carrying the alternative allele over the one carrying the reference allele, to measure effect sizes of *cis*-eQTLs using the aFC (v.0.3) tools⁴³. We further applied the statistical fine-mapping method, dag-g (v.1.0.0)⁴⁴, to infer multiple independent casual *cis*-eQTLs of a gene in a tissue. The dag-g approach employed a Bayesian variable selection model, using a signal-level posterior inclusion probability to measure the strength of each association signal (SNPs in LD). We set a cutoff of 0.1 (signal-level posterior inclusion probability > 0.9) as the inclusion threshold to detect representative/independent eQTLs for the target eGene. To analyze pairwise tissue similarity in QTLs, we calculated π_i statistics, defined as the proportion of true positive QTLs identified in the first tissue (Discovery tissue) among all tested gene–variant pairs in second tissue (Validation tissue), using the Storey and Tibshirani *q*-value approach.¹³

Meta-analysis of *cis*-eQTLs of muscle samples from three subspecies. Data from muscle samples were available from three subspecies: *B. indicus* ($n = 51$), *B. taurus* ($n = 505$) and their crosses ($n = 108$). To explore the similarity and variability of *cis*-eQTLs among subspecies, we conducted *cis*-eQTL mapping using muscle samples from each of the subspecies separately. We then conducted a meta-analysis to integrate *cis*-eQTL results from three subspecies using the METAL (v.2020-05-05) tool⁴⁵. We obtained *Z*-scores (the sum of weighted effect sizes) of SNPs from the meta-analysis. Weights were proportional to the square-root of the number of individuals in each subspecies⁴⁵. We employed plink (v.1.90)³¹ to test the SNP × breed interaction in muscle samples, and adjusted the *P* values to FDR using Benjamini–Hochberg procedure. We took FDR < 0.05 as the significant threshold.

cis-sQTL mapping and tissue-sharing patterns. In each of the 23 distinct tissues, we applied a linear regression model, implemented in FastQTL⁴², to test for associations of genotypes within 1 Mb up- and downstream of target intron clusters and their corresponding intron excision ratios. We used the first five genotype PCs to account for the effect of ancestry, and ten PEER factors to adjust for the effect of unknown confounding variables. We applied the permutation pass mode (--permute 1000 10000) in FastQTL⁴² to obtain beta approximated permutation *P* values, followed by multiple-test correction with the FDR method. We considered sQTL–intron pairs with FDR < 0.05 as significant, and defined sGene as genes containing a significant sQTL in any introns. We employed MashR (v.0.2.57)⁴⁶ to analyze tissue-sharing patterns of QTLs³, and considered the local false sign rate < 0.05 as significant.

trans-QTL mapping. We conducted *trans*-eQTLs for 15 tissues with at least 100 samples each. We filtered genomic variants using a more stringent threshold than *cis*-eQTL mapping to partially account for the reduction in statistical power. We obtained mappability of variants based on *k*-mer lengths of 36 and 75 following the procedure described in https://wiki.bits.vib.be/index.php/Create_a_mappability_track. Briefly, we calculated the mappability of variants with 36 and 75 *k*-mer based on ARS-UCD1.2 using a fast mapping-based algorithm⁴⁷, allowing for two mismatches. For each gene, we averaged the mappability across exons with 72 *k*-mer length and untranslated regions with 36 *k*-mer length. We excluded any variants within repeats (RepeatMasker and simple repeats), and further removed variants with mappability < 1, based on a *k*-mer length of 75. After filtering, we kept SNPs with MAF > 0.05 and at least ten minor alleles within each tissue for association testing.

We used two methods to detect *trans*-eQTLs for protein-coding genes with an average mappability of ≥ 0.8. First, we associated the normalized expression of target genes with genotypes on other autosomal chromosomes using a linear regression model in MatrixQTL (v.2.3)⁴⁸, while adjusting for the same covariates as in *cis*-eQTL analysis. We further removed *trans*-eQTL–gene pairs that were cross-mappable

to reduce false positives⁴⁹. Second, we employed a linear mixed model (by fitting a polygenic effect with the genetic relationship matrix to further account for the complex relatedness among individuals) in GCTA (v.1.93.3beta)⁵⁰ for *trans*-eQTL and *trans*-sQTL mapping. For both methods, we adjusted *P* values for multiple testing using the Benjamini–Hochberg method to obtain FDR. We considered gene–variant pairs with FDR < 0.05 as significant. To conduct an internal validation of *trans*-eQTL mapping, we randomly and evenly divided blood and muscle samples into two groups. We conducted *trans*-eQTL mapping in the first group using the linear mixed model to detect significant *trans*-eQTL–gene pairs, and then repeated in the second group.

TWAS and colocalization of *cis*-eQTLs and GWAS loci. To associate gene expression in a tissue with complex traits, we conducted a single-tissue TWAS analysis using S-PrediXcan (v.0.6.1)⁵¹ by prioritizing GWAS summary statistics for 43 agronomic traits of economic importance in cattle (Supplementary Table 10), including reproduction ($n = 11$), production (milk-relevant; $n = 6$), body type ($n = 17$) and health (immune/metabolic-relevant; $n = 9$). For body conformation (type), reproduction and production traits, we conducted a single-marker GWAS by fitting a linear mixed model in 27,214 U.S. Holstein bulls¹⁸. For health traits, we conducted GWAS using the same method in a subset (ranging from 11,880 for hypocalcemia to 24,699 for livability) of the 27,214 available bulls¹⁹. We constructed a Nested Cross Validated Elastic Net prediction model using genotype and expression data. We included subspecies, ten PEER factors and corresponding genotype PCs in the model to adjust for unknown confounding variables and underlying population structure. For each trait, we conducted TWAS in each of the same 23 distinct tissues as in *cis*-eQTL mapping. We considered genes with Bonferroni corrected $P < 0.05$ as significant. We visualized the Manhattan plots of *P* values of all tested genes using ggplot2 (v.3.3.2) in R (v.3.4.1). In addition, we further employed S-MultiXcan (v.0.6.1)⁵² to conduct multi-tissue TWAS analysis, and considered gene–trait pairs with Bonferroni threshold $P < 4 \times 10^{-6}$ (0.05/13,024) significant.

To detect the shared causal variants of gene expression and complex traits, we conducted a colocalization analysis of *cis*-eQTLs from 23 distinct tissues and GWAS loci of 43 agronomic traits using fastENLOC (v.1.0)⁵³. Briefly, we split the imputed GWAS summary statistics into approximately LD-independent regions, and each region was considered as a GWAS locus. The LD-independent regions were generated from genotypes of 886 Holstein animals from run7 of the 1000 Bull Genomes Project because the GWAS summary statistics were from the U.S. Holstein population. In each GWAS locus of a trait with suggestive significant SNPs ($P < 10^{-5}$), we considered a gene with *rcp* > 0.5 as significant. We further conducted colocalization analysis using Coloc (v.5.1.0)⁵⁴ with the function coloc.abf. We obtained posterior probability values for the H_4 case (PPH4); that is, both traits (GWAS trait and eQTLs) are associated and share a single causal variant. We kept the tissue–trait–gene triples with PPH4 > 0.8 for downstream analysis.

Other downstream bioinformatics analysis. We used Genomic Association Tester (GAT v.1.3.4)⁵⁵ 1,000 permutations to estimate the functional enrichment of QTLs in particular genomic regions, for example, chromatin states and methylation elements. We considered enrichments with FDR < 0.05 as significant. We used the R package, ClusterProfiler (v.3.0.4)⁵⁶, to annotate the function of genes based on the Gene Ontology database from Bioconductor (org.Bt.eg.db v3.11.4). We considered Gene Ontology terms with FDR < 0.05 as significant.

Statistics and reproducibility. No statistical method was used to predetermine sample size. We used all data passing standard quality controls, resulting in 7,180 samples. For RNA-seq samples, we filtered out samples with clean read counts ≤ 500K or uniquely mapping rates < 60%, resulting in 7,180 samples. For genotypes, we filtered out SNPs with MAF < 0.05 or imputation $DR^2 < 0.8$, resulting in 3,824,444 SNPs used for QTL mapping. For the QTL mapping in each tissue, we set an IBS distance cutoff of 0.85 to deem two samples as duplicates and kept one of them for analysis. Details of data exclusions are available in the Methods section (Quantification of gene expression, and genotyping and imputation). For all the boxplots, horizontal lines inside the boxes show the medians. Box bounds show the lower quartile (Q1, the 25th percentile) and the upper quartile (Q3, the 75th percentile). Whiskers are minima (Q1 – 1.5 × IQR) and maxima (Q3 + 1.5 × IQR), where IQR is the interquartile range (Q3–Q1). Outliers were not shown in the boxplots. The experiments were not randomized because all the datasets are publicly available and from observational studies. The investigators were not blinded to allocation during the experiments and outcome assessment because the data are not from controlled randomized studies.

Ethics. Ethical approval for this project was obtained from the US Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center's Institutional Animal Care and Use Committee (Protocol 16-016).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All raw data analyzed in this study are publicly available for download without restrictions from SRA (<https://www.ncbi.nlm.nih.gov/sra/>) and BIGD (<https://www.bigsdb.org/>)

bigd.big.ac.cn/bioproject/) databases. Details of RNA-seq, WGBS and WGS can be found in Supplementary Tables 1, 2 and 15, respectively. All processed data, the full summary statistics of QTL mapping are available at <https://cgtex.roslin.ed.ac.uk/>.

Code availability

All the computational scripts and codes for RNA-seq and DNA methylation data quantification, quality control, gene expression normalization, genotype imputation, QTL mapping, functional enrichment, TWAS and colocalization are available at both the web portal of CattleGTEx (<https://cgtex.roslin.ed.ac.uk/>) and the github website (<https://github.com/shuliliu/cattleGTEx>, <https://doi.org/10.5281/zenodo.6510550>)⁵⁷.

References

25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
26. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
27. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
28. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
29. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
30. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
31. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
32. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
33. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat. Commun.* **7**, 12817 (2016).
34. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Ramirez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
37. Servant, N. et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
38. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
39. Song, Q. et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS ONE* **8**, e81148 (2013).
40. Liu, H. et al. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res.* **44**, 75–94 (2016).
41. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
42. Ongen, H., Buil, A., Brown, A. A., Dermizakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
43. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).
44. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.* **98**, 1114–1129 (2016).
45. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
46. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
47. Derrien, T. et al. Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
48. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
49. Saha, A. & Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors. *F1000Res* **7**, 1860 (2018).
50. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
51. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
52. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* **15**, e1007889 (2019).
53. Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* **13**, e1006646 (2017).
54. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
55. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046–2048 (2013).
56. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
57. Liu, S. et al. A multi-tissue atlas of regulatory variants in cattle. Code resource at github website. *GitHub*: <https://github.com/shuliliu/cattleGTEx>; *Zenodo*: <https://doi.org/10.5281/zenodo.6510550>

Acknowledgements

This work was supported in part by Agriculture and Food Research Initiative (AFRI) grant numbers 2016-67015-24886, 2019-67015-29321 and 2021-67015-33409 from the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA) Animal Genome and Reproduction Programs, and US–Israel Binational Agricultural Research and Development (BARD) grant number US-4997-17 from the BARD Fund. L.F. was partially funded through Health Data Research UK (HDRUK) award HDR-9004 and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 801215. A.T. acknowledged funding from the Biotechnology and Biological Sciences Research Council through program grants BBS/E/D/10002070 and BBS/E/D/30002275, Medical Research Council research grant MR/P015514/1 and HDRUK award HDR-9004. O.C.-X. was supported by MR/R025851/1. R.X. was supported by Australian Research Council's Discovery Projects (DP200100499). Y. Yu. was supported by the National Science Foundation of China–Pakistan Science Foundation Joint Project (31961143009) and National Key R&D Program of China (2021YFD1200900 and 2021YFD1200903). L.M. was supported in part by AFRI grant numbers 2020-67015-31398 and 2021-67015-33409 from the NIFA. G.E.L., B.D.R. and C.P.V.T. were supported by appropriated project 8042-31000-001-00-D, 'Enhancing Genetic Merit of Ruminants Through Improved Genome Assembly, Annotation, and Selection' of the Agricultural Research Service (ARS) of the USDA. C.-J.L. was supported by appropriated project 8042-31310-078-00-D, 'Improving Feed Efficiency and Environmental Sustainability of Dairy Cattle through Genomics and Novel Technologies' of ARS-USDA. J.B.C. was supported by appropriated project 8042-31000-002-00-D, 'Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals' of ARS-USDA. This research used resources provided by the SCINet project of the ARS-USDA project number 0500-00093-001-00-D. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer. All the funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank US dairy producers for providing phenotypic, genomic and pedigree data through the Council on Dairy Cattle Breeding under ARS-USDA Material Transfer Research Agreement 58-8042-8-007. Access to 1000 Bull Genomes Project data was provided under ARS-USDA Data Transfer Agreement 15443. International genetic evaluations were calculated by the International Bull Evaluation Service (Interbull; Uppsala, Sweden).

Author contributions

L.F., A.T. and G.E.L. conceived and designed the project. S.L., Y.G., O.C.-X., S.W., L.F., R.X., W.C., B.L., C.X., Y. Yao, Z.Y. and X.L. performed bioinformatic analyses. O.C.-X., L.F., Y. Yu, E.P.-C., K.D., K.R., C.L., A.J.C., P.N., D.R., B.D.R., C.P.V.T., P.M.V., S.Z., L.M., J.B.C., G.E.L. and A.T. contributed to the resource generation. S.L., L.F., Y.G., G.E.L. and A.T. wrote the manuscript. All authors read, edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

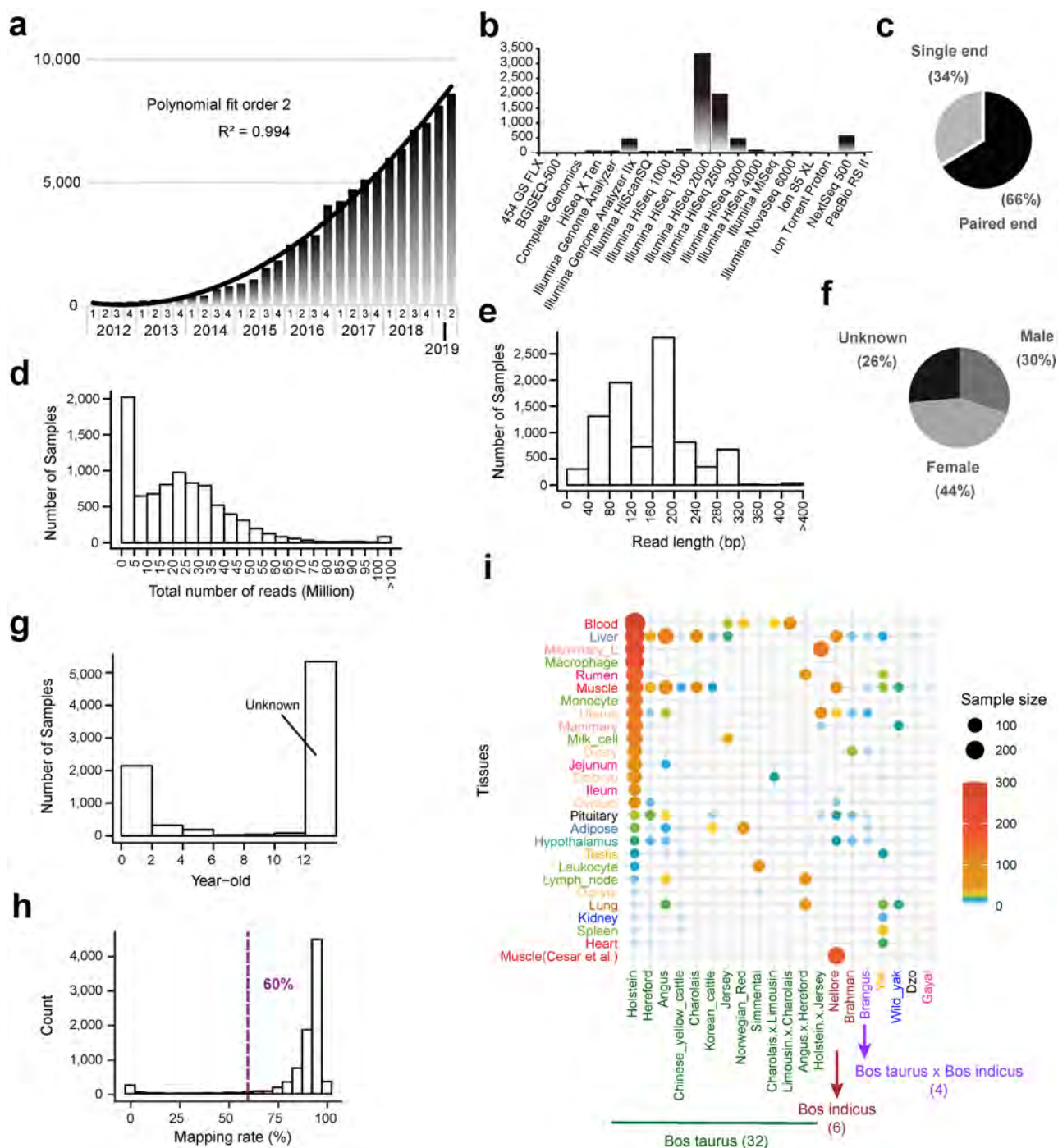
Extended data is available for this paper at <https://doi.org/10.1038/s41588-022-01153-5>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-022-01153-5>.

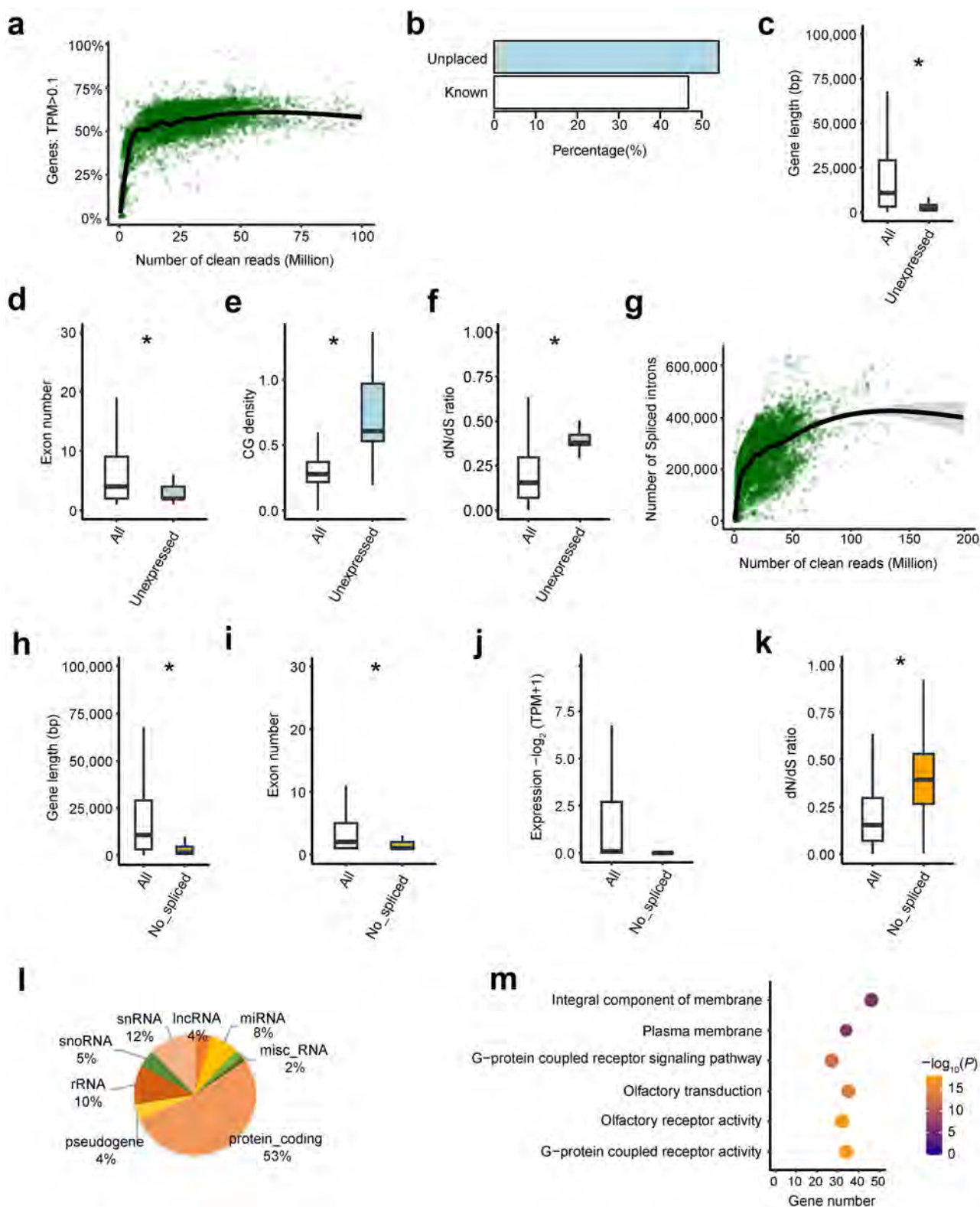
Correspondence and requests for materials should be addressed to George E. Liu, Albert Tenesa or Lingzhao Fang.

Peer review information *Nature Genetics* thanks Ben Hayes and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

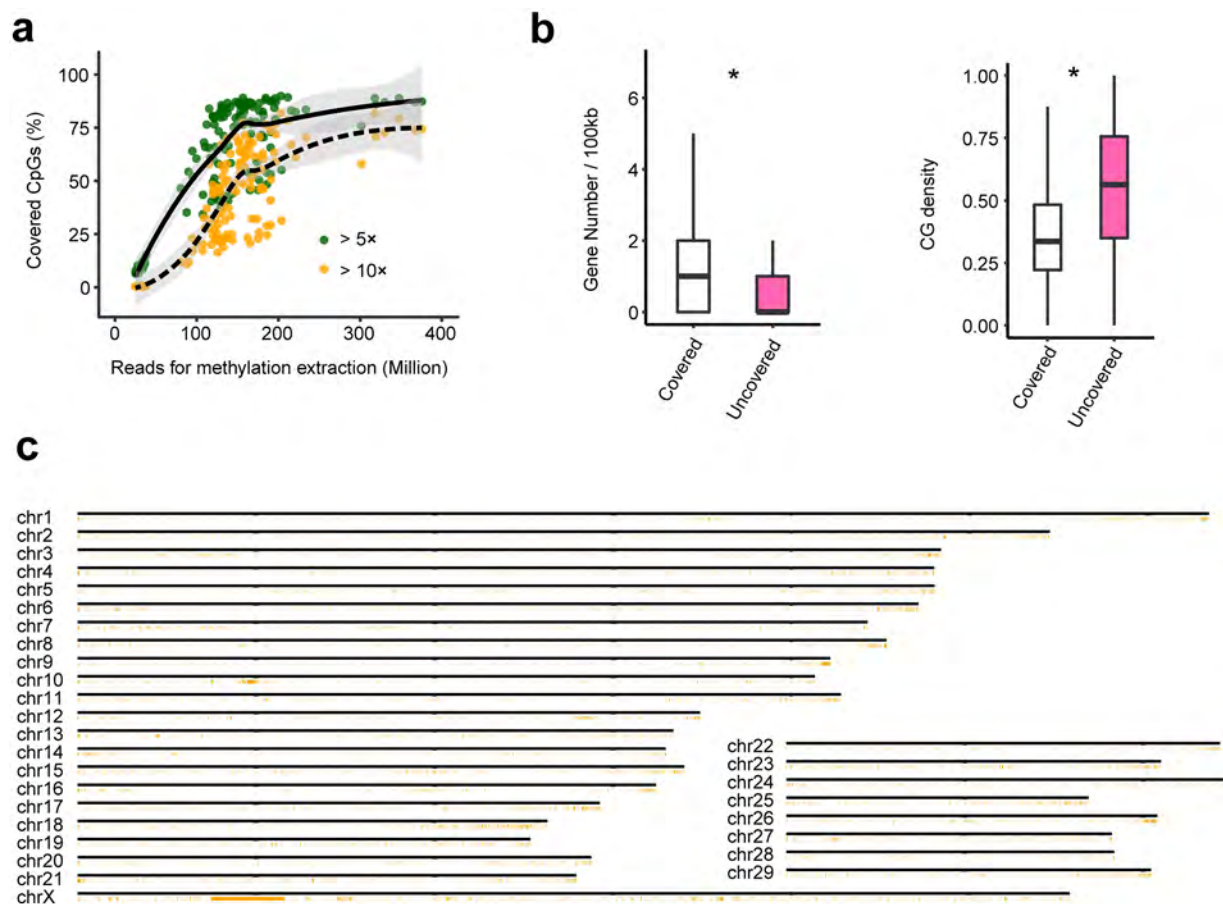


Extended Data Fig. 1 | Data summary of publicly available RNA-Seq data. (a) The number of publicly available RNA-Seq samples increases rapidly over years by fitting a second order polynomial model. (b) Distribution of sequence platforms of all 8,536 RNA-Seq samples. (c) Percentage of RNA-seq with single or paired reads. (d) Distribution of numbers of clean reads across all samples. (e) Distribution of read lengths. (f) Distribution of sexes. (g) Distribution of ages (Year-old). (h) Distribution of uniquely mapping rates. (i) Distribution of major tissues and breeds/ancestries in the 7,180 high quality RNA-Seq datasets (clean read > 500,000 & mapping rate > 60%).

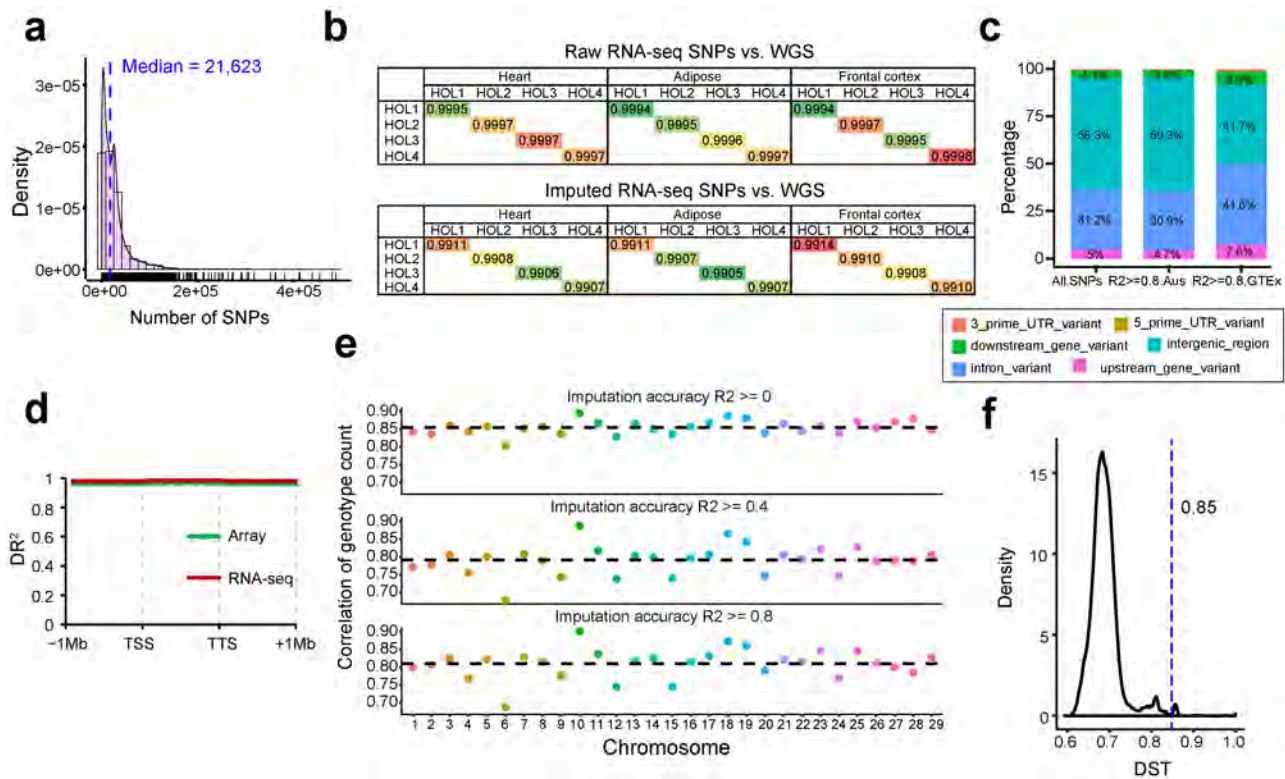


Extended Data Fig. 2 | See next page for caption.

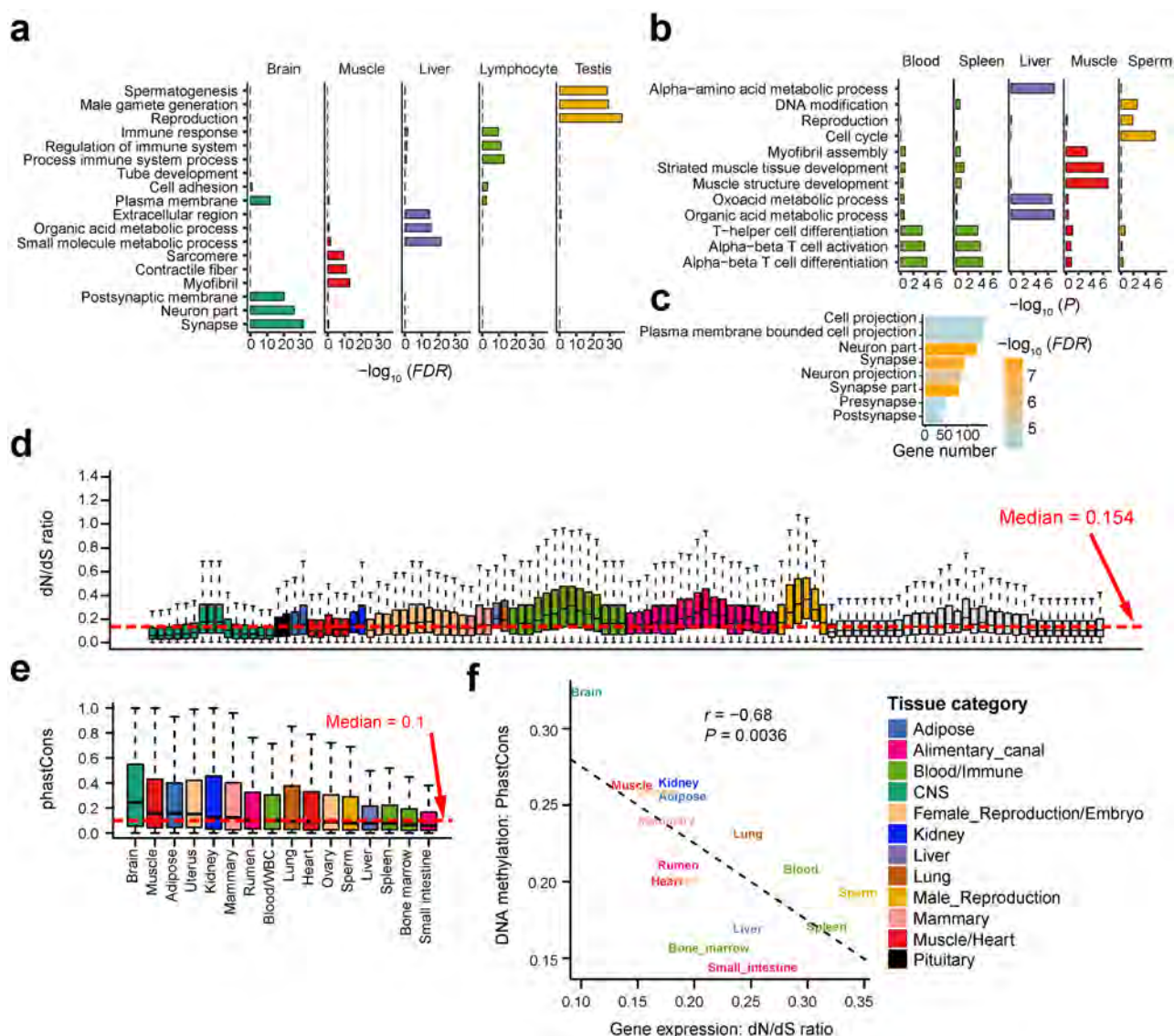
Extended Data Fig. 2 | Gene expression and alternative splicing profiles across samples. (a) Number of expressed genes (Transcripts per Million, TPM > 0.1) increases rapidly with the increasing number of clean reads across all 8,536 samples, reaching a plateau at 50 million reads. The black line is the smoothed curve fitted by a generalized additive model using `geom_smooth` function from `ggplot2` (v3.3.6) in R (v3.4.1). The shaded area around the lines represents the 95% confidence interval for the fitted values (the line). **(b)** The percentage of unexpressed genes (TPM < 0.1 across all samples) on known chromosomes (Known) and unplaced scaffolds (Unplaced, 54.10%). **(c-f)** Compared to expressed genes, the unexpressed genes have shorter gene length (df = 21,921, $P = 2.2 \times 10^{-4}$) **(c)**, fewer exons (df = 27,675, $P = 2.5 \times 10^{-5}$) **(d)**, higher CG density (df = 21,921, $P = 1.5 \times 10^{-103}$) **(e)**, and higher dN/dS ratio (df = 19,718, $P = 5.4 \times 10^{-21}$) **(f)**. **(g)** The number of spliced introns increases rapidly with the increasing number of clean reads across samples, reaching a plateau at 100 million reads. The smoothed curve and the shaded band are obtained using the same method as in **(a)**. **(h-k)** Compared to all genes, genes without spliced introns in any tissues have shorter gene length (df = 22,320, $P = 2.9 \times 10^{-18}$) **(h)**, fewer exons (df = 17,690, $P = 7.4 \times 10^{-52}$) **(i)**, lower expression levels (median gene expression levels across samples, df = 28,479, $P = 0.35$) **(j)**, and higher dN/dS ratio (df = 19,921, $P = 3.7 \times 10^{-32}$) **(k)**. All the P values above are obtained based on the two-sided Welch two sample t -test, and * indicates $P < 0.05$. **(l)** Distribution of gene types for those without spliced introns. **(m)** Significant terms ($P < 0.05$) of Gene Ontology for genes without spliced introns based on the hypergeometric test.



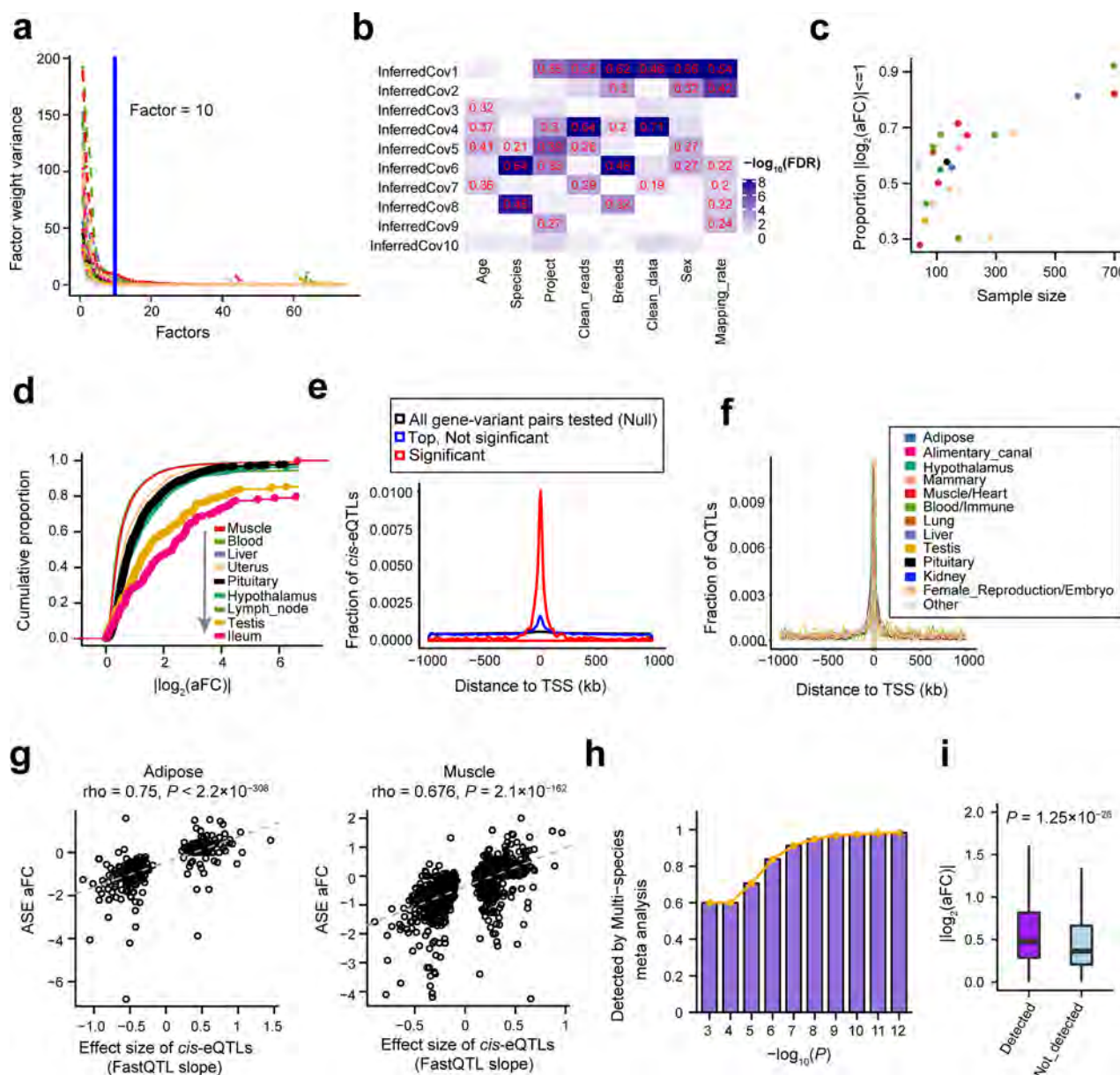
Extended Data Fig. 3 | DNA methylation profiles of 144 WGBS samples. (a) The percentage of covered CpGs (read depth $\geq 5\times$ or $\geq 10\times$) in the entire genome increases rapidly with the increasing number of reads used for methylation extraction, approximately reaching a plateau at 200 million reads. The black solid line and dash line are the smoothed curves fitted by a generalized additive model using `geom_smooth` function from `ggplot2` (v3.3.6) in R (v3.4.1) for read depth $\geq 5\times$ and $\geq 10\times$, respectively. The shaded area around the lines represents the 95% confidence interval for the fitted values (the lines). **(b)** Compared to covered CpGs (Covered), the uncovered CpGs (read depth $< 5\times$ across all samples, Uncovered) tend to be located within gene deserts ($df = 15,074,753$, $P < 2.2 \times 10^{-308}$) and regions with higher CG density ($df = 15,074,753$, $P < 2.2 \times 10^{-308}$). All the P values above are obtained based on the two-sided Welch two sample t -test, and * indicates $P < 0.05$. **(c)** Distribution of uncovered CpGs ($< 5\times$) along the entire genome.



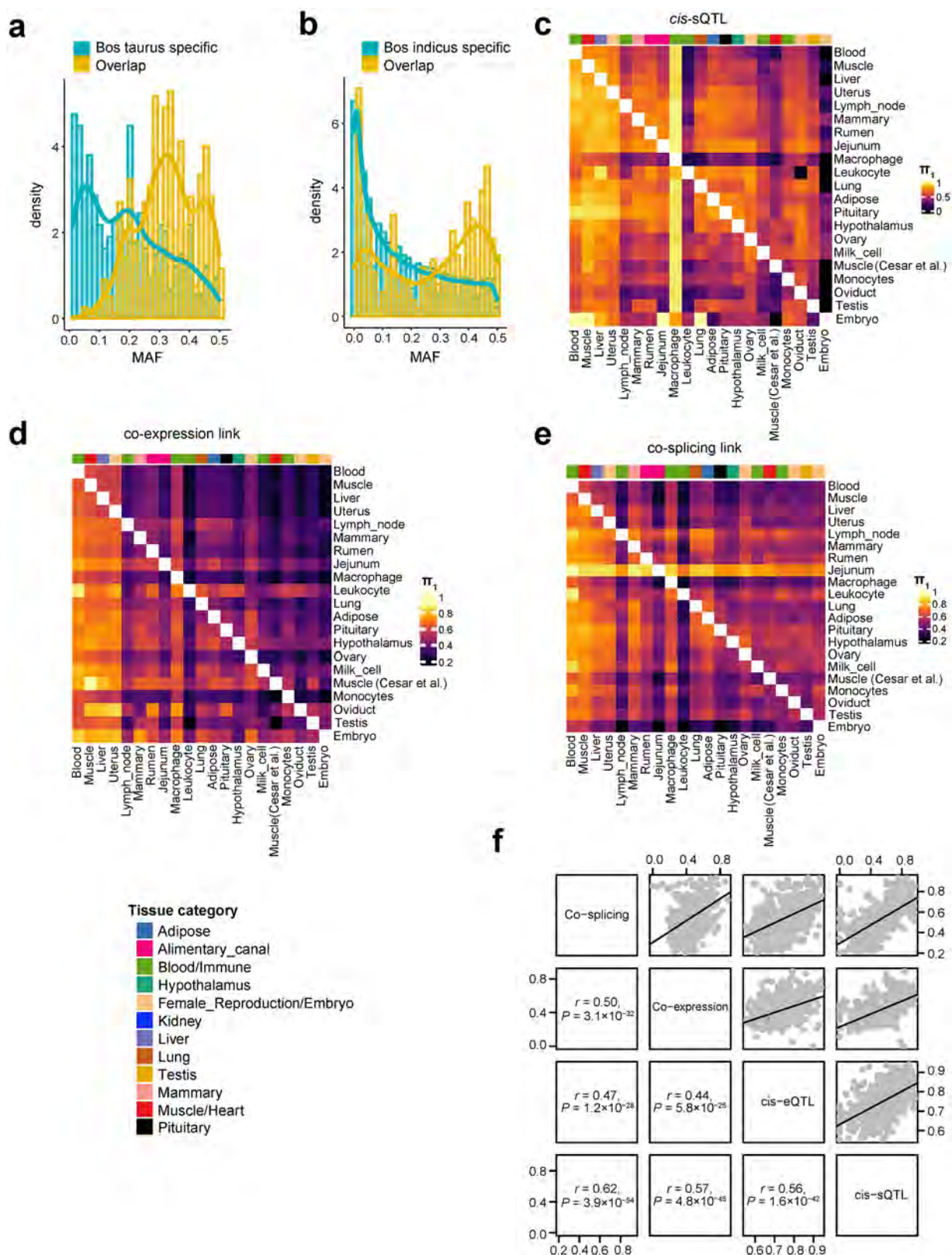
Extended Data Fig. 4 | Genotyping and imputation of variants from RNA-Seq data. (a) Distribution of numbers of SNPs directly called from RNA-Seq data across all 8,536 samples. (b) Concordance rates between genotypes (mean = 78,587, range = 47,407–113,868) called from RNA-Seq data and imputed genotypes (mean = 2.50 million, range = 1.20–2.73 million) in three tissues and those called from whole genome sequencing (WGS) data across four Holstein (HOL) animals. (c) Proportion of variants within functional categories using different imputation accuracy cutoffs. These results are derived from 109 Holstein animals with both RNA-seq and 50 K SNP array. 'All.SNPs' are those 31,377,923 imputed variants common in the two imputation processes (that is, the genotype imputation based on RNA-Seq SNPs and that based on SNP array). 'imp.acc ≥ 0.80.Aus' are those imputed based on 50 K SNP array genotypes (Australian HOL animals) and variants with imputation accuracy $DR^2 > 0.80$ were selected ($n = 16,501,943$). 'imp. acc ≥ 0.80.GTEX' are those in the CattleGTEx data where the imputation was based on RNA-seq SNPs and variants with imputation accuracy $DR^2 > 0.80$ were selected ($n = 5,292,828$). (d) Comparison of DR^2 of SNPs imputed from SNP array (50 K) and those imputed from RNA-Seq SNPs along 1 Mb up-/down- stream of gene body. The up-/down- stream is divided into windows of 100 kb length, while the gene body region of each gene is evenly divided into 10 windows. The DR^2 values of SNPs within each window are then averaged for plotting. (e) Pearson correlations of genotype counts between variants imputed from RNA-Seq SNPs and those from 50 K SNP arrays across different imputation quality cutoffs and chromosomes. The horizontal dashed line in each graph indicates the mean of correlations across chromosomes. (f) Distribution of identity by state (IBS) distance between all sample pairs. The IBS distance is calculated using PLINK v1.90 to measure the average proportion of alleles shared between samples. The sample pairs with IBS distance > 0.85 are considered as duplicated samples.



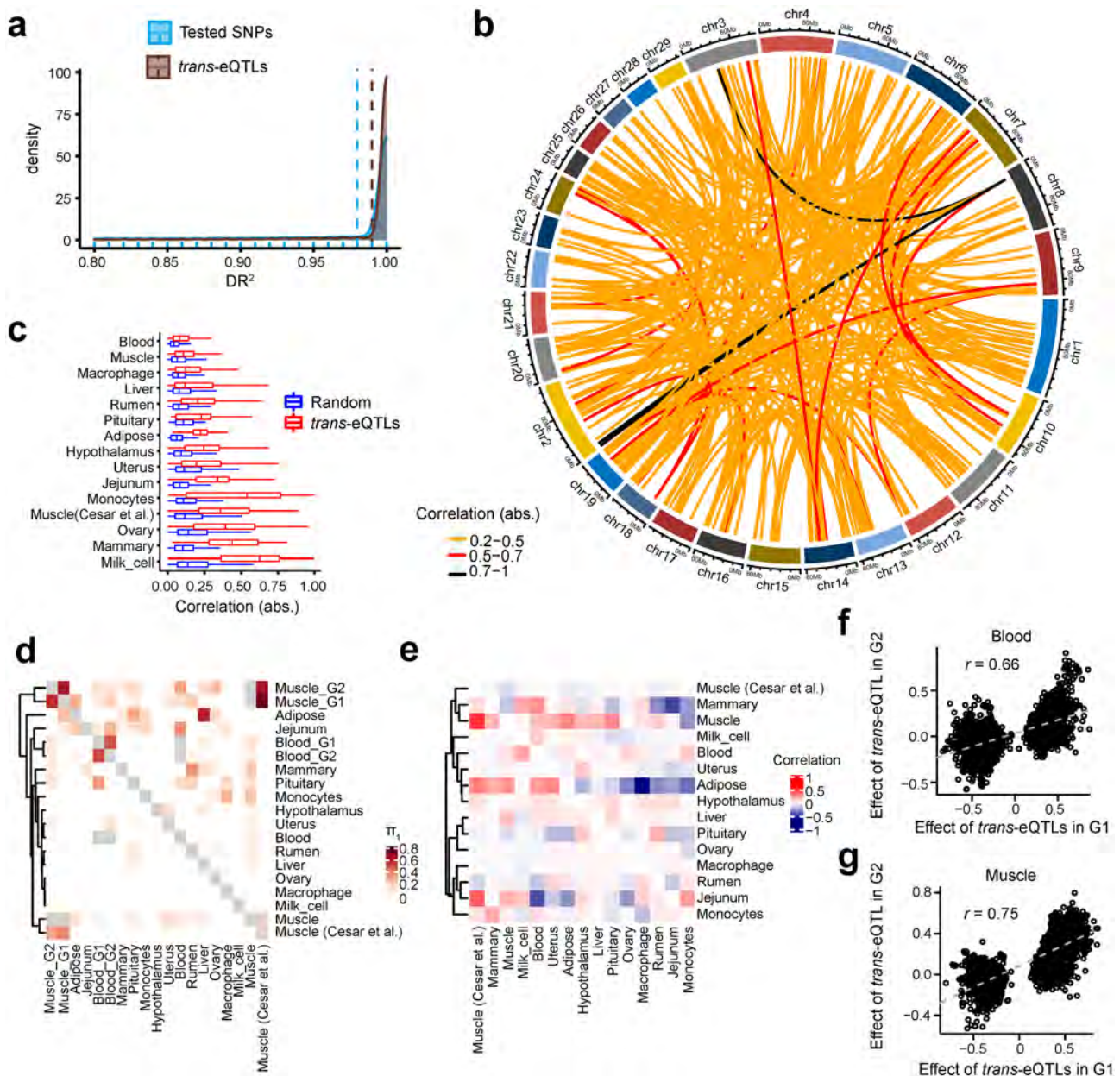
Extended Data Fig. 5 | Functional annotation of tissue-specific genes and their sequence conservation in mammals. (a) Significant Gene Ontology (GO) terms for genes with tissue-specific expression, based on the one-sided Fisher's exact test using ClusterProfiler v3.0.4. FDR is obtained after the Benjamini-Hochberg correction for the raw P value. **(b)** Significant GO terms for genes with tissue-specific hypomethylated promoters ($P < 0.05$). **(c)** Significant GO terms for genes with brain-specific spliced introns (Benjamini-Hochberg corrected P (FDR) < 0.05 after correction). **(d)** dN/dS ratio (between cattle and humans) of orthologous genes with tissue-specific expression across tissues. The red dash line indicates median value of 0.154. **(e)** PhastCons scores of regions with tissue-specific hypomethylation across tissues. PhastCons scores were obtained from UCSC website and calculated on the basis of DNA sequences of 46 placental mammals. The red dash line indicates the median value of 0.1. **(f)** The Pearson's correlation ($r = -0.68$, the two-sided Student's t -test: $P = 0.0036$) between PhastCons scores of tissue-specific DNA methylation regions and dN/dS ratios of tissue-specific expressed genes across 16 common tissues. **(f)** has the same color key as **(d)**.



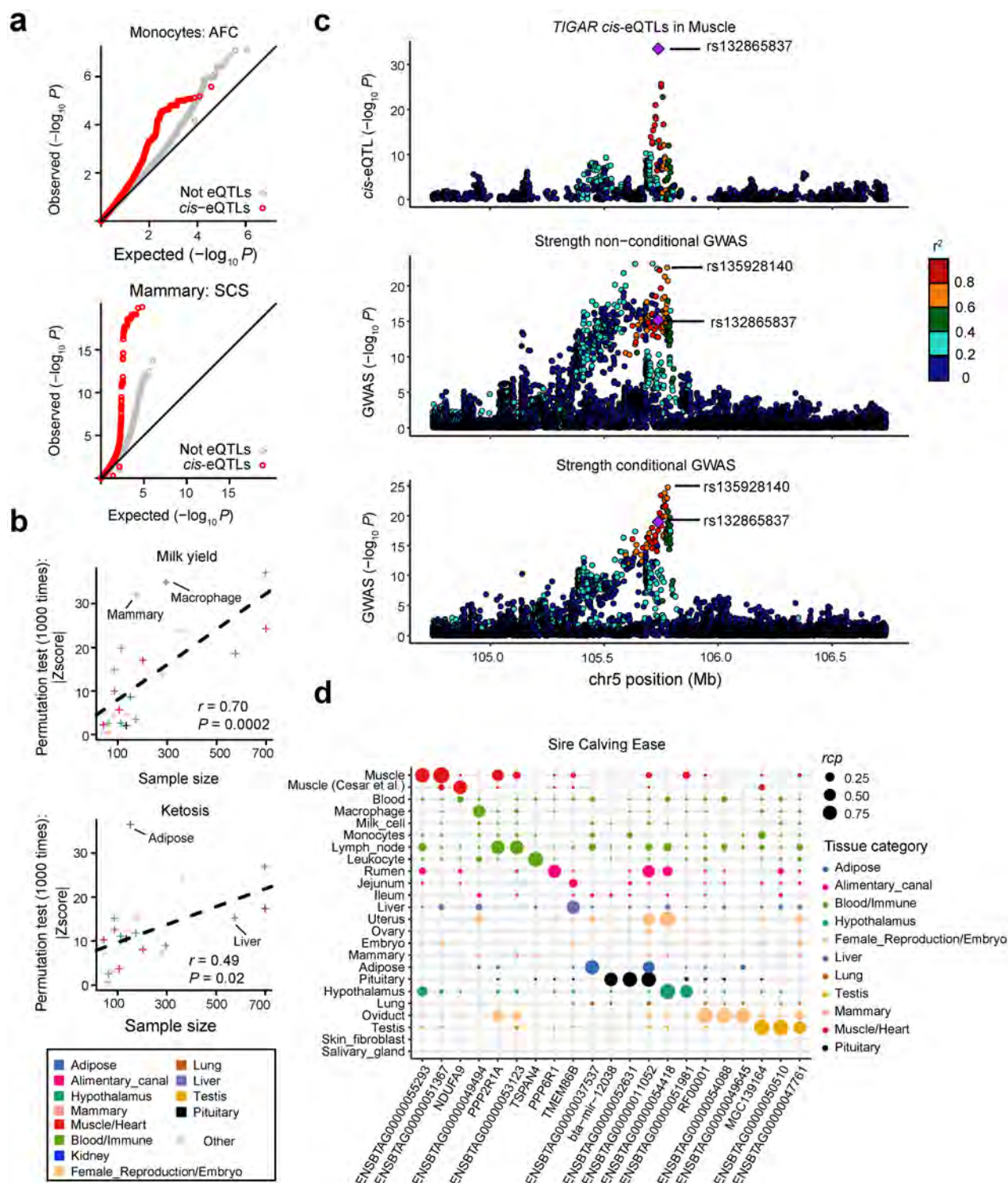
Extended Data Fig. 6 | Characterization of *cis*-eQTLs across tissues. (a) Factor weight variance as a function of PEER factors computed up to 75 factors for each of 23 distinct tissues. Factor weight variances become small for most of tissues when the number of inferred hidden PEER factors reaches 10. (b) Pearson's correlation between inferred factors and known covariates in adipose. The color in each cell denotes $-\log_{10}(\text{FDR})$ after the Benjamini-Hochberg correlation of P values (the two-sided Student's t -test). Only significant correlations ($\text{FDR} < 0.05$) are shown in cells. (c) The proportion of *cis*-eQTLs with $|\log_2(\text{aFC})| \leq 1$ over all *cis*-eQTLs as a function of sample size across 23 distinct tissues. $|\log_2(\text{aFC})|$, that is, the \log_2 transformed allelic fold change, which is used to measure the effect size of *cis*-eQTL. (d) The *cis*-eQTL cumulative proportion plot of $|\log_2(\text{aFC})|$ distribution across 9 tissues with variable sample sizes. The arrow indicates tissues in legend were listed from largest to smallest sample size. (e) Distribution of *cis*-eQTLs around TSS (1 Mb up- and down-stream) in adipose. All gene-variant pairs tested as null; 'Significant' indicates the top eQTLs for significant eGenes; 'Top, Not significant' indicates the top associated SNP for non-significant genes (non-eGenes). (f) Distribution of *cis*-eQTLs around the TSS (1 Mb up- and down-stream) across all 23 distinct tissues. (g) Correlation of effect sizes (fastQTL slope) of *cis*-eQTLs and aFC of matched loci tested by allelic specific expression (ASE) analysis in adipose (Spearman's $\rho = 0.75$, the two-sided Student's t -test: $P < 2.2 \times 10^{-308}$) and muscle (Spearman's $\rho = 0.68$, the two-sided Student's t -test: $P = 2.1 \times 10^{-162}$). (h) Percentage of *cis*-eQTLs in the combined muscle data that are replicated in multi-subspecies meta-analysis at different P -value cutoffs used for defining *cis*-eQTLs. The *cis*-eQTLs with higher significant levels are more likely to be specifically detected in the combined population. (i) Effects sizes ($|\log_2(\text{aFC})|$) of *cis*-eQTLs specifically detected in combined population are significantly (the two-sided Welch two sample t -test: $P = 1.25 \times 10^{-26}$) smaller than those that are replicated in multi-breed meta-analysis.



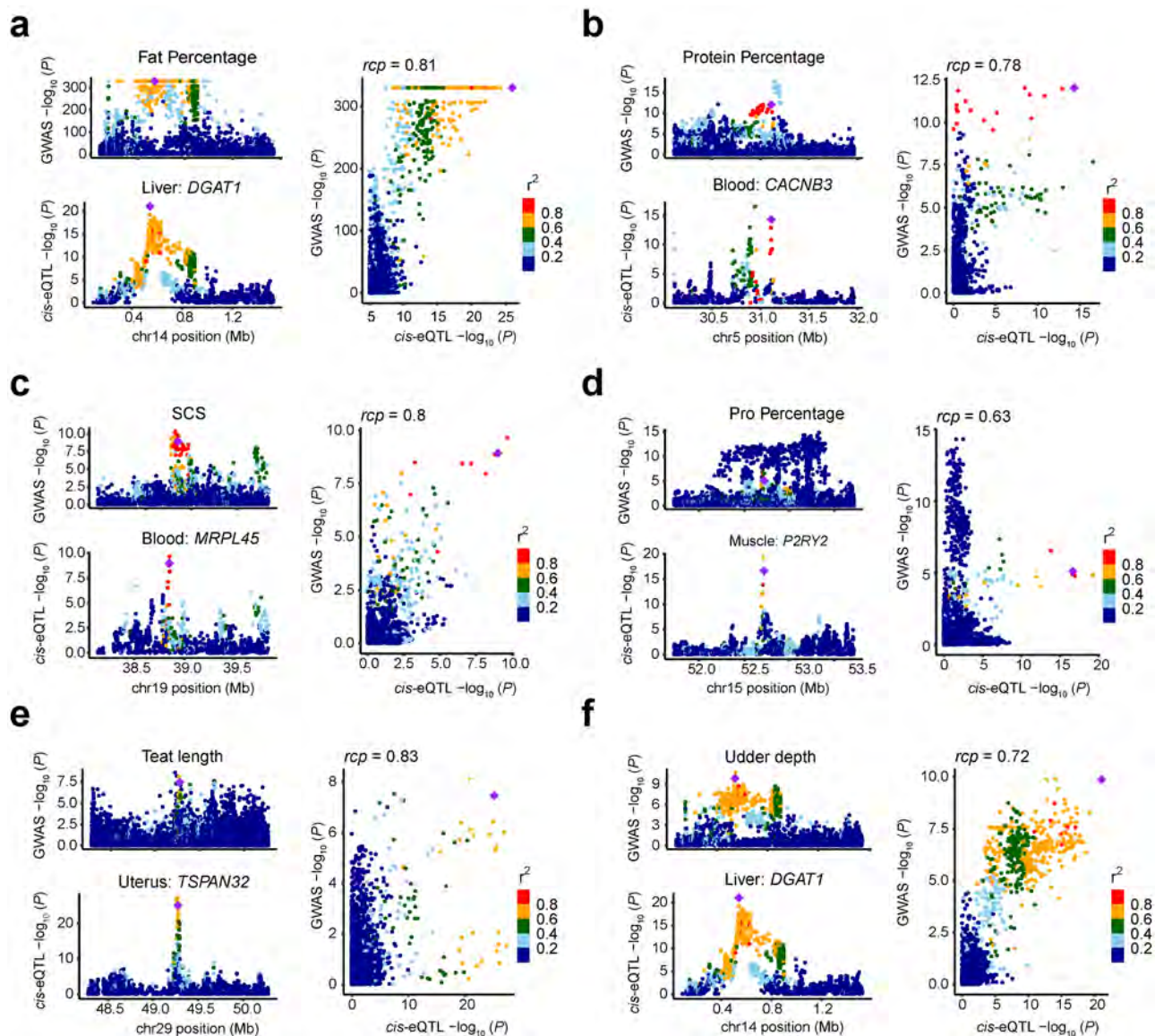
Extended Data Fig. 7 | Sharing of QTLs across ancestries and tissues. (a) Distribution of minor allele frequency (MAF) of loci with *bos taurus* specific ($n=202,583$) and overlapped ($n=459$) *cis*-eQTLs in *bos indicus* population. (b) Distribution of MAF of loci with *bos indicus* specific ($n=437$) and overlapped *cis*-eQTLs in *bos taurus* population. (c) Tissue-sharing patterns of *cis*-sQTL. (d) The gene co-expression patterns across tissues. (e) The co-splicing patterns of spliced introns across tissues. The π_1 values are calculated to measure the replication rates between tissues. (f) The Pearson's correlation of π_1 values of matched tissue-pairs (that is, values in the tissue-sharing heatmaps above) across four data types. The P value is obtained by the two-sided Student's t -test.



Extended Data Fig. 8 | Characterization and internal replications of trans-eQTLs. (a) Comparison of imputation quality (DR^2) of *trans*-eQTLs and all tested SNPs. Dashed lines are median values of DR^2 . (b) Inter-chromosomal linkage disequilibrium (LD) (the genotype correlation in absolute values) between *trans*-eQTLs and *cis*-eQTLs of the same genes in muscle. (c) Comparison of LD of *cis*-eQTLs vs. *trans*-eQTLs of the same genes and that of *cis*-eQTLs vs. random SNPs with matched minor allele frequency (MAF) and chromosomes. The comparisons of all tissues are statistically significant ($P < 0.05$, the two-sided Student's *t*-test). Box plots depict the interquartile range (IQR), whiskers depict $1.5 \times IQR$. (d) Tissue-sharing patterns (π_1 statistics) of *trans*-eQTLs across tissues and replicates. Muscle_G1 ($n = 435$) and Muscle_G2 ($n = 435$) are two replicates of muscle samples by dividing the whole muscle samples randomly into two groups. Similarly, Blood_G1 ($n = 349$) and Blood_G2 ($n = 349$) are two replicates of blood samples. (e) Pearson correlations of effect sizes (β values) of *trans*-eQTLs in one tissue (x-axis) and those of matched SNPs in another tissue (y-axis). (f) Pearson correlation of effect sizes (β values) of *trans*-eQTLs ($n = 5,782$) in blood tissue in Group1 (G1, $n = 349$) and those of matched SNPs in Group2 (G2, $n = 349$) ($r = 0.66$, the two-sided Student's *t*-test: $P < 2.2 \times 10^{-308}$). (g) Pearson correlation of effect sizes (β values) of *trans*-eQTLs ($n = 4,344$) in muscle tissue in Group1 (G1, $n = 435$) and those of matched SNPs in Group2 (G2, $n = 435$) ($r = 0.75$, the two-sided Student's *t*-test: $P < 2.2 \times 10^{-308}$).



Extended Data Fig. 9 | Associations of cis-eQTLs and GWAS loci for important agronomic traits in cattle. (a) cis-eQTLs discovered in monocytes, and mammary gland show enrichments for top SNPs (top 10%) associated with age at first calving (AFC) ($P = 0.001$, the two-sided permutation test with 1,000 times), and somatic cell score (SCS) ($P = 0.001$, the two-sided permutation test with 1,000 times) respectively, compared to the null expectation (shown in gray) defined by 'Not eQTLs'. (b) Pearson correlation between z-scores from permutation tests (1000 times) and sample sizes of cis-eQTL tissues for milk yield trait (top, $r = 0.70$, the two-sided Student's t -test: $P = 0.0002$) and ketosis trait (bottom, $r = 0.49$, the two-sided Student's t -test: $P = 0.02$). (c) An example of a colocalization of cis-eQTLs of *TIGAR* gene in muscle and GWAS loci of strength in cattle on chromosome 5. Four independent GWAS signals (that is, rs210875465, rs381714832, rs1115089453 and rs135928140) are located within the region. The All-but-One conditional analysis across the individual GWAS signals shows that only rs135928140 in strength GWAS is colocalized with cis-eQTLs of *TIGAR* in muscle, when conditioning on the remaining three signals. The colocalized SNP (that is, rs132865837) of *TIGAR* in muscle is in LD ($r^2 = 0.49$) with the GWAS loci rs135928140. (d) Colocalization between GWAS loci of sire calving ease (Sire_Calv_Ease) in cattle and cis-eQTLs across 23 distinct tissues.



Extended Data Fig. 10 | Locuscompare plots for six colocalized events detected by two TWAS methods (S-PrediXcan and MetaXcan), fastENLOC and Coloc simultaneously. (a) eQTLs of *DGAT1* colocalized with GWAS signals of fat percentage in liver. **(b)** eQTLs of *CACNB3* colocalized with GWAS signals of protein percentage in blood. **(c)** eQTLs of *MRPL45* colocalized with GWAS signals of somatic cell score (SCS) in blood. **(d)** eQTLs of *P2RY2* colocalized with GWAS signals of protein percentage in muscle. **(e)** eQTLs of *TSPAN32* colocalized with GWAS signals of Teat length in uterus. **(f)** eQTLs of *DGAT1* colocalized with GWAS signals of udder depth in liver.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☐ ☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

All raw data analyzed in this study are publicly available for download without restrictions from SRA (<https://www.ncbi.nlm.nih.gov/sra/>) and BIGD (<https://bigd.big.ac.cn/bioproject/>) databases. We downloaded them by June 24th 2019.

Data analysis

All the computational scripts and codes for RNA-seq and DNA methylation data quantification, quality control, gene expression normalization, genotype imputation, QTL mapping, functional enrichment, TWAS and colocalization are available at both the web portal (<https://cgtex.roslin.ed.ac.uk/>) and the github website with DOI-minting (Zenodo, <https://zenodo.org/record/6510550#.Ym-FANrMKUk>) (<https://github.com/shuliliu/cattleGTEx>).

For RNA-Seq data analysis, we used Trimmomatic (v0.39), STAR (v2.7.0), Stringtie (v2.1.1), featureCounts (v1.5.2), Leafcutter (v0.2.9), (GATK) (v4.0.8.1), GATK ASEReadCounter tool (v4.0.8.1), and Beagle (v5.1) for quality control, mapping, gene expression quantification, alternative splicing, SNP calling, ASE analysis, and genotype imputation, respectively.

For WGBS, we used Trim Galore v0.4.0, Bismark (v0.14.5) and SMART(v2) for quality control, read mapping and DNA methylation level extraction, and hypomethylation region detection, QTL respectively.

For Hi-C, we used Trim Galore (v0.4.0), BWA(v0.7.17), HiCExplorer v3.4.1, HiC-Pro (v2.11.4) and FitHiC (v2.0.7) for quality control, read mapping, Hi-C contact matrix construction, and significant intra-chromosome contact detection, respectively.

For QTL mapping, we used FastQTL (v2.184), aFC (v0.3), dap-g (v1.0.0), METAL, MashR (v0.2.57), MatrixQTL (v2.3) and GCTA (v1.93.3beta) for cis-QTL mapping, effect size estimation, fine-mapping, meta-analysis, tissue-sharing pattern estimation, and trans-QTL mapping, respectively.

For TWAS, we used S-PrediXcan (v0.6.1) and S-MultiXcan (v0.6.1) for single-tissue and multi-tissue TWAS analysis, respectively.

For the colocalization analysis between cis-QTL and GWAS loci, we used both fastENLOC (v1.0) and Coloc (v5.1.0).

For enrichment analysis, we used GAT(v1.3.4) and ClusterProfiler (v3.0.4) for eQTLs and genes functional annotation, respectively.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All raw sequencing data analyzed in this study are publicly available in NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>). Details of these data can be found in Supplementary Table 1-2, 14. All processed data, the full summary statistics of QTL mapping are available at <https://cgtex.roslin.ed.ac.uk/>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No power calculation was needed in advance in this study. In total, we analyzed all 11,642 RNA-Seq runs (downloaded by July, 2019) from SRA (n = 11,513, https://www.ncbi.nlm.nih.gov/sra/) and BIGD databases (n = 129, https://bigd.big.ac.cn/bioproject/), yielding 8,536 unique RNA-Seq samples. After filtering the samples with low quality (see below), all the remaining samples have been used for analysis.
Data exclusions	<p>Full details of data exclusions for each analysis can be found in the Methods as well.</p> <p>We filtered out RNA-Seq samples with clean read counts $\leq 500K$ or uniquely mapping rates $< 60\%$, resulting in 7,264 samples. We further excluded samples with obvious clustering errors (e.g., samples labeled as liver that were not clustered with other liver samples), resulting in 7,180 samples for subsequent analysis.</p> <p>For cis-QTLs detection, we excluded tissues with less than 40 individuals, resulting in 23 distinct tissues for cis-QTL mapping. While, for trans-QTLs we excluded tissues with less than 100 individuals, resulting in 15 tissues for trans-QTL mapping.</p>
Replication	To conduct an internal validation for both cis-eQTL and trans-eQTLs, we randomly and evenly divided the blood samples into two groups (G1 and G2), each of which has 349 samples. We then conducted trans-eQTL analysis using the linear mixed model in each group separately. We observed that $\pi_1 = 0.48$ between G1 and G2, and the the beta-values of trans-eQTLs (n = 5782) in G1 were significantly correlated ($r = 0.66$ with $p < 1e-22$) with those of the matched variants in G2. We observed similar results for muscle samples ($\pi_1 = 0.62$, and $r = 0.75$ with $p < 1e-22$). While, the internal replication (π_1) of cis-eQTLs were 0.97 and 0.95 for blood and muscle, respectively.
Randomization	All the datasets are from observation studies and we used all samples publicly available after data exclusions listed above. Therefore, Randomization were not relevant in this study. Samples were grouped by tissue types.
Blinding	Blinding were not relevant in this study, as we re-analyzed all the publicly available RNA-seq data rather than data from controlled randomized studies in a uniform pipeline.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

- | | |
|-------------------------------------|--|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Human research participants |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |

Methods

- | | |
|-------------------------------------|---|
| n/a | Involved in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

Article

Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods

Yahui Gao ^{1,2}, Li Ma ²  and George E. Liu ^{1,*} 

¹ Animal Genomics and Improvement Laboratory, Beltsville Agricultural Research Center, Agricultural Research Service, U.S. Department of Agriculture, Beltsville, MD 20705, USA; gyhalvin@gmail.com

² Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA; lima@umd.edu

* Correspondence: george.liu@usda.gov; Tel.: +1-301-504-9843

Abstract: Structural variations (SVs), as a great source of genetic variation, are widely distributed in the genome. SVs involve longer genomic sequences and potentially have stronger effects than SNPs, but they are not well captured by short-read sequencing owing to their size and relevance to repeats. Improved characterization of SVs can provide more advanced insight into complex traits. With the availability of long-read sequencing, it has become feasible to uncover the full range of SVs. Here, we sequenced one cattle individual using 10× Genomics (10 × G) linked read, Pacific Biosciences (PacBio) continuous long reads (CLR) and circular consensus sequencing (CCS), as well as Oxford Nanopore Technologies (ONT) PromethION. We evaluated the ability of various methods for SV detection. We identified 21,164 SVs, which amount to 186 Mb covering 7.07% of the whole genome. The number of SVs inferred from long-read-based inferences was greater than that from short reads. The PacBio CLR identified the most of large SVs and covered the most genomes. SVs called with PacBio CCS and ONT data showed high uniformity. The one with the most overlap with the results obtained by short-read data was PB CCS. Together, we found that long reads outperformed short reads in terms of SV detections.

Keywords: cattle; structural variation; long-read sequencing



Citation: Gao, Y.; Ma, L.; Liu, G.E.

Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods.

Genes **2022**, *13*, 828. <https://doi.org/10.3390/genes13050828>

Academic Editor: Qiuyue Liu

Received: 15 April 2022

Accepted: 4 May 2022

Published: 6 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Unraveling the genetic underpinnings of phenotypic variation relies on a comprehensive knowledge of all forms of genetic variation. The exploitation of genetic variation has mainly focused on single-nucleotide polymorphisms (SNPs) and small insertions or deletions (indels, <50 bp), with a minor emphasis on larger variations such as copy number variations (CNV) and other structural variations (SV). SVs are most commonly defined as genomic changes of at least 50 bp in size, and they are difficult to detect precisely. Although there exist fewer SVs in the genome relative to SNPs and indels, SVs can impact more base pairs, thus being more likely to affect the phenotype [1,2]. While short-read sequencing technologies can detect SVs, they have various weaknesses. Since short reads (<1 kb) are typically smaller than or similar in size to SVs, a wide collection of indirect methods has been developed to infer SVs, including the use of split reads, read pairs, read depths, and local de novo assembly. On the other hand, linked reads provide long-range (100+ kb) information to short reads, bringing the reads into phase for haplotype-specific deletion detection, large SV detection [3–5], and diploid de novo assembly [6]. Long reads (>> 1 kb) spanning more SVs allow further SV detection, with mapped reads [7,8], local assembly after phasing long reads [9], and global de novo assembly [10,11]. Currently, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the most commonly employed technologies to produce long reads. Single-molecule real-time (SMRT) sequencing, developed by PacBio, can yield reads of tens of kilobases using either continuous long reads (CLR) or circular consensus sequencing (CCS) mode, which achieves high-quality

genome assembly. ONT enables direct and real-time sequencing of long DNA or RNA by analyzing the current interference caused by the molecules as they pass through the protein nanopore. To date, these sequencing methods have enabled the improved genome assemblies for many species, including humans [12,13], cattle [14–16], buffalo [17], pigs [18,19], sheep [20], and goats [21]. To study the effects of these methods on SV detection in humans, Aganezov et al. [22] performed whole-genome sequencing of the SKBR3 breast cancer cell line and patient-derived tumor and normal organoids from two breast cancer patients using Illumina/10× Genomics, PacBio, and ONT sequencing. They inferred SVs and large-scale CNVs and showed that long-read sequencing enables more accurate and sensitive SV detection. In dairy cattle, Couldrey et al. [15] detected CNVs using PacBio long-read and Illumina sequencing. In this study (Figure 1), we sequenced one cattle individual using cutting-edge technologies, i.e., 10× Genomics (10 × G), PromethION (ONT), PacBio continuous long reads (PB CLR), and PacBio circular consensus sequencing (PB CCS). We then evaluated various methods using these data from the same lung DNA sample for their abilities for the SV detection.

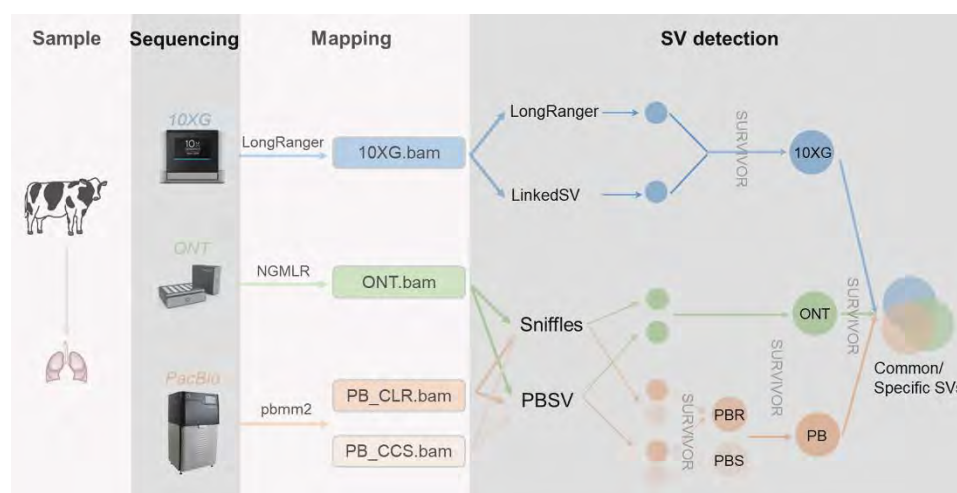


Figure 1. Sample collecting, sequencing, and mapping pipeline.

2. Materials and Methods

Under the approval of the US Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center’s Institutional Animal Care and Use Committee (Protocol 16-016), lung tissue was collected and then snap-frozen in liquid N₂ immediately after excision and kept at −80 °C until use. The high-molecular-weight (HMW) DNA for lung tissue was extracted according to the MagAttract HMW DNA Kit (Cat. No. 67563, QIAGEN, Valencia, CA, USA). The quality of DNA samples was evaluated using the 2100 Bioanalyzer and the 4200 TapeStation (both from Agilent Technologies, Santa Clara, CA, USA), including degradation, potential RNA contamination, purity (OD₂₆₀/OD₂₈₀), and concentration using spectrophotometers of Qubit (Thermo Fisher Scientific, Waltham, MA, USA) and NanoDrop (NanoDrop Technologies, Rockland, DE, USA) to meet the demands for library construction.

The HMW DNA was sequenced using the Linked-Reads method developed by 10× Genomics [4], and standard protocols were followed in this study. By using microfluidics to segment and barcode HMW DNA, 10× Genomics can provide long-range information for short reads of the genome. We then aligned 10 × G short reads with LongRanger [23] v2.1.6 and used LongRanger [23] v 2.1.6 and LinkedSV [24] with the recommended settings to call SVs, respectively. DNA was prepared using standard ONT methods and sequenced on a PromethION device. We aligned ONT long reads with NGMLR [8] v0.2.7 and run Sniffles [8] v1.0.11 and PBSV v2.2.0 (<https://github.com/PacificBiosciences/pbsv>, accessed on 3 May 2022) with default settings for SV inference. PacBio sequencing was carried out on a Pacific Biosciences Sequel II platform using two modes, i.e., continuous long

reads (CLR) and circular consensus sequencing (CCS). We aligned the long reads with pbmm2 v1.3.0 (<https://github.com/PacificBiosciences/pbmm2>, accessed on 3 May 2022) and run Sniffles v1.0.11 [8] and PBSV v2.2.0 (<https://github.com/PacificBiosciences/pbsv>, accessed on 3 May 2022) with default settings for SV inference. We mapped all reads against the latest cattle genome reference ARS-UCD1.2 [25] and performed follow-up SV detection. We computed the alignment coverage by SAMtools [26] v1.9 depth command. For each sequencing technology, we merged the SVs generated by different callers with the SURVIVOR [27] v1.0.7 into a $10 \times$ G, ONT, and PacBio technology-specific SV call sets. We then ran the SURVIVOR merge module with a maximum allowed distance of 200 bp and minimum SV size set to 30 bp regardless of SV types, as different methods may assign different types.

3. Results

3.1. SV Inference

A total of 1,577,259,728 (Table 1) short reads were generated through $10 \times$ Genomics, representing $55 \times$ coverage of the genome. The LongRanger alignment resulted in 97.14% (Table 1) of the reads mapping to the ARS-UCD1.2 cattle genome reference [25]. There was a total of 8315 and 6453 putative SVs identified by LongRanger and LinkedSV, respectively (Table 2). The SVs identified by LongRanger ranged in size from 49 bp to 1.59 Mb with an average size of 4481 bp (Table S1). For LinkedSV, the size ranged from 39 bp to 2.39 Mb, and the average size was 3180 bp (Table S1). The distribution of SVs across the genome was shown in Figure 2. After merging using SURVIVOR, the total quantity of SVs was 10,439 (114 duplications and 10,325 deletions) (Table 2), covering 53 Mb of the whole genome (Table S1).

Table 1. Yield and alignment coverage statistics for the cattle lung sample across various sequencing platforms.

Platform	$10 \times$ G	PromethION	PacBio CLR	PacBio CCS
Number of reads	1,577,259,728	1,618,623	11,178,388	2,875,796
Mapped reads	1,532,221,733	1,488,641	11,178,388	2,875,796
Mapping rate (%)	97.14	91.97	100	100
Depth	$55 \times$	$11 \times$	$40 \times$	$6 \times$
Read min length	19	70	53	74
Read max length	150	248,333	369,285	47,915
Read mean length	133.94	28,191.59	25,259.03	8763.78

Table 2. Statistics over SVs identified by various methods.

Platform	Method	DEL	DUP	Total
$10 \times$ G	LongRanger	8242	73	8315
	LinkedSV	6415	38	6453
	Merge	10,325	114	10,439
ONT	PBSV	26,397	2888	29,285
	Sniffles	3497	168	3665
	Merge	13,472	1881	15,353
PB_CLR	PBSV	885	169	1054
	Sniffles	1340	1238	2578
	Merge	1800	1162	2962
PB_CCS	PBSV	23,353	6569	29,922
	Sniffles	190	99	289
	Merge	15,601	3891	19,492
Merge	SURVIVOR	16,289	4875	21,164



Figure 2. Individualized cattle SV map. The tracks under every black bar represent the SVs for 10 × G_LongRanger, 10 × G_LinkedSV, CCS_PBSV, CCS_Sniffles, CLR_PBSV, CLR_Sniffles, ONT_PBSV and ONT_Sniffles (in order from top to bottom). Red means deletion, and green means duplication.

Oxford nanopore sequencing generated 1,618,623 sequences representing approximately $11\times$ coverage of the genome (Table 1). The distribution of sequence lengths (70–248,333 bp) was shown in Figure S1a, with an average length of 28,191.59 bp (Table 1). A total of 91.97% (Table 1) of the reads were mapped to the cattle genome assembly. Sniffles and PBSV identified 3665 and 29,285 SVs, respectively (Table 2). The identified SVs ranged from 32 bp to 2.62 Mb (mean size = 5676 bp) and 9 bp to 0.1 Mb (mean size = 592 bp) (Table S1), and their distribution across the whole genome was shown in Figure 2. The merging total number of SVs was 15,353 (1881 duplications and 13,472 deletions) (Table 2), covering 34 Mb of the whole genome (Table S1).

PacBio CLR sequencing yielded a total of 11,178,388 reads, representing 40-fold genome coverage, and they distributed in length between 53 and 369,285 bp (Figure S1b), with an average of 25,259.03 bp (Table 1). All reads were mapped to the cattle reference genome by pbmm2 (Table 1). Sniffles and PBSV identified 2578 and 1054 SVs, respectively (Table 2). The SV sizes identified by Sniffles ranged from 35 bp to 2.62 Mb, with an average size of 36,485 bp (Figure 2 and Table S1). For PBSV, the sizes ranged from 14 bp to 96 kb, and the mean size was 2377 bp (Figure 2 and Table S1). A total 2962 (1162 duplications and 1800 deletions) events covering 92 Mb (Table S1) of the whole genome were identified after merging (Table 2).

PacBio CCS sequencing generated 2,875,796 reads, representing $6\times$ coverage of the genome. The distribution of sequence length (74–47,915 bp) is illustrated in Figure S1c, with an average length of 8763.78 bp (Table 1). All reads were mapped to the cattle reference genome by pbmm2 (Table 1). Sniffles and PBSV identified 289 and 29,922 putative SVs, respectively (Table 2). The SV sizes identified by Sniffles ranged from 34 bp to 3.6 Mb and had a mean size of 72,166 bp (Figure 2, Table S1). For PBSV, the sizes ranged from 8 bp to 100 kb, and the mean size was 722 bp (Figure 2 and Table S1). The total merging number of SVs was 19,492 (3891 duplications and 15,601 deletions) (Table 2), covering 41 Mb of the whole genome (Table S1).

3.2. SV Overlap

In general, the total amount of SVs derived from short reads is much smaller than for the long-read-based inferences (Table 2). Most of the SVs were located between 50 bp to 200 bp, but long-read-based inferences can detect more large SVs (Figure 3a). Overall, these results show that across SVs accounts and sizes, long-read-based SV inference outperforms that of short reads. Between 45% and 60% of variants were called in at least one of the long-read data types, both of which were supported (Figure 3b). SVs called using PacBio CCS and ONT data showed high concordance (Figure 3b). The highest overlap with the results obtained from the short-read data was the PacBio CCS.

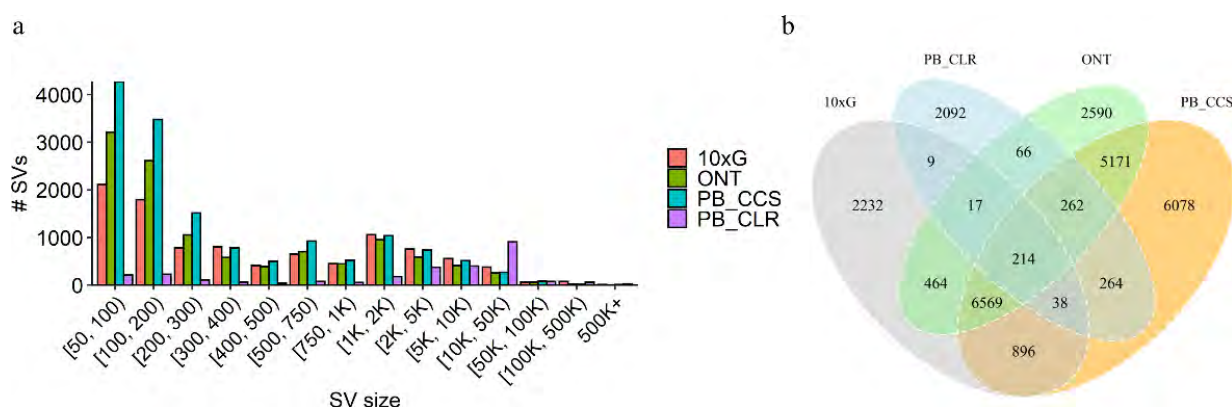


Figure 3. (a) Size distribution for SVs inferred from either long reads or Illumina/10 × G short reads. (b) Comparison between the four SV datasets.

4. Discussion

The long reads generated by the third-generation sequencing technology can span tens of thousands of base pairs, which are tremendously serviceable in filling gaps in current references [28,29] and for the assembly of complicated genomic regions [29,30]. Meanwhile, they can also be helpful for the identification of large SVs. In this study, we presented a comparison of four sequencing datasets from the same cattle lung DNA sample. We sequenced the genome with Illumina/10 × G, ONT, and PacBio (CLR and CCS) sequencing technologies and subsequently analyzed for structural variations. We observed comparisons between various SV methods and how SV results differ for different sequencing technologies.

We identified a total of 21,164 SVs, which amount to 186 Mb covering 7.07% of the whole genome (Table 2). In general, except for PB CLRs, the number of SVs inferred from long-read-based inferences was greater than that of short-reads (Table 2). The CLR detected the least number of SVs, probably due to insufficient coverage, but it identified the most of large SVs and covered the most genomes (Figure 3a). When using 10× linked reads, we obtained 10,439 SVs, but there were 8207 SVs shared between short- and long-read technologies (Figure 3a). We showed that SVs called with PacBio CCS versus ONT data show high concordance, with more than 90% of SVs called with one platform also being called with the other (Figure 3a), which is consistent with human results [22]. Our results indicated a concordance between SVs inferred with ONT and PacBio CCS.

With the advancement of long-read sequencing, the higher quality of the reference assembly could further benefit the identification of SVs. Leonard et al. showed that 20× for HiFi or 60× for ONT sequencing was sufficient to produce two haplotype-resolved assemblies while retaining over 90% accuracy in detecting SVs when integrated into pangenomes [31]. With a combination of PacBio HiFi, Hi-C, and ONT ultra-long read sequencing, we could soon routinely obtain a Telomere-to-Telomere (T2T) assembly for livestock, as recently demonstrated for humans [32].

5. Conclusions

In this study, we generated four sequencing datasets and compared the SV results based on them. For each dataset, we identified SVs using two programs. Our results indicated a concordance between SVs inferred with ONT and PacBio CCS. The one with the most overlap with the results obtained by short-read data is PB CCS. Together, we found that long reads performed better than short reads in terms of SV detections.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/genes13050828/s1>. Figure S1: Length distribution for reads from ONT and PacBio sequencing runs. Table S1. Summary of identified SVs using different methods.

Author Contributions: Conceptualization, Y.G. and G.E.L.; methodology, Y.G.; software, Y.G.; formal analysis, Y.G.; resources, L.M.; writing—original draft preparation, Y.G.; writing—review and editing, G.E.L.; supervision, G.E.L.; funding acquisition, L.M. and G.E.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by USDA National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) grant numbers 2016-67015-24886, 2019-67015-29321, and 2021-67015-33409.

Institutional Review Board Statement: The animal study protocol was approved by the US Department of Agriculture, Agricultural Research Service, Beltsville Agricultural Research Center's Institutional Animal Care and Use Committee (Protocol 16-016).

Informed Consent Statement: Not applicable.

Data Availability Statement: The SV calls reported in this article are available in the Supplemental Material, and sequencing data are available upon request for research purposes.

Acknowledgments: We thank Reuben Anderson, Mary Bowman, Donald Carbaugh, Christina Clover, Sarah McQueeney, Mary Niland, Derek Bickhart, Tim Smith, and Research Animal Services for the technical assistance. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture (USDA). The USDA is an equal opportunity provider and employer.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Chaisson, M.; Sanders, A.D.; Zhao, X.; Malhotra, A.; Porubsky, D.; Rausch, T.; Gardner, E.J.; Rodriguez, O.L.; Guo, L.; Collins, R.L.; et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **2019**, *10*, 1784. [[CrossRef](#)] [[PubMed](#)]
2. Chiang, C.; Scott, A.J.; Davis, J.R.; Tsang, E.K.; Li, X.; Kim, Y.; Hadzic, T.; Damani, F.N.; Ganel, L.; GTEx Consortium; et al. The impact of structural variation on human gene expression. *Nat. Genet.* **2017**, *49*, 692–699. [[CrossRef](#)] [[PubMed](#)]
3. Karaoglanoglu, F.; Ricketts, C.; Ebre, E.; Rasekh, M.E.; Hajirasouliha, I.; Alkan, C. VALOR2: Characterization of large-scale structural variants using linked-reads. *Genome Biol.* **2020**, *21*, 72. [[CrossRef](#)] [[PubMed](#)]
4. Marks, P.; Garcia, S.; Barrio, A.M.; Belhocine, K.; Bernate, J.; Bharadwaj, R.; Bjornson, K.; Catalanotti, C.; Delaney, J.; Fehr, A.; et al. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **2019**, *29*, 635–645. [[CrossRef](#)] [[PubMed](#)]
5. Spies, N.; Weng, Z.; Bishara, A.; McDaniel, J.; Catoe, D.; Zook, J.M.; Salit, M.; West, R.B.; Batzoglou, S.; Sidow, A. Genome-wide reconstruction of complex structural variants using read clouds. *Nat. Methods* **2017**, *14*, 915–920. [[CrossRef](#)] [[PubMed](#)]
6. Weisenfeld, N.I.; Kumar, V.; Shah, P.; Church, D.M.; Jaffe, D.B. Direct determination of diploid genome sequences. *Genome Res.* **2017**, *27*, 757–767. [[CrossRef](#)] [[PubMed](#)]
7. Cretu Stancu, M.; van Roosmalen, M.J.; Renkens, I.; Nieboer, M.M.; Middelkamp, S.; de Ligt, J.; Pregno, G.; Giachino, D.; Mandrile, G.; Espejo Valle-Inclan, J.; et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **2017**, *8*, 1326. [[CrossRef](#)]
8. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **2018**, *15*, 461–468. [[CrossRef](#)]
9. Chaisson, M.J.; Huddleston, J.; Dennis, M.Y.; Sudmant, P.H.; Malig, M.; Hormozdiari, F.; Antonacci, F.; Surti, U.; Sandstrom, R.; Boitano, M.; et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **2015**, *517*, 608–611. [[CrossRef](#)]
10. Chin, C.S.; Peluso, P.; Sedlazeck, F.J.; Nattestad, M.; Concepcion, G.T.; Clum, A.; Dunn, C.; O'Malley, R.; Figueroa-Balderas, R.; Morales-Cruz, A.; et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **2016**, *13*, 1050–1054. [[CrossRef](#)]
11. Koren, S.; Rhie, A.; Walenz, B.P.; Dilthey, A.T.; Bickhart, D.M.; Kingan, S.B.; Hiendleder, S.; Williams, J.L.; Smith, T.; Phillippy, A.M. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **2018**, *36*, 174–182. [[CrossRef](#)] [[PubMed](#)]
12. Ebert, P.; Audano, P.A.; Zhu, Q.; Rodriguez-Martin, B.; Porubsky, D.; Bonder, M.J.; Sulovari, A.; Ebler, J.; Zhou, W.; Serra Mari, R.; et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **2021**, *372*, eabf7117. [[CrossRef](#)] [[PubMed](#)]
13. Quan, C.; Li, Y.; Liu, X.; Wang, Y.; Ping, J.; Lu, Y.; Zhou, G. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. *Genome Biol.* **2021**, *22*, 159. [[CrossRef](#)]
14. Low, W.Y.; Tearle, R.; Liu, R.; Koren, S.; Rhie, A.; Bickhart, D.M.; Rosen, B.D.; Kronenberg, Z.N.; Kingan, S.B.; Tseng, E.; et al. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat. Commun.* **2020**, *11*, 2071. [[CrossRef](#)]
15. Couldrey, C.; Keehan, M.; Johnson, T.; Tiplady, K.; Winkelman, A.; Littlejohn, M.D.; Scott, A.; Kemper, K.E.; Hayes, B.; Davis, S.R.; et al. Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. *J. Dairy Sci.* **2017**, *100*, 5472–5478. [[CrossRef](#)] [[PubMed](#)]
16. Lamb, H.J.; Ross, E.M.; Nguyen, L.T.; Lyons, R.E.; Moore, S.S.; Hayes, B.J. Characterization of the poll allele in Brahman cattle using long-read Oxford Nanopore sequencing. *J. Anim. Sci.* **2020**, *98*, skaa127. [[CrossRef](#)]
17. Ananthasayanam, S.; Kothandaraman, H.; Nayee, N.; Saha, S.; Baghel, D.S.; Gopalakrishnan, K.; Peddamma, S.; Singh, R.B.; Schatz, M. First near complete haplotype phased genome assembly of River buffalo (*Bubalus bubalis*). *bioRxiv* **2020**, 618785. [[CrossRef](#)]
18. Zhou, R.; Li, S.T.; Yao, W.Y.; Xie, C.D.; Chen, Z.; Zeng, Z.J.; Wang, D.; Xu, K.; Shen, Z.J.; Mu, Y.; et al. The Meishan pig genome reveals structural variation-mediated gene expression and phenotypic divergence underlying Asian pig domestication. *Mol. Ecol. Resour.* **2021**, *21*, 2077–2092. [[CrossRef](#)] [[PubMed](#)]
19. Ma, H.; Jiang, J.; He, J.; Liu, H.; Han, L.; Gong, Y.; Li, B.; Yu, Z.; Tang, S.; Zhang, Y.; et al. Long-read assembly of the Chinese indigenous Ningxiang pig genome and identification of genetic variations in fat metabolism among different breeds. *Mol. Ecol. Resour.* **2022**, *22*, 1508–1520. [[CrossRef](#)]

20. Li, R.; Gong, M.; Zhang, X.; Wang, F.; Liu, Z.; Zhang, L.; Xu, M.; Zhang, Y.; Dai, X.; Zhang, Z.; et al. The first sheep graph-based pan-genome 1 reveals the spectrum of structural variations and their effects on tail phenotypes. *bioRxiv* **2021**, 472709. [[CrossRef](#)]
21. Bickhart, D.M.; Rosen, B.D.; Koren, S.; Sayre, B.L.; Hastie, A.R.; Chan, S.; Lee, J.; Lam, E.T.; Liachko, I.; Sullivan, S.T.; et al. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **2017**, *49*, 643–650. [[CrossRef](#)] [[PubMed](#)]
22. Aganezov, S.; Goodwin, S.; Sherman, R.M.; Sedlazeck, F.J.; Arun, G.; Bhatia, S.; Lee, I.; Kirsche, M.; Wappel, R.; Kramer, M.; et al. Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res.* **2020**, *30*, 1258–1273. [[CrossRef](#)] [[PubMed](#)]
23. Zheng, G.X.; Lau, B.T.; Schnall-Levin, M.; Jarosz, M.; Bell, J.M.; Hindson, C.M.; Kyriazopoulou-Panagiotopoulou, S.; Masquelier, D.A.; Merrill, L.; Terry, J.M.; et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **2016**, *34*, 303–311. [[CrossRef](#)] [[PubMed](#)]
24. Fang, L.; Kao, C.; Gonzalez, M.V.; Mafra, F.A.; Pellegrino da Silva, R.; Li, M.; Wenzel, S.S.; Wimmer, K.; Hakonarson, H.; Wang, K. LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. *Nat. Commun.* **2019**, *10*, 5585. [[CrossRef](#)]
25. Rosen, B.D.; Bickhart, D.M.; Schnabel, R.D.; Koren, S.; Elsik, C.G.; Tseng, E.; Rowan, T.N.; Low, W.Y.; Zimin, A.; Couldrey, C.; et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **2020**, *9*, giaa021. [[CrossRef](#)]
26. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)]
27. Jeffares, D.C.; Jolly, C.; Hoti, M.; Speed, D.; Shaw, L.; Rallis, C.; Balloux, F.; Dessimoz, C.; Bähler, J.; Sedlazeck, F.J. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **2017**, *8*, 14061. [[CrossRef](#)]
28. English, A.C.; Richards, S.; Han, Y.; Wang, M.; Vee, V.; Qu, J.; Qin, X.; Muzny, D.M.; Reid, J.G.; Worley, K.C.; et al. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **2012**, *7*, e47768. [[CrossRef](#)]
29. Rhoads, A.; Au, K.F. PacBio Sequencing and Its Applications. *Genom. Proteom. Bioinform.* **2015**, *13*, 278–289. [[CrossRef](#)]
30. English, A.C.; Salerno, W.J.; Hampton, O.A.; Gonzaga-Jauregui, C.; Ambreth, S.; Ritter, D.I.; Beck, C.R.; Davis, C.F.; Dahdouli, M.; Ma, S.; et al. Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC Genom.* **2015**, *16*, 286. [[CrossRef](#)]
31. Leonard, A.S.; Crysanto, D.; Fang, Z.H.; Heaton, M.P.; Ley, B.L.V.; Herrera, C.; Bollwein, H.; Bickhart, D.M.; Kuhn, K.L.; Smith, T.P.L.; et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. *bioRxiv* **2021**, 466900. [[CrossRef](#)]
32. Nurk, S.; Koren, S.; Rhie, A.; Rautiainen, M.; Bzikadze, A.V.; Mikheenko, A.; Vollger, M.R.; Altemose, N.; Uralsky, L.; Gershman, A.; et al. The complete sequence of a human genome. *Science* **2022**, *376*, 44–53. [[CrossRef](#)] [[PubMed](#)]

RESEARCH

Open Access



Single-cell transcriptomic and chromatin accessibility analyses of dairy cattle peripheral blood mononuclear cells and their responses to lipopolysaccharide

Yahui Gao^{1,2†}, Jianbin Li^{1*†}, Gaozhan Cai^{1,3†}, Yujiao Wang¹, Wenjing Yang⁴, Yanqin Li¹, Xiuxin Zhao³, Rongling Li¹, Yundong Gao¹, Wenbin Tuo⁵, Ransom L. Baldwin VI², Cong-jun Li^{2*}, Lingzhao Fang^{6*} and George E. Liu^{2*}

Abstract

Background: Gram-negative bacteria are important pathogens in cattle, causing severe infectious diseases, including mastitis. Lipopolysaccharides (LPS) are components of the outer membrane of Gram-negative bacteria and crucial mediators of chronic inflammation in cattle. LPS modulations of bovine immune responses have been studied before. However, the single-cell transcriptomic and chromatin accessibility analyses of bovine peripheral blood mononuclear cells (PBMCs) and their responses to LPS stimulation were never reported.

Results: We performed single-cell RNA sequencing (scRNA-seq) and single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) in bovine PBMCs before and after LPS treatment and demonstrated that seven major cell types, which included CD4 T cells, CD8 T cells, and B cells, monocytes, natural killer cells, innate lymphoid cells, and dendritic cells. Bioinformatic analyses indicated that LPS could increase PBMC cell cycle progression, cellular differentiation, and chromatin accessibility. Gene analyses further showed significant changes in differential expression, transcription factor binding site, gene ontology, and regulatory interactions during the PBMC responses to LPS. Consistent with the findings of previous studies, LPS induced activation of monocytes and dendritic cells, likely through their upregulated TLR4 receptor. NF-κB was observed to be activated by LPS and an increased transcription of an array of pro-inflammatory cytokines, in agreement that NF-κB is an LPS-responsive regulator of innate immune responses. In addition, by integrating LPS-induced differentially expressed genes (DEGs) with large-scale GWAS of 45 complex traits in Holstein, we detected trait-relevant cell types. We found that selected DEGs were significantly associated with immune-relevant health, milk production, and body conformation traits.

*Correspondence: msdljb@163.com; Congjun.Li@usda.gov; Lingzhao.fang@igmm.ed.ac.uk; George.Liu@usda.gov

†Yahui Gao, Jianbin Li and Gaozhan Cai contributed equally to this work.

¹ Institute of Animal Science and Veterinary Medicine, Shandong Academy of Agricultural Sciences, No.202, Gongyebei Road, Jinan 250100, China

² Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, MD 20705, USA

⁶ MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Conclusion: This study provided the first scRNA-seq and scATAC-seq data for cattle PBMCs and their responses to the LPS stimulation to the best of our knowledge. These results should also serve as valuable resources for the future study of the bovine immune system and open the door for discoveries about immune cell roles in complex traits like mastitis at single-cell resolution.

Keywords: Cattle, Peripheral blood mononuclear cell, Lipopolysaccharide, Single-cell RNA-seq, Single-cell ATAC-seq

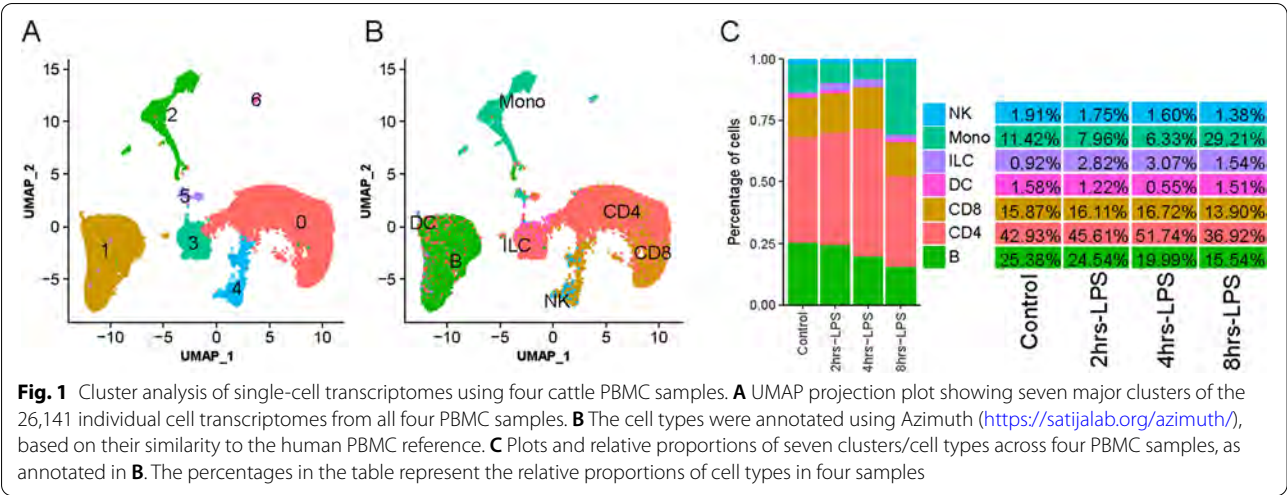
Introduction

Mastitis is the most severe economic and health problem associated with dairy cow herds, affecting milk yield, milk composition, and productive life. Gram-negative bacteria are one of the important pathogens in cattle causing severe diseases, including mastitis and digestive tract infections. Lipopolysaccharides (LPS), also known as endotoxins, are components of the outer membrane of Gram-negative bacteria and crucial mediators of chronic inflammation in cattle suffering from clinical and subclinical infections caused by the bacteria. LPS exposure can result in elevated levels of local or systemic inflammation, which could compromise animal wellbeing and productivity [1, 2]. In mammals, the innate immune system serves as the first line of defense involving sensing pathogen-associated molecular patterns (PAMPs) and launching innate immune responses against the infections. LPS, a PAMP of the Gram-negative bacteria, is a highly potent activator of the innate immune system, eliciting strong inflammatory responses in infected animals [3]. The cells of the innate immune system, including monocytes (Mono), dendritic cells (DC), and granulocytes, function as the first line of defense upon encounter of infectious agents. Phagocytic macrophages, cytotoxic natural killer (NK) cells, and $\gamma\delta$ T cells also play a crucial role in the innate immunity [4, 5]. Studies have been conducted to demonstrate the mechanisms by which LPS modulates the immune responses in vivo and in vitro. LPS can activate cellular responses by binding to the TLR4/CD14/MD2 receptor complex and activating pro-inflammatory transcription factors [6, 7]. Activated monocytes and DCs release nitric oxide, interleukin-1 (IL-1), IL-6, tumor necrosis factor-alpha (TNF α), and other factors [8]. Additionally, the innate immune cells such as monocytes and DC play a crucial role in bridging the innate and acquired immunities by responding to various PAMPs and serving as antigen-presenting cells (APCs) in the context of major histocompatibility complexes (MHC) [9]. The APCs must be adequately activated and conditioned upon their engagement with T cells, resulting in T cell activation in the presence of a cytokine and cell surface costimulatory molecule milieu, which is essential for the development of recall T cell responses required for host defense and

protection. Surface marker genes on many immune cell types, like B cells and T cells, have been extensively studied [10]. For example, based on the expression levels of CD14 and CD16, monocytes can be divided into two types in the human blood [11].

A bulk human RNA-seq study demonstrated that LPS-responsive genes could be characterized as two co-regulated programs, i.e., the “antiviral-like” program and “inflammatory-like” program, based on their expression profiles [12]. The antiviral program is mainly mediated by interferon regulatory factors (IRFs). In contrast, the inflammatory program is primarily mediated by the Nuclear factor kappa-light-chain-enhancer of activated B cells (NF- κ B) [12]. Single-cell-based analyses have been used to define human and mouse immune cells [13–15] and their responses to LPS. Additionally, single-cell RNA-seq studies further partitioned the inflammatory program genes into two modules, a peaked inflammatory module consisting of genes such as *TNF*, *IL1B*, and *CXCL2* that responded rapidly, yet transiently, when stimulated by LPS, and a sustained inflammatory module which included genes such as *Mmp14*, *Marco*, and *IL6*, exhibiting a continued rise in expression under LPS stimulation [13, 16].

The cell types and functions of cattle peripheral blood mononuclear cells (PBMCs) have been extensively studied [17–20]. In general, cattle PBMCs, similar to those of mammals, consist of primarily T and B cells, NK cells, monocytes, and DC [17, 20]. Cattle PBMC composition is unique in that young calves have higher levels of gamma/delta ($\gamma\delta$) T cell receptor (TCR) positive T cells in comparison to those of humans and mice [18]. However, large-scale single-cell analyses in cattle PBMCs have never been reported. There is a need to document the gene transcriptional, chromatin accessibility, and gene-based changes in PBMCs at the single-cell resolution before and after LPS stimulation. These studies will permit investigators to interrogate complex cellular regulations and interactions and delineate cell differentiation and lineage relationships within a sample of heterogeneous cell populations at the single-cell level. They will facilitate further understanding of LPS-mediated bovine PBMC responses and complement the existing methodologies determining PBMC cell types and functions. This is particularly important in cattle or



other livestock species. There is a general lack of critical immunological reagents for thorough profiling of cell phenotypes, activation status, and cytokine production. This study presents the first cattle single-cell PBMC profiling and their responses to LPS stimulation *in vitro*. The analyses of scRNA-seq data of the present study demonstrate robust clustering and assignment of naïve bovine PBMC populations and cell type-specific responses to LPS at the single-cell level. This study reports trait-relevant cell types and genes underlying complex traits by integrating LPS-induced DEGs with large-scale GWAS of 45 complex Holstein traits.

Results

Data generation and quality assessment

Using the 10× Genomics Chromium Controller [21], we performed scRNA-seq and scATAC-seq of Holstein PBMC samples treated without (Control) or with LPS for 2 h (2 h-LPS), 4 h (4 h-LPS), and 8 h (8 h-LPS). We sequenced a total of 30,756 single cells with approximately 62,254 reads per cell (Table S1). After quality filtering and integration, we obtained 26,141 single cells, corresponding to a median of 4,581 unique molecular identifiers per cell and ~15,000 total genes in the whole population. Overall, we obtained 7,107 (Control), 9,174 (2 h-LPS), 6,741 (4 h-LPS), and 3,119 (8 h-LPS) cells.

Cell clustering and cell type assignment

Using Seurat v3.2 [22], we performed a graph-based clustering on cells according to the gene expression profiles. After visualizing the Uniform Manifold Approximation and Projection (UMAP) plots, we found that the single-cell transcriptomes of the four samples analyzed were similar (Figure S1A and 1B), indicating a high degree of reproducibility among them. We obtained a total of 7

distinct clusters designated by Cluster (C)0, C1, C2, C3, C4, C5, and C6 (Fig. 1A). We utilized canonical marker genes of immune cells derived from published literature and the online database PanglaoDB Field [23] to assign cell types. Based on gene expression patterns, we generated violin plots (Figure S2A) and UMAP projections (Figure S2B) for each gene. We then assigned immune cell types in cattle PBMC samples based on the combined unique patterns of these cell marker gene expressions as shown in parentheses (Figure S2A, Table S2). For example, CD4 T cells (*CD4*, *CD5*, and *LEF1*) and CD8 T cells (*CD8A* and *CD8B*) in C0, B cells (*MS4A1*, *CD79A*, *CD79B*, and *VPREB3*) in C1, monocytes (*CD14*, *S100A12*, *ADGRE1*, *MEFV*, and *HCK*) in C2 and C5, innate lymphoid cells (ILCs) (*SLC4A4*, *PLIN3*, and *COL5A1*) in C3, NK cells (*GNLY*, *NKG7*, *CTSW*, *PRF1*, and *IL2RB*) in C4, and DCs (*IRF8* and *CD83*) in C1 and C6. Of note, some combinations of these top marker genes were uniquely expressed in only one cell type, such as *CD14* and *S100A12* (*S100* calcium-binding protein A12) in monocytes, whereas *IRF8* (interferon regulatory factor 8) and *CD83* (nuclear receptor subfamily 4, group A, member 3) in DCs. However, some known marker genes were not detected, such as *FCER1A*, which is considered a gene marker for cDC2. *FCER1G*, a related gene coding for the gamma chain of the high-affinity receptor for the Fc fragment of IgE (FCER), was detected as a DEG in all cell types except for CD8 T cells (Table S2). We further confirmed the above cell type assignments using two other methods: Azimuth (Fig. 3A) and SingleR (Figure S3B). With Azimuth [24], we generated cell-type annotation results at three resolutions: low, medium, and high (Figure S3A, Table S3). As shown in Figure S3A, we detected Treg, TEM, and TCM cells, as well as naïve and memory B cells. Additionally, we assigned cell types using

SingleR [25] and the human cell reference datasets, Blueprint and Encode (Table S4). By combining all three cell assignment efforts, consistent assignment results were demonstrated across three different methods (Tables S2, S3, and S4), where the main cell types were CD4 T cells and CD8 T cells for C0, B cells, and DCs for C1, monocytes for C2 and C5, and NK cells and CD8 T cells for C4 (Fig. 1B). We also successfully assigned CD14 monocytes and CD16 monocytes in C2 using SingleR (Figures S2A) or Azimuth (Figures S3A), separately.

Cross-species comparison

To verify our cell clustering and assignments, we compared results between the cattle and human PBMCs. We downloaded the scRNA-seq dataset of the human PBMC from the GSE96583 [14, 15] and performed a joint Seurat clustering analysis with our Control cattle PBMC sample [22]. Plotting the single-cell transcriptomes via UMAP projection yielded largely overlapping distributions of cells from cattle and human samples (Figures S4A, B, and C), validating our scRNA-seq data generation, processing, clustering, and cell type assignment. With Azimuth [24], we obtained 13,601 individual cell transcriptomes of eight-cell types from the two samples (Figure S5A and B). The UMAP plot distribution reflected that the main cell types were CD4 T cells, CD8 T cells, B cells, NK cells, monocytes, DCs, and other minor populations (Figure S5A), confirming the seven cell types identified in our cattle Control sample. We also calculated the correlation between paired clusters of humans and cattle based on the top 2000 variable gene expressions. We showed the correlations were higher than 0.4 between humans and cattle, indicating a high similarity of these two species (Figure S4D). In summary, the analysis produced seven major cell types and their corresponding subtypes: CD4 T cells (CD4 Naïve, CD4 TCM, CD4 TEM, and Treg), CD8 T cells (CD8 Naïve, CD8 TCM, and CD8 TEM), B cells (B intermediate, B memory, B naïve, and plasmablast), monocytes (CD14 Mono and CD16 Mono), NK cells, ILCs, and DCs (cDC and pDC) (Fig. 1B). We will focus on these seven cell types for the subsequent sections unless specified otherwise.

Cell cycle analysis for PBMC

We performed the cell cycle analyses to calculate their cell cycle indices (i.e., the ratio of actively proliferating cells of each feature, such as different samples and different developmental stages) and explore cell proliferation status, using sets of 43 G1/S and 55 G2/M genes (Table S5). The expression profiles of cell cycle-related genes revealed that the cell cycle indices were 50.63%, 45.61%, 60.48%, 22.19%, 38.18%, 37.13%, and 36.30% for CD4 T cells, CD8 T cells, B cells, monocytes, NK cells, ILCs,

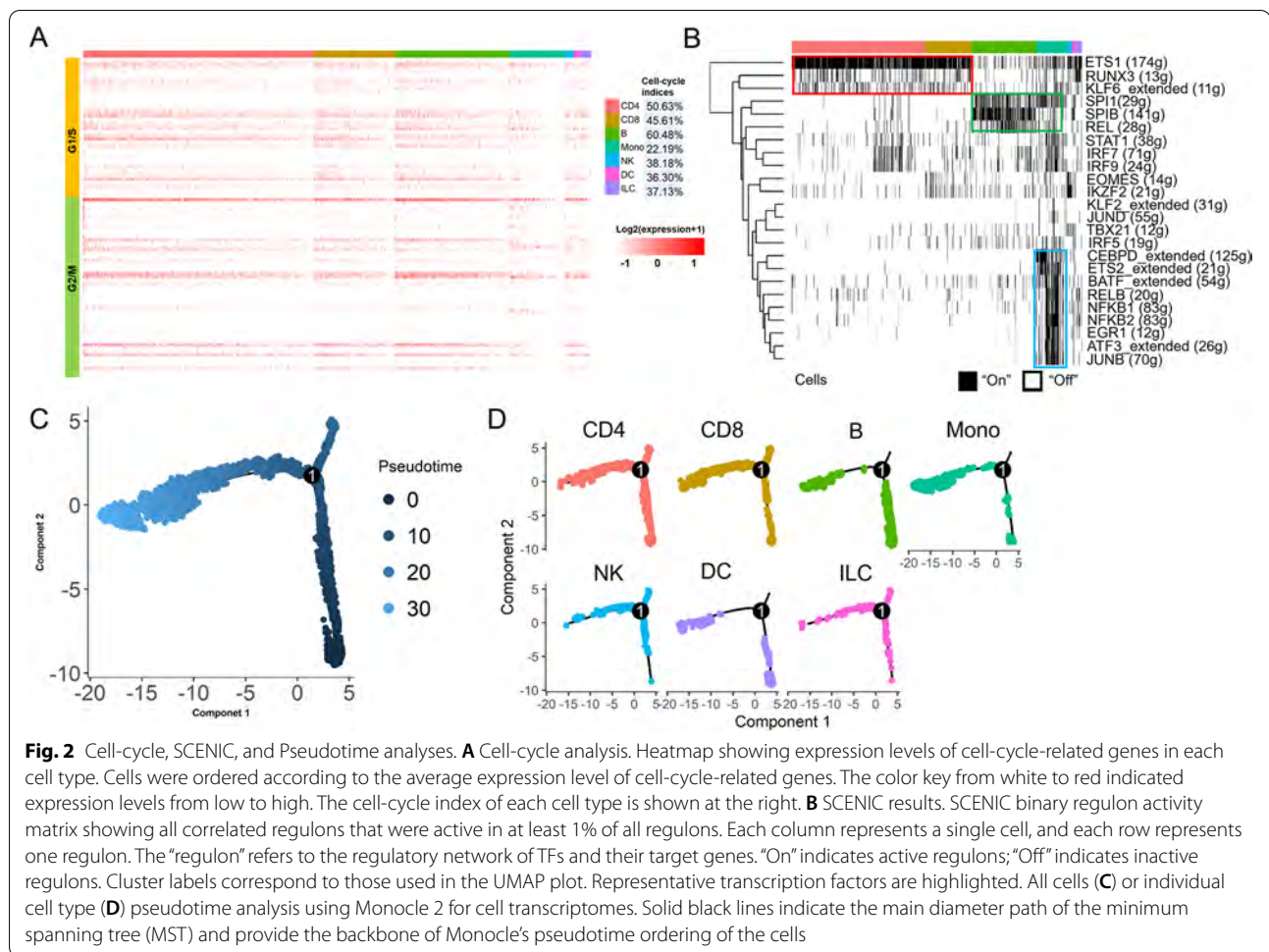
and DCs, respectively (Fig. 2A). Over LPS treatment time points, we found that monocyte cell cycle indices were 3.61%, 2.30%, 1.91%, and 31.88% in Control, 2 h-LPS, 4 h-LPS, and 8 h-LPS, respectively (Figure S6A). The cell cycle indices revealed that monocyte cell cycle progression was upregulated, suggesting that monocyte proliferation was dramatically activated during the early LPS treatment.

Transcription factor analysis for PBMCs

To understand LPS-induced transcriptional activities of PBMC transcription factors (TF), we performed a transcription factor analysis using SCENIC [26] to identify regulators and gene regulatory networks. Through this analysis, we identified 24 active regulons in cattle PBMCs (Fig. 2B). Most of the regulons are related to immune functions in the differentiation and proliferation of T cells and B cells (ETS1, RUNX3, KLF6, SPI1, SPIB) or involved in mediating immune and inflammatory responses (REL, STAT1, IRF7, IRF9, EOMES, IKZF2, KLF2). The count range of target genes of these regulons was between 11 and 174 (Table S6). SCENIC analysis revealed several critical transcriptional regulators modulating cell type-specific gene regulatory networks. For all PBMCs, especially in CD4 T cells, CD8 T cells, B cells (to a lesser extent), monocytes, NK cells, DCs, and ILCs, we identified several universal TFs like ETS1, RUNX3, and KLF6_extended, as shown in Fig. 2B (red rectangle). We detected PU.1/SPI1, SPIB, and REL, primarily in B cells and monocytes (Fig. 2B, green rectangle). In CD4 T cells, CD8 T cells, B cells, monocytes, NK cells, and DCs, specific TFs, such as IRF5, IRF7, IRF9, and STAT1, were identified. For monocytes, we further identified their specific TFs, including CEBPD_extended, ETS2_extended, BATF_extended, IKZF2, NFKB1, NFKB2, and RELB, EGR1, ATF3_extended, and JUNB (Fig. 2B, blue rectangle). Therefore, TFs, as essential regulators of gene expression, are also marker genes for identifying cell types.

Pseudotime analysis

To understand the developmental states of monocytes and DCs, we conducted a pseudotime analysis to infer cell trajectories using Monocle 2 [27]. Following a “developmental/transitional” path according to their transcriptomic similarity, we identified one significant and long-trajectory branch, with which cells are ordered in an arrangement from proximal to distal distribution (Fig. 2C). Combining with the pseudotime values (Table S7), we observed that the long-trajectory tree rooted from the bottom right to the top left, covering CD4 T cells, CD8 T cells, B cells, NK cells, ILCs, monocytes, and DCs (Fig. 2D). Larger portions of monocytes and

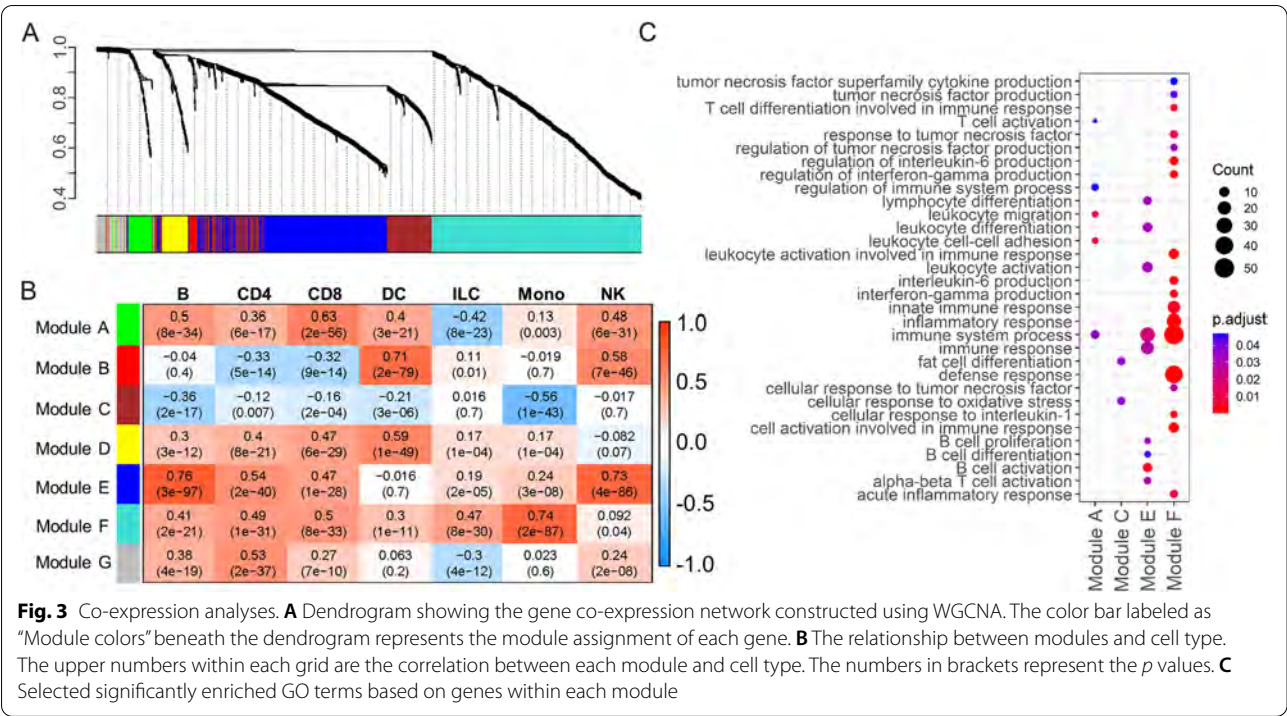


DCs were observed at the top left end of the trajectory. The path also appeared to agree with our Seurat cluster results, i.e., monocytes in C2 and C5, while DCs in C6. Thus, those monocytes and DCs with the highest pseudotime scores might represent their terminal developmental states.

Co-expression analyses

To systematically investigate the genetic program dynamics, we performed a weighted gene co-expression network analysis (WGCNA) [28] using the top 2,000 marker genes reported by Seurat. Seven gene modules were identified by WGCNA (Fig. 3A), each containing gene sets that tend to be co-expressed (Table S8). To assign co-expressed gene functions to cell types, we calculated the correlation between each module (module eigengene) and each cell type (UMI) and generated a correlation heatmap in Fig. 3B. We then performed GO analyses for genes in each module to investigate their biological functions (Fig. 3C, Table S9). For example, Module E genes (blue) were enriched for immune responses, lymphocyte

activation, differentiation, proliferation, and migration, especially with B cells and alpha-beta TCR T cells. Module E was also more correlated with B cells, NK cells, CD4 cells, and CD8 T cells. Module A genes (green) were enriched for the G protein-coupled receptor signaling pathway, kinase regulator activity, chemokine-mediated signaling pathway, regulation of chemotaxis, leukocyte adhesion and migration, regulation of cell death, calcium ion transport, and T cell activation. Module A was more correlated with CD8 T cells, B cells, NK cells, and CD4 cells. Module F genes (turquoise) were enriched for multiple GO terms, including (1) cellular response to LPS, LPS-mediated signaling pathways, innate immune responses, regulation of adaptive immune responses, leukocyte differentiation and adhesion, regulation of CD4+alpha-beta TCR T cell activation, T-helper cell differentiation, macrophage migration, positive regulation of cytokine production, regulation of cell death; (2) positive regulation of interferon- γ production, positive regulation of interleukin-6 production, regulation of interleukin-1 β production and response, cellular



response to fibroblast growth factor (FGF) stimulus, p38 mitogen-activated protein kinases (MAPK) cascade, and tumor necrosis factor (TNF) superfamily cytokine production; and (3) cell chemotaxis, response to reactive oxygen species, positive regulation of endopeptidase activity, response to glucocorticoid, collagen metabolic process, regulation of translation and transcription, lysosome, and osteoclast differentiation. Module F was mainly correlated with monocytes, followed by CD8 T cells, CD4 T cells, ILCs, and B cells (Fig. 3C, Table S9). Therefore, our co-expression analyses identified critical gene sets corresponding to cell type-differential functions.

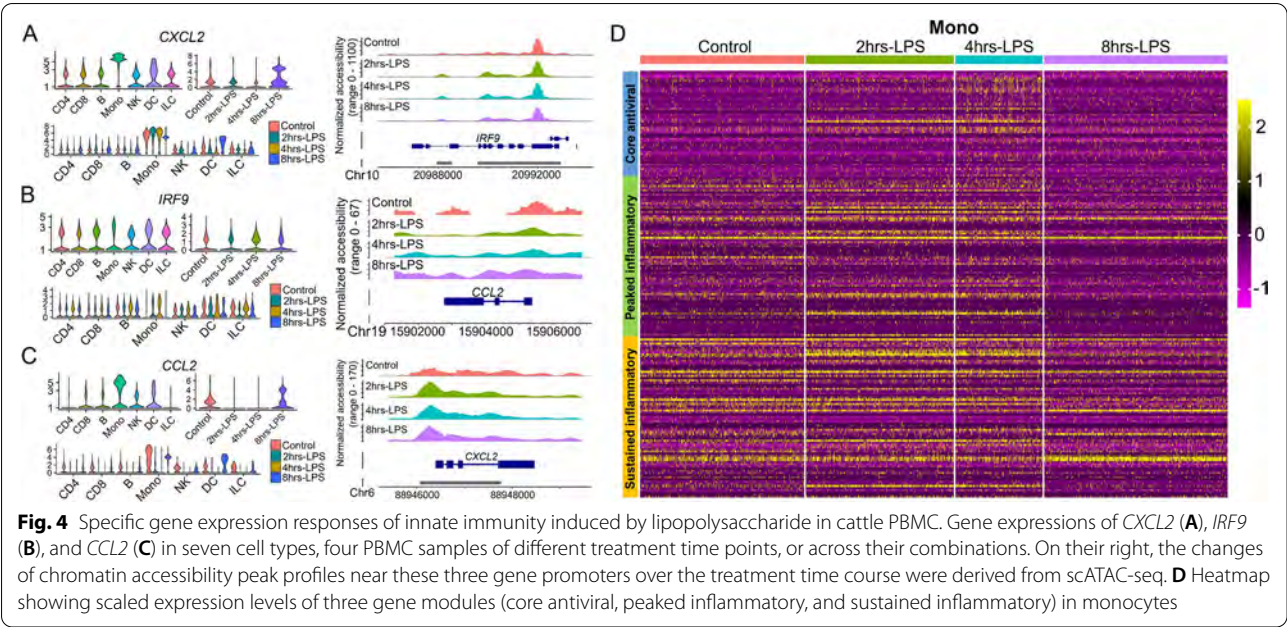
Marker gene expression for PBMC clusters

Marker gene expression analysis was aimed to determine the expressions of essential known marker genes and their nearby chromatin accessibilities in several cell types. Based on the Seurat results, we obtained distinct sets of marker genes among these cell types (Table S2). For example, *CXCL2* (C-X-C motif ligand 2) expression was higher in monocytes than others (Fig. 4A). When we analyzed cell type-specific responses over time, we found that *CXCL2* expression was higher in monocytes than other cell types; its expression was elevated in Control, 2 h-LPS, and 4 h-LPS samples decreased in 8 h-LPS. Correspondingly, we also detected increased levels of chromatin accessibility in the *CXCL2* promoter (Fig. 4A). A similar pattern was also found for *CXCL5* (Figure S7B).

When we plotted individual or combined marker gene expression over time, *IRF9* was expressed higher in DCs than other cell types (ANOVA test, $p < 2 \times 10^{-16}$). However, due to the small cell count of DCs, we did not detect significant differences in gene expression or chromatin openness over time points (Fig. 4B). *CCL2* (C-C motif ligand 2, encoded by the negative-sense strand) expression was higher in Control and 8 h-LPS than 2 h-LPS and 4 h-LPS, which were in line with higher chromatin accessibility in Control and 8 h-LPS (Fig. 4C). Also, in monocytes, we detected *IL1B* expression, which was decreased from early (2 h-LPS, 4 h-LPS) to late time points (8 h-LPS), while in DCs and ILCs, *IL1B* expression was increased (Figure S7A). Hence, we found a consistent correlation between expression and chromatin accessibility for selected marker genes.

Gene expression patterns during LPS treatment

In humans, Shalek et al. [13] used the single-cell gene expression profiles to partition the LPS-responsive genes into two programs: the antiviral programs and the inflammatory programs, which include three modules: the core antiviral module (enriched for annotated antiviral and interferon response genes), the peaked inflammatory module and the sustained inflammatory module. We obtained these three human LPS-responsive gene lists and plotted the expression patterns of the bovine ortholog genes from monocytes with or without LPS treatment (Fig. 4D). The analysis showed that the gene



expression sustained until four hours post LPS treatment for the sustained inflammatory module and then decreased slightly at eight h. But for the core antiviral module and the peaked inflammatory module, gene expression was increased from Control to 4 h-LPS and fell to Control levels in 8 h-LPS. These results were consistent with the observation in human PBMCs that the antiviral and inflammation responses mainly occurred early but decreased in the late-stage [13]. Therefore, we observed similar gene expression patterns for those three modules in cattle and humans during LPS treatment.

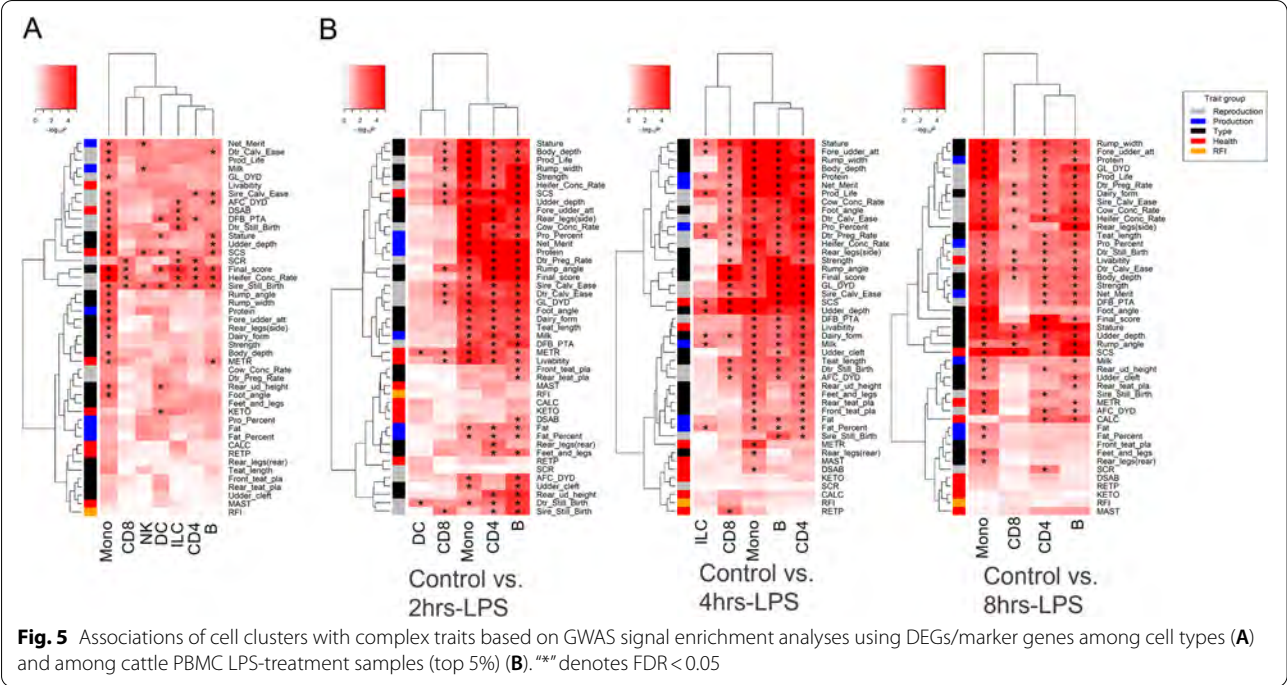
Trait-relevant cell clusters

Using edgeR [29], we detected thousands of marker genes among seven cell types (Table S10-S16). Using a permutation-based marker-set test approach (Methods), we tested the enrichment of 45 GWAS signals within these marker genes of distinct cell types ($FDR < 0.05$) (Fig. 5A). Reproduction traits were significantly associated with all cell types, reflecting the potential functions of these cell types related to fertility and tissue development. Since all cell types in the present study were immune cells, their high correlation with reproduction traits confirmed our previous findings [30]. Additionally, health traits, such as SCS (somatic cell score, an indicator of mastitis), were associated with most cell types, confirming that these cell types have a role in immunity and tissue integrity. Body conformation traits were also significantly associated with monocytes.

Moreover, based on the marker genes reported by edgeR between cell clusters across the LPS-untreated (Control) and LPS-treated (2 h-LPS, 4 h-LPS, and 8 h-LPS) PBMC samples, we also detected similar results (Fig. 5B). In all three comparisons, we found that the cell types with the most DEGs were monocytes, CD4 T cells, and B cells. Generally, all cell types were significantly associated with reproduction, body conformation, and health traits. In both Control vs. 2 h-LPS and Control vs. 4 h-LPS comparisons, monocytes were associated with health traits, especially immune traits, such as SCS and Livability, but not with the health traits relating to metabolic diseases.

Discussion

In the current cattle single-cell analyses, we successfully detected and confirmed seven major cell types (including CD4 T cells, CD8 T cells, B cells, monocytes, NK cells, ILCs, and DCs), as well as their responses to LPS challenge in vitro using scRNA-seq and scATAC-seq. We characterized these cells and their genes in detail. Our bioinformatic analyses indicated that LPS could increase PBMC cell cycle progression, cellular differentiation, and chromatin accessibility. Our gene analyses further showed significant changes in differential expression, transcription factor binding site, gene ontology, and regulatory interactions during the PBMC responses to LPS. These results of cattle PBMC generally agreed with the existing human and cattle studies [2, 13, 16]. The reactions to LPS treatment include innate immunity activation of monocytes and dendritic cells, featuring the



antiviral program mediated by interferon regulatory factors (IRFs) and the inflammatory program mediated by NF- κ B and pro-inflammatory cytokines such as CCL2 and CXCL2. LPS induced activation of monocytes and dendritic cells, likely through their upregulated TLR4 receptor. NF- κ B was observed to be activated by LPS and increased transcriptions of an array of pro-inflammatory cytokines, in agreement that NF- κ B is an LPS-responsive regulator of innate immune responses.

For example, our transcription factor analysis discovered crucial TFs, like NFKB1, NFKB2, RELB, and others for monocytes (Fig. 2B blue rectangle). We also compared the expression patterns of NFKB1 and NFKBIZ (Figure S7C and D). NFKB1 displayed a universal gene expression pattern in all cell types over all time points, while NFKBIZ was mainly detected in monocytes, DCs, and ILCs, especially in 2 h-LPS and 4 h-LPS. It is generally accepted that NF- κ B is a known pleiotropic TF present in almost all cell types and is involved in many biological processes such as inflammation, immunity, differentiation, cell growth, tumorigenesis, and apoptosis. Moreover, we found other TFs, such as IRF5, IRF7, IRF9, and STAT1 (Fig. 2B). Earlier bulk studies have shown that IRF5, IRF7, and IRF9 belong to the interferon response factor (IRF) family. After activation via the JAK-STAT signaling pathway, these TFs bind specifically to the interferon consensus sequence (ICS) in the upstream promoters [31] and

regulate transcription of interferons and inflammatory cytokines [32]. They control many aspects of innate and adaptive immune responses, including responding to pathogens to induce pro-inflammatory responses and regulating immune cell differentiation. Therefore, our single-cell analyses confirmed the previous bulk study results of these critical TFs. In our DEG analyses, we pinpointed many factors like monocyte chemotactic protein-1 (CCL2) and monocyte chemotactic protein-3 (CCL7), which can regulate the chemotaxis and other functions of monocytes [33]. CCL2 is a chemokine that belongs to the CC chemokine family [34]. CCL2 is also called monocyte chemoattractant protein 1 (MCP1) and small inducible cytokine A2. CCL2 recruits monocytes, memory T cells, and dendritic cells to the sites of the inflammation [35]. CXCL2 is another small cytokine belonging to the CXC chemokine family. It activates cells via binding to a cell surface chemokine receptor CXCR2 [36].

Additionally, our previous studies using bulk RNA-seq data demonstrated that the immune system was significantly associated with many health and fertility traits in the cattle [30, 37]. This study further detected trait-relevant cell types by integrating LPS-induced DEGs with large-scale GWAS of 45 complex traits in Holstein. We found that selected DEGs were significantly associated with immune-relevant health, milk production, and body conformation traits.

Limitations and future directions

Some essential marker genes are not detected in this study. These can be due to methodological noise, where a gene is expressed but not detected by the sequencing technology, and/or due to the biological absence of expression. Moreover, we did see discrepancies in cell-type assignments using different methods. For example, SingleR assigned C5 and C6 as monocytes, while marker gene expressions and Azimuth annotated them as DCs and macrophages. These are not surprising partially because monocytes, some DCs, and macrophages are closely related, such that *in silico* predictions may not be reliable. We also checked the relative portion changes among the seven cell types across different time points during LPS treatment (Fig. 1C). We found monocytes decreased first from 11.42% in Control to 7.96% in 2 h-LPS and 6.33% in 4 h-LPS and then increased to 29.21% in 8 h-LPS (Table S1). This corresponded to that monocyte's cell cycle index increased over the LPS treatment time course. However, it is noted that these cell number changes were from one-time measurement and may be impacted by the Azimuth cell type assignment. T cells also changed gene expression and cell activation, resulting from bystander effects secondary to the monocyte response. In addition, T cells may respond to LPS because a recent report shows that TLR2/4 are expressed by bovine T cells [26]. There are also known differences in PBMCs of these two mammalian species, which we did not detect. For example, besides common α and β T cells, γ and δ T cells typically represent 1–10% of circulating T lymphocytes in adult human individuals and approximately 10–25% in adult cattle. This number can be as high as 40% in the young calves [38]. Previous work has also shown that human and bovine $\gamma\delta$ T cells can be directly activated by LPS, suggesting an innate role of $\gamma\delta$ T cells [39]. We were unable to demonstrate a sufficient number of $\gamma\delta$ T cells for analysis in this study because adult cattle have much lower levels of circulating $\gamma\delta$ T cells. Our ability of $\gamma\delta$ T cell assignment was also undermined, probably because we used human reference cell types to assign cattle cells. These designations might be biased towards human-specific features and functions. Therefore, more dedicated experiments are warranted to investigate the roles of ruminant-specific $\gamma\delta$ T cells in cattle.

Conclusions

The functional results inferred from these single-cell-based data sets were consistent with previous findings. They revealed new findings in LPS-driven cell proliferation and differentiation, differential gene transcription, and correlation between DEGs and production traits in cattle. Single-cell analyses provide an unprecedented

opportunity to dissect cell lineages and heterogeneity and understand their identity, differentiation, and function. The successful applications of these new technologies in farm animals like cattle indicated that some research bottleneck problems could be alleviated, e.g., only limited immunological reagents are available in cattle. This study provides an initial example for cattle single-cell analysis. It opens the door for discoveries about the roles of cell types and marker genes in complex traits at single-cell resolution.

Materials and methods

Sample collection

All samples were collected with the approval of the Dairy Cattle Research Centre in Shandong Academy of Agricultural Sciences under Protocol 20–123, and all experiments were carried out in compliance with the ARRIVE guidelines.

Four 2-year old Holstein female lactating cattle were used for blood collection from the tail vein in Jinan Jiabao Dairy Co., Ltd. After pooling; four whole blood samples included either no LPS treatment—control sample CO, or three treated samples with LPS (2 μ g/ml, Product Number: L2880, Sigma-Aldrich, Saint Louis, MO, USA) for 2 h (2 h-LPS), 4 h (4 h-LPS), and 8 h (8 h-LPS) at 37 °C. PBMCs were isolated by centrifugation of whole blood on Hanks' Balanced Salt Solution (Solarbio; Beijing, China) at 500 g for 20 min at room temperature.

Single-cell isolation, scRNA-seq, and scATAC-seq library preparation and sequencing

After cell isolation, scRNA-seq Library for 10 \times Genomics v3 chemistry was generated following the Chromium Single Cell 3' Reagent Kits v3 User Guide: CG000183 Rev C. In brief, cells were barcoded and mixed with reverse transcriptase into a Gel Beads-In-Emulsions (GEMs), then R1 (read 1 primer sequence) was put into the molecules during GEM incubation. P5, P7, a sample index, and Read 2 primer sequence were included during library construction via end repair, A tailing, adaptor ligation, and PCR. The final libraries containing the P5 and P7 primers were used in Illumina bridge amplification.

For scATAC-seq, PBMC nuclei were prepared for library preparation sequencing. Library generation was accomplished following the Chromium Single Cell ATAC Reagent Kits v1.1 User Guide: CG000209 Rev D. Concisely, Nuclei suspensions were incubated in a Transposition Mix that includes a Transposase, which preferentially fragmented the DNA in open regions of the chromatin. Instantaneously, adapter sequences were added to the ends of the DNA fragments. Nuclei were barcoded into a Gel Beads-In-Emulsions (GEMs), a sample index, P7, and Read 2 sequence were added

during library construction via PCR. In the same way, the scATAC-seq libraries contained the P5 and P7 primers used in Illumina bridge amplification. Finally, scRNA-seq and scATAC-seq libraries were sequenced on the Illumina Novaseq 6000 platform (Illumina, San Diego, CA, USA) with double-end 150 bp.

Generation of single-cell transcriptomes

10X Genomics raw data were handled by the Cell Ranger Single-Cell Software Suite (release 3.1.0) and Cell Ranger “*mkfastq*” was used to demultiplex raw base-call files into FASTQ files followed by using Cell Ranger “*count*” to perform alignment, filtering, barcode counting, and UMI counting. Using default parameters, the raw reads were aligned to the ARS-UCD1.2 cattle reference genome [40] by Cell Ranger “*pipeline*” using default parameters. The results are summarized in Supplemental Table S1. All downstream single-cell analyses were accomplished using the Seurat 3.2 [22] R package v3.6.3.

Quality control, dimension reduction, and cell clustering

Seven thousand one hundred seven (Control), 9,174 (2 h-LPS), 6,741 (4 h-LPS), and 3,119 (8 h-LPS) cells passed the quality control thresholds. All genes expressed in fewer than three cells were removed. The cut-off of the number of gene expressions per cell was set at 200 as low and < 3,000 as high; UMI counts less than 200; the percent of mitochondrial-DNA derived gene-expression < 20%. LogNormalize method of the “Normalization” function was used to determine the expression value of genes. We then constricted the corrected expression matrix to the subsets of HVG, centered, and scaled values before performing dimension-reduction and clustering. We selected 2,000 genes as HVG using the “*FindVariableFeatures*” function with default parameters. The “*RunPCA*” function was used to perform the principal components analysis (PCA) on the single-cell expression matrix with genes restricted to HVG. Using a permutation test implemented by the “*JackStraw*” function, we determined the number of significant principal components (PC). The top 12 PCs were used for clustering and UMAP analysis. The weighted Shared Nearest Neighbor (SNN) graph-based clustering method executed by the “*FindNeighbors*” function was used to find clusters. We utilized the “*FindClusters*” function to conduct the cell-clustering analysis by inserting cells into a graph structure in the PCA cluster. Based on the number of cells in our study, we set the parameter resolution to 0.05. Visualization of the cells was performed using the UMAP algorithm as implemented by the Seurat “*RunUMAP*” function. With default parameters, canonical cell-type marker genes maintained across conditions were identified using the “*FindConservedMarkers*” function.

Assigning cell type labels to single-cell clusters

We utilized two methods to label the cell clusters identified by Seurat. First, we projected the PBMC data onto an annotated PBMC CITE-Seq reference dataset [41] using Azimuth [24]. Each cell received an assignment and prediction score to a cell class in the reference. We normalized data using the “*SCTransform*” function [42] and then found anchors between reference and query using “*FindTransferAnchors*.” Here we used a precomputed supervised PCA (spca) transformation. We then transferred cell type labels and protein data from the reference to the query using “*MapQuery*.” Additionally, we used SingleR [25] to annotate raw expression data for the filtered cells with default parameters using the Blueprint [43] and Encode [44] human cell atlases.

Pseudotime trajectory analysis

For trajectory analysis, we used Monocle 2 [27] to order cells in pseudotime based on their transcriptional similarities, with UMI counts modeled using a negative binomial distribution. First, we integrated the preprocessed Seurat objects into Monocle 2 utilizing the “*newCellDataSet*” function. We then determined the differentially expressed genes or marker genes using the “*differentialGeneTest*” function. We next reduced the dimensionality of the data to two dimensions using the discriminative dimensionality reduction with trees (DDRTree) method implemented in the “*reduceDimension*” function. Finally, after pseudotime calculations were made for each cell, we projected clusters derived from the Seurat object onto the minimum spanning tree upon cell order using the “*plot_cell_trajectory*” function.

Cell-cycle analysis

Sets of 43 G1/S and 55 G2/M genes [45] were used in the cell-cycle analysis. To calculate the ratio of actively proliferating cells of each feature, such as different clusters and different time points, we first calculated the total expression levels of all 98 cell-cycle genes in every single cell, and only cells with mean expression levels higher than the average values of all clusters were regarded as actively proliferating.

Single-cell regulatory network inference and clustering (SCENIC) analysis

We conducted SCENIC analysis on cells after filtering for each major cell type using the R package SCENIC v1.1.2 [26], a computational workflow that predicts TF activities from scRNA-seq data. Briefly, SCENIC infers co-expression modules between TF and candidate target genes using machine learning regression techniques (e.g., random forest or gradient boosting machines), pruned based on the enrichment of the TF motif around the TSS of the

potential target genes, resulting in regulons. Based on the AUCell algorithm, SCENIC calculates each regulator's activity in single-cell transcriptomes to obtain the corresponding area under the curve (AUC) scores, which are used to rank the cells for a given regulon and determine a threshold for active or inactive expression. Then the network activity was converted into ON/OFF, thus making the final output binary (binary regulon activity matrix). Individual regulons were constructed from the scRNA-seq data. Regions for TF searching were restricted to a 10 kb distance centered on the transcriptional start site (TSS) or 500 bp upstream of the TSS. First, TF-gene co-expression modules were defined in a data-driven manner with GENIE3 v1.8.0. Subsequently, those modules were refined via RcisTarget by keeping only those genes that contain the respective transcription factor binding site (TFBS). Once the regulons were constructed, the method AUCell scored individual cells by assessing for each TF separately whether target genes were enriched in the top quantile of the cell signature.

Weighted gene co-expression network analysis

WGCNA was performed with functions in the WGCNA v1.69 R package following the previously published study by Tosches and colleagues [46]. According to the methods, the analyses were performed on pseudocells, calculated as averages of 100 cells randomly chosen within each cluster. DC was not included due to its small cell number. The top 2,000 highly variably expressed genes determined in Seurat were used for analysis. Briefly, the topological overlap matrix (TOM) was constructed with softPower and was set to 2. The hub genes for each module were identified as module eigengene. The GO enrichment analysis was performed by ClusterProfiler [47] R package using hub gene data sets, and the Benjamini-Hochberg method was employed for multiple test correction. GO terms with a *P*-value lower than 0.05 were considered as significantly enriched.

Gene differential expression analysis

To get the lists of marker genes, we first extracted the genes' UMIs across cells within each cluster and then assigned cells to each sample. Based on the gene \times cells matrix, we utilized edgeR [29] to detect DEGs for each cluster in each pairwise comparison among Control, 2 h-LPS, 4 h-LPS, and 8 h-LPS (Tables S10-15).

Single-cell ATAC-seq alignment and data processing

For scATAC-seq analyses, we aligned the sequence using the 10 \times Genomics Cell Ranger ATAC pipeline (version 1.2) against the UCD-ARS1.2 genome. The "Cell Ranger Aggr" function normalizes the number of confidently mapped reads per cell across the libraries. We

processed the data with Seurat and the additional package Signac (v1.1.0) [48]. We first computed quality control (QC) metrics and removed the cells with the number of expressed genes < 500 . We then normalized the filtered data by the "RunTFIDF" function and removed features in less than 20 cells with the "FindTopFeatures" function. We next ran singular value decomposition (SVD) using "LSI" with the features selected above. Next, we performed graph-based clustering by "FindNeighbors" and "FindClusters" functions using the first 30 dimensions of reduction as an input. Finally, the read coverage of regions near specific genes in each group was plotted by the "CoveragePlot" function. On average, 3,798 fragments per cell were obtained, and 4,200 cells were recovered.

GWAS signal enrichment analysis

Details of the single-marker GWAS and fine-mapping analyses designed for the body type, reproduction, and production traits from 27,214 U.S. Holstein bulls, intended for health traits from 11,880–24,699 bulls, and feed efficiency (i.e., RFI) from 3,947 Holstein cows were previously reported [30, 49–51]. As the complex traits being explored were highly polygenic, the sum-based marker-set test methodology shown in Eq. 1 was utilized as in QGG package v1.0 [52] to establish whether GWAS signals were enhanced in marker genes of distinct cell clusters and DEGs of six pairwise comparison groups (Control vs. 2 h-LPS, Control vs. 4 h-LPS, Control vs. 8 h-LPS, 2 h-LPS vs. 4 h-LPS, 2 h-LPS vs. 8 h-LPS, 4 h-LPS vs. 8 h-LPS). We included 20-kb windows around gene regions to identify the potential *cis*-regulatory variants. Previous studies indicated that this method had at best equal power compared to other commonly used GWAS signal enrichment methods in humans [37, 53], *Drosophila melanogaster* [54], and livestock [55–57], especially for the highly polygenic traits.

$$T_{sum} = \sum_{i=1}^{m_f} b^2 \quad (1)$$

In this equation, m_f is the number of genomic markers within a list of genes (marker genes of each cell cluster or DEGs from pairwise comparisons in each cell cluster), and b is the marker weight from single-marker GWAS. We restricted marker-set sizes and linkage disequilibrium patterns among markers by utilizing the genotype cyclical permutation strategy [52]. We first organized marker effects (i.e., b^2) utilizing their chromosome positions (i.e., $b_1^2, b_2^2, \dots, b_{m-1}^2, b_m^2$). We then at random designated one marker (i.e., b_k^2) from this vector as the first place, and altered the remaining ones to new positions while retaining their original orders (i.e., $b_k^2, b_{k+1}^2, \dots, b_{m-1}^2, b_m^2, b_1^2, \dots, b_{k-1}^2$) to conserve LD patterns among markers. We determined a new summary statistic for an allocated list

of genes using their original chromosome locations. To attain an empirical *P*-value for the list of genes, we went over this permutation procedure 10,000 times. We used a one-tailed test of the proportion of random summary statistics greater than that observed.

Cross-species comparison

We downloaded a single-cell RNA-seq dataset of human PBMC from GSE96583. We first merged expression matrices of the two species (cattle and human) based on the intersection of the detected homologous genes. Next, we performed expression matrix preprocessing separately for the two species, followed by integrating three datasets using functions in Seurat v3.2 [22]. The top 13 PCs were selected, and the resolution was set to 0.18 to yield 13 cell clusters.

Abbreviations

CCL2: C–C Motif Chemokine Ligand 2/Chemotactic Protein-1; CCL7: C–C Motif Chemokine Ligand 7/Chemotactic Protein-3; CD14: Cluster of differentiation 14; CD83: Nuclear receptor subfamily 4, group A, member 3; CXCL2: C–X–C motif ligand 2; DC: Dendritic cell; DEGs: Differentially expressed genes; FGF: Fibroblast growth factor; GO: Gene ontology; GRN: Gene regulatory network; HVG: Highly variable genes; ICS: Interferon consensus sequence; IκB: NF-κB inhibitor; IL-1: Interleukin-1; IRF: Interferon response factor; IRF8: Interferon regulatory factor 8; LPS: Lipopolysaccharide; MAPK: P38 mitogen-activated protein kinases; MCP1: Monocyte chemoattractant protein 1; MHC: Major histocompatibility complex; NF-κB: Nuclear factor kappa-light-chain-enhancer of activated B cells; NK: Natural killer; PBMC: Peripheral blood mononuclear cell; S100A12: S100 calcium-binding protein A12; scATAC-seq: Single-cell sequencing assay for transposase-accessible chromatin; scRNA-seq: Single-cell RNA sequencing; SCENIC: Single-Cell rEGulatory Network Inference and Clustering; SCS: Somatic cell score; STAT: Signal transducer and activator of transcription; TF: Transcription factor; TNF: Tumor necrosis factor; TNFα: Tumor necrosis factor-α; UMAP: Uniform Manifold Approximation and Projection; WGCNA: Weighted gene co-expression network analysis.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08562-0>.

Additional file 1: Figure S1. Cell clustering. **Figure S2.** Cell marker gene expression. **Figure S3.** Cell type annotation using Azimuth and SingleR (based on the human Blueprint and Encode cell atlas references). **Figure S4.** Comparative analyses between cattle and human PBMC. **Figure S5.** Comparative analyses between cattle and human PBMC under two resolutions using Azimuth. **Figure S6.** Cell cycle analysis for cattle PBMC. **Figure S7.** Gene expression of innate immunity genes and transcription factors during lipopolysaccharide treatments in cattle PBMC. **Figure S8.** Heatmap showing scaled expression levels of three gene modules (core antiviral, peaked inflammatory, and sustained inflammatory) in CD4 cells, CD8 cells, B cells, separately and jointly with monocytes.

Additional file 2: Table S1. Summary of scRNA-seq and scATAC-seq dataset. **Table S2.** Marker genes of each cell type. **Table S3.** Cattle PBMC cell type annotation under three resolutions using Azimuth. **Table S4.** Cattle PBMC cell type annotation using SingleR. **Table S5.** The expression of 93 cell cycle-related genes in each cell. **Table S6.** The summary information of TF identified by SCENIC. **Table S7.** Single cell's pseudotime value obtained from Monocle2. **Table S8.** Gene list of each module identified

by WGCNA. **Table S9.** Enrichment results of each module identified by WGCNA.

Additional file 3: Table S10. Differentially expressed genes with each cell cluster between CO and T1 identified by edgeR. **Table S11.** Differentially expressed genes with each cell cluster between CO and T2 identified by edgeR. **Table S12.** Differentially expressed genes with each cell cluster between CO and T3 identified by edgeR. **Table S13.** Differentially expressed genes with each cell cluster between T1 and T2 identified by edgeR. **Table S14.** Differentially expressed genes with each cell cluster between T1 and T3 identified by edgeR. **Table S15.** Differentially expressed genes with each cell cluster between T2 and T3 identified by edgeR. **Table S16.** Human and cattle PBMC cell type annotation under three resolutions using Azimuth.

Acknowledgements

We thank Kun Zhao and Research Animal Services staff at Jinan Jiabao Dairy Co., Ltd. This research used resources provided by the SCINet project of the USDA ARS project number 0500-00093-001-00-D.

Authors' contributions

JBL, CJL, LF, and GEL conceived and designed the experiments. GC, YW, and JBL collected samples and/or generated NGS data. YG, GC, YW, YL, XZ, RL, YG, WT, RL, and GEL performed in silico prediction and computational analyses. GEL, WT, YG, and CJL wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported in part by China Agriculture Research System of MOF and MARA (CARs-36), Agricultural Scientific and Technological Innovation Project of Shandong Academy of Agricultural Sciences (Cxgc2016A04), Breeding and demonstration of high yield and β-casein A2 dairy cows (2019B10018), USDA National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) grant numbers 2016–67015-24886 and 2019–67015-29321, and the US-Israel Binational Agricultural Research and Development (BARD) Fund grant number US-4997–17. L. Fang was partially funded through the HDR-UK award HDR-9004 and the Marie Skłodowska-Curie grant agreement No [801215].

Availability of data and materials

The accession number for the scRNA-seq data reported in this study is GEO: GSE166473. The GWAS summary statistics for all complex traits have been submitted to Figshare, i.e., body type, production, and reproduction traits under <https://figshare.com/s/ea726fa95a5bac158ac1>, and the remaining ones under <https://figshare.com/s/94540148512dddf7ed32>. All scripts and source codes can be found in the Supplemental Material and in <https://github.com/YahGao/PBMC-scRNA-seq>.

Declarations

Ethics approval and consent to participate

All samples were collected with the approval of the ethics committee, Dairy Cattle Research Centre in Shandong Academy of Agricultural Sciences under Protocol 20–123, and all experiments were carried out in compliance with the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines and regulations. No human or human objects were directly involved in this experiment. Human PBMC scRNA-seq dataset was downloaded from the database, GSE96583, NCBI, NIH [14, 15].

Consent for publication

Not applicable.

Competing interests

All authors declare no potential conflict of interest.

Author details

¹Institute of Animal Science and Veterinary Medicine, Shandong Academy of Agricultural Sciences, No.202, Gongyebei Road, Jinan 250100, China. ²Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, MD

20705, USA. ³Shandong Ox Livestock Breeding Co., Ltd, Jinan 250100, China. ⁴College of Animal Science and Technology, China Agricultural University, Beijing 100193, China. ⁵Animal Parasitic Diseases Laboratory, BARC, USDA-ARS, Beltsville, MD 20705, USA. ⁶MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.

Received: 27 September 2021 Accepted: 13 April 2022

Published online: 30 April 2022

References

- Rietschel ET, Kirikae T, Schade FU, Mamat U, Schmidt G, Loppnow H, Ulmer AJ, Zähringer U, Seydel U, Di Padova F, et al. Bacterial endotoxin: molecular relationships of structure to activity and function. *Faseb j*. 1994;8(2):217–25.
- Li CJ, Li RW, Elsasser TH, Kahl S. Lipopolysaccharide-induced early response genes in bovine peripheral blood mononuclear cells implicate GLG1/E-selectin as a key ligand-receptor interaction. *Funct Integr Genomics*. 2009;9(3):335–49.
- Wong HR, Odoms K, Sakthivel B. Divergence of canonical danger signals: the genome-level expression patterns of human mononuclear cells subjected to heat shock or lipopolysaccharide. *BMC Immunol*. 2008;9:24.
- Iwasaki A, Medzhitov R. Toll-like receptor control of the adaptive immune responses. *Nat Immunol*. 2004;5(10):987–95.
- Martinez J, Huang X, Yang Y. Direct action of type I IFN on NK cells is required for their activation in response to vaccinia viral infection in vivo. *J Immunol*. 2008;180(3):1592–7.
- Poltorak A, He X, Smirnova I, Liu MY, Van Huffel C, Du X, Birdwell D, Alejos E, Silva M, Galanos C, et al. Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science*. 1998;282(5396):2085–8.
- Mazgaeeen L, Gurung P. Recent advances in lipopolysaccharide recognition systems. *Int J Mol Sci*. 2020;21(2):379.
- Schletter J, Heine H, Ulmer AJ, Rietschel ET. Molecular mechanisms of endotoxin activity. *Arch Microbiol*. 1995;164(6):383–9.
- Arango Duque G, Descoteaux A. Macrophage cytokines: involvement in immunity and infectious diseases. *Front Immunol*. 2014;5:491.
- Davis JM 3rd, Knutson KL, Strausbauch MA, Crowson CS, Therneau TM, Wettstein PJ, Matteson EL, Gabriel SE. Analysis of complex biomarkers for human immune-mediated disorders based on cytokine responsiveness of peripheral blood cells. *J Immunol*. 2010;184(12):7297–304.
- Ziegler-Heitbrock L, Ancuta P, Crowe S, Dalod M, Grau V, Hart DN, Leenen PJ, Liu YJ, MacPherson G, Randolph GJ, et al. Nomenclature of monocytes and dendritic cells in blood. *Blood*. 2010;116(16):e74–80.
- Amit I, Garber M, Chevrier N, Leite AP, Donner Y, Eisenhaure T, Guttman M, Grenier JK, Li W, Zuk O, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*. 2009;326(5950):257–63.
- Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublotte JT, Yosef N, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*. 2014;510(7505):363–9.
- Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;36(1):89–94.
- Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, et al. Construction of a human cell landscape at single-cell level. *Nature*. 2020;581(7808):303–9.
- Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublotte JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. 2013;498(7453):236–40.
- Morrison WJ, Baldwin CL, MacHugh ND, Teale AJ, Goddeeris BM, Ellis J. Phenotypic and functional characterisation of bovine lymphocytes. *Prog Vet Microbiol Immunol*. 1988;4:134–64.
- Hein WR, Mackay CR. Prominence of gamma delta T cells in the ruminant immune system. *Immunol Today*. 1991;12(1):30–4.
- Brown WC, Rice-Ficht AC, Estes DM. Bovine type 1 and type 2 responses. *Vet Immunol Immunopathol*. 1998;63(1):45–55.
- Ziegler-Heitbrock L. Monocyte subsets in man and other species. *Cell Immunol*. 2014;289(1–2):135–9.
- Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
- Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive Integration of single-cell data. *Cell*. 2019;177(7):1888–1902 e1821.
- Franzén O, Gan LM, Björkgren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database (Oxford)*. 2019;2019:baz046.
- Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zagar M, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573–87 e29.
- Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol*. 2019;20(2):163–72.
- Aibar S, Gonzalez-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine JC, Geurts P, Aerts J, et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*. 2017;14(11):1083–6.
- Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*. 2017;14(10):979–82.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139–40.
- Fang L, Cai W, Liu S, Canela-Xandri O, Gao Y, Jiang J, Rawlik K, Li B, Schroeder SG, Rosen BD, et al. Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res*. 2020;30(5):790–801.
- Weisz A, Marx P, Sharf R, Appella E, Driggers PH, Ozato K, Levi BZ. Human interferon consensus sequence binding protein is a negative regulator of enhancer elements common to interferon-inducible genes. *J Biol Chem*. 1992;267(35):25589–96.
- Taniguchi T, Ogasawara K, Takaoka A, Tanaka N. IRF family of transcription factors as regulators of host defense. *Annu Rev Immunol*. 2001;19:623–55.
- Ziegler-Heitbrock L. The CD14+ CD16+ blood monocytes: their role in infection and inflammation. *J Leukoc Biol*. 2007;81(3):584–92.
- Xu LL, Warren MK, Rose WL, Gong W, Wang JM. Human recombinant monocyte chemotactic protein and other C-C chemokines bind and induce directional migration of dendritic cells in vitro. *J Leukoc Biol*. 1996;60(3):365–71.
- Carr MW, Roth SJ, Luther E, Rose SS, Springer TA. Monocyte chemoattractant protein 1 acts as a T-lymphocyte chemoattractant. *Proc Natl Acad Sci U S A*. 1994;91(9):3652–6.
- Wolpe SD, Sherry B, Juers D, Davatelis G, Yurt RW, Cerami A. Identification and characterization of macrophage inflammatory protein 2. *Proc Natl Acad Sci U S A*. 1989;86(2):612–6.
- Liu S, Yu Y, Zhang S, Cole JB, Tenesa A, Wang T, McDanel TG, Ma L, Liu GE, Fang L. Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of complex traits in cattle and human. *BMC Biol*. 2020;18(1):80.
- Wilson RA, Zolnai A, Rudas P, Frenyo LV. T-cell subsets in blood and lymphoid tissues obtained from fetal calves, maturing calves, and adult bovine. *Vet Immunol Immunopathol*. 1996;53(1–2):49–60.
- Hedges JF, Lubick KJ, Jutila MA. Gamma delta T cells respond directly to pathogen-associated molecular patterns. *J Immunol*. 2005;174(10):6045–53.
- Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elisk CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience*. 2020;9(3):giaa021.
- Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, Satija R, Smibert P. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*. 2017;14(9):865–8.

42. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 2019;20(1):296.
43. Stunnenberg HG. International human epigenome C, hirst M: the international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell.* 2016;167(5):1145–9.
44. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57–74.
45. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell.* 2015;161(5):1202–14.
46. Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, Laurent G. Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science.* 2018;360(6391):881.
47. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16(5):284–7.
48. Stuart T, Srivastava A, Lareau C, Satija R. Multimodal single-cell chromatin analysis with Signac. *Nat Methods.* 2021;18(11):1333–41.
49. Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun Biol.* 2019;2(1):212.
50. Li B, Fang L, Null DJ, Hutchison JL, Connor EE, VanRaden PM, VandeHaar MJ, Tempelman RJ, Weigel KA, Cole JB. High-density genome-wide association study for residual feed intake in Holstein dairy cattle. *J Dairy Sci.* 2019;102(12):11067–80.
51. Freebern E, Santos DJA, Fang L, Jiang J, Parker Gaddis KL, Liu GE, VanRaden PM, Maltecca C, Cole JB, Ma L. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. *BMC Genomics.* 2020;21(1):41.
52. Rohde PD, Fourie Sørensen I, Sørensen P. qqg: an R package for large-scale quantitative genetic analyses. *Bioinformatics.* 2019;36(8):2614–5.
53. Rohde PD, Demontis D, Cuyabano BCD, Børklum AD, Sørensen P. Covariance Association Test (CVAT) identifies genetic markers associated with schizophrenia in functionally associated biological processes. *Genetics.* 2016;203(4):1901–13.
54. Sørensen IF, Edwards SM, Rohde PD, Sørensen P. Multiple trait covariance association test identifies gene ontology categories associated with chill coma recovery time in *Drosophila melanogaster*. *Sci Rep.* 2017;7(1):2413.
55. Sarup P, Jensen J, Ostensen T, Henryon M, Sørensen P. Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs. *BMC Genet.* 2016;17(1):11.
56. Fang L, Sahana G, Su G, Yu Y, Zhang S, Lund MS, Sørensen P. Integrating sequence-based GWAS and RNA-seq provides novel insights into the genetic basis of mastitis and milk production in dairy cattle. *Sci Rep.* 2017;7(1):45560.
57. Fang L, Sahana G, Ma P, Su G, Yu Y, Zhang S, Lund MS, Sørensen P. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol.* 2017;49(1):44.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



RESEARCH

Open Access



Towards the detection of copy number variation from single sperm sequencing in cattle

Liu Yang^{1,2†}, Yahui Gao^{1,3†}, Adam Oswalt⁴, Lingzhao Fang⁵, Clarissa Boschiero¹, Mahesh Neupane¹, Charles G. Sattler⁴, Cong-jun Li¹, Eyal Seroussi⁶, Lingyang Xu⁷, Lv Yang⁸, Li Li², Hongping Zhang², Benjamin D. Rosen¹, Curtis P. Van Tassell¹, Yang Zhou⁸, Li Ma^{3*} and George E. Liu^{1*}

Abstract

Background: Copy number variation (CNV) has been routinely studied using bulk-cell sequencing. However, CNV is not well studied on the single-cell level except for humans and a few model organisms.

Results: We sequenced 143 single sperms of two Holstein bulls, from which we predicted CNV events using 14 single sperms with deep sequencing. We then compared the CNV results derived from single sperms with the bulk-cell sequencing of one bull's family trio of diploid genomes. As a known CNV hotspot, segmental duplications were also predicted using the bovine ARS-UCD1.2 genome. Although the trio CNVs validated only some single sperm CNVs, they still showed a distal chromosomal distribution pattern and significant associations with segmental duplications and satellite repeats.

Conclusion: Our preliminary results pointed out future research directions and highlighted the importance of uniform whole genome amplification, deep sequence coverage, and dedicated software pipelines for CNV detection using single cell sequencing data.

Keywords: Cattle, Single sperm sequencing, Copy number variation

Background

Copy number variation (CNV) is defined as deletions, insertions, and duplications ranging from 50 base pairs (bp) to 5 million base pairs (Mbp) between any individuals [1]. CNV has been extensively studied in multiple species for its functional impacts on gene expression, such as altering gene dosage, disrupting coding sequence, or perturbing long-range gene regulation [2]. To date, CNV has been investigated in humans [1, 3–7], mice [8–10], and domesticated animals [11–20]. In cattle, we and others

reported germline/inherited and somatic CNV using microarrays and short-read sequencing in breeds like Angus, Holstein, Hanwoo, Brown Swiss, Simmental, and Qinchuan [19, 21–28].

Recent breakthroughs in the development and application of single-cell sequencing technologies provide an avenue for dissecting population lineages and heterogeneity and understanding cell identity, differentiation, and function [29–34]. Single-cell DNA-seq (scDNA-seq) technologies produce data, which is ideal for detecting CNV or abnormal chromosome numbers (aneuploidy) on the single-cell level [35–37]. Because copy number aberrations (CNAs), which are pathogenic CNVs, play an important role in the initiation and progression of cancer, they have been intensively studied using single-cell sequencing in humans [38, 39]. Currently, multiple

*Correspondence: lima@umd.edu; George.Liu@ars.usda.gov

[†]Liu Yang and Yahui Gao contributed equally to this work.

¹ Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, MD 20705, USA

³ Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

analysis tools are available for detecting CNVs in human scDNA-seq data, as reviewed recently [40].

However, no report has been published on the CNV identification on the single-cell level in livestock, including cattle. Here we sequenced and analyzed 143 single sperm genomes from two Holstein bulls, identifying thousands of candidate CNV events. We attempted to validate the single-sperm sequencing-based CNV results using the data derived from the diploid genome sequencing of one bull's family trio. Since one mechanism of CNV formation is non-allelic homologous recombination (NAHR), a recent paper reported that NAHR leads to over two-thirds of the structural variation detected within the human genome [41]. We also investigated CNVs and their associated segmental duplications [2]. To the best of our knowledge, this is the first reported trial of single sperm genome sequencing in livestock, highlighting future CNV detection directions using scDNA-seq data and opening the door for studying individual sperm genome and male infertility.

Results

Sequencing of haploid sperms and diploid trio

Sequencing of sperms

We chose two bulls with different fertility capabilities (See [Methods](#)). Using the MALBAC method [42], we amplified and sequenced a total of 156 single sperm cells manually picked from two Holstein bulls' semen. After quality control filtering, 143 sperm data (71 for Sample1 and 72 for Sample2) were kept for downstream analyses. The sequenced sperms had an average of $1.79 \times$ genome coverage, and 16 of them were at $\sim 4 \times$ genome coverage, achieving an overall coverage of $\sim 11.40\%$ to $\sim 41.35\%$ of the genome, respectively (Table S1). On average, 98.18% of sequencing reads from single sperms were mapped on the bovine ARS-UCD1.2 genome.

Sequencing of the trio

For Samples1's family trio diploid genomes, we sequenced bulk DNA samples extracted from ear punches of Sample1, its sire Sample1-sire, and dam Sample1-dam to approximately 40, 10, and $20 \times$ genome coverage, respectively, with over 99% genome mapping rate and covering 96% genome sequence (Table S2).

Segmental duplication analysis

Delineation of the recent duplication events at the genomic-sequence level, particularly sequences located at their junctions [43], may provide insight into their mechanism of origin. Because SDquest can detect recent and ancient segmental duplication (SegDup) [44], we applied it to the latest bovine ARS-UCD1.2 genome assembly. A total of 27,560 pairwise SegDup sequence

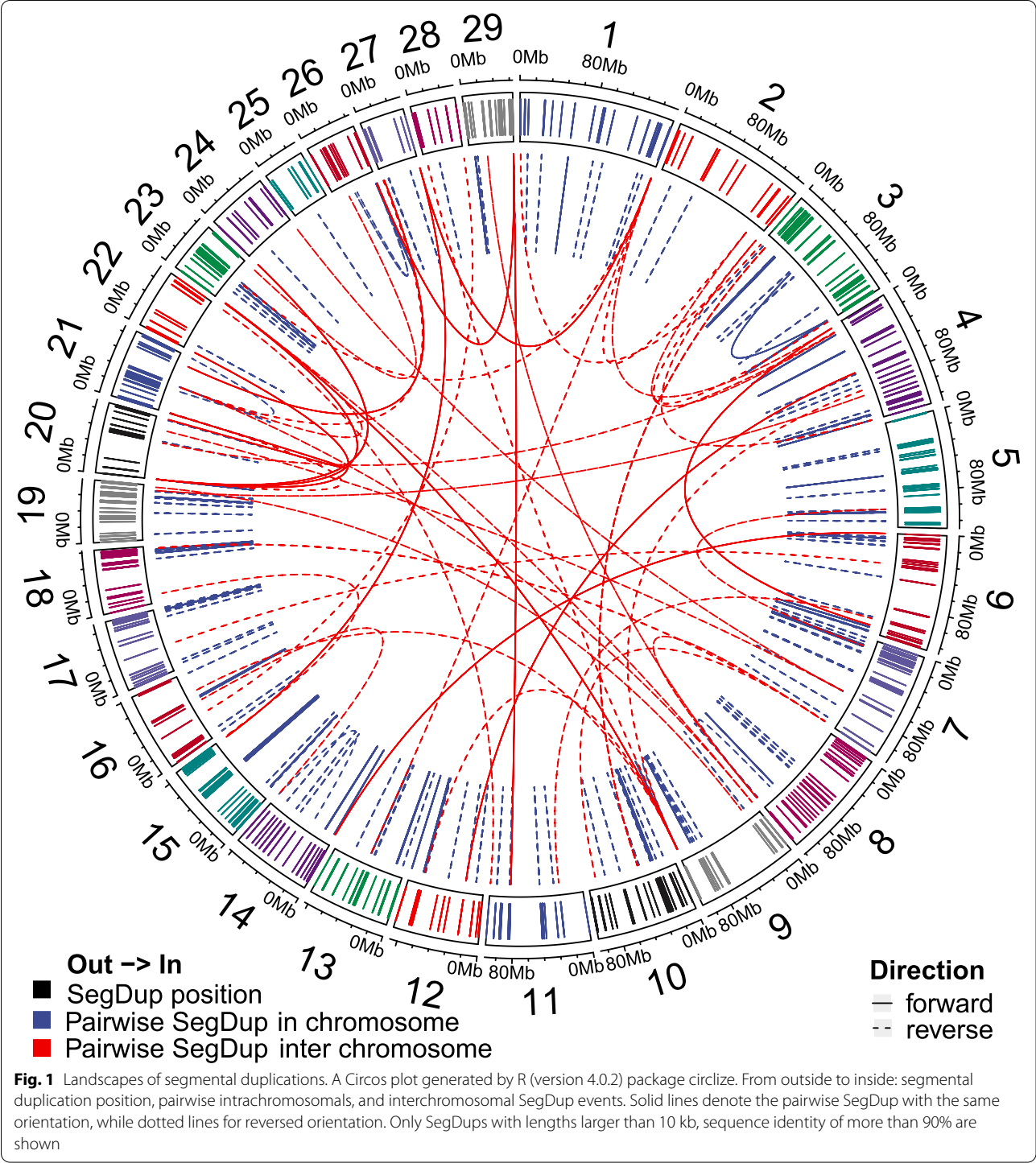
fragments were reported by SDquest, with 49,126 unique nonredundant fragment regions (Table S3). Among them, 12,400 (44.99%) and 3,374 (12.24%) pairwise SegDups have sequence identity larger than 80% and 90%, respectively. Also, 17,477 (63.41%) pairwise SegDup sequence fragments are reversed in their orientations on the chromosomes, while 16,621 (60.31%) are interchromosomally distributed (Fig. 1). After merging neighboring pairwise SegDup sequence fragments, we detected a total of 9,445 SegDup regions, covering 2.89% of the bovine genome (71,877,120 bp) (Table S4). As shown in Table S5, chr3 has the highest count of SegDup regions (600), chr5 has the largest length of SegDup regions (5,004,378 bp), and chr29 has the largest percentage of SegDup coverage (7.42%).

Following our previous study [45], we analyzed repetitive sequence contents in and near SegDup regions (Table S6, Methods). We evaluated the repeat content of duplicated sequence, 20 kb flanking sequence, and the whole genome. As reported before [35], SINE Alu repeats were associated with human segmental duplications, but we did not find SINE enrichment was enriched for bovine segmental duplications. However, we detected two clear patterns regarding repeat content. While LINE content remains similar, DNA and SINE repeat content of most duplications are reduced. We observed a reverse trend for LTR and satellite repeat sequences, even though the fold change for LTR is only 1.25 (Table S6, Random simulation test, P -value < 0.001). Bovine segmental duplications show a 2.84-fold enrichment for satellite repeat content and a 2.03-fold elongation for satellite repeat average length over the genome average (Table S6), agreeing with our earlier observation [45].

We also performed gene annotation for those SegDup regions and found 3,724 SegDups overlapping with 2,969 genes, which were significantly enriched (adjusted P -value < 0.05) in the GO term of GTPase activity and 12 KEGG pathways, such as metabolism of xenobiotics by cytochrome P450 and antigen processing and presentation (Table S7), again agreeing with our previous cattle results and the results from other species [8–10, 45]. When compared with the cattle QTL database [46], we found a total of 837 QTLs intersected with 425 SegDups. We also found that eight QTLs were significantly enriched (adjusted P -value < 0.05 after the Benjamini–Hochberg correction for multiple testing) for animal reproduction and health traits, such as conception rate, inseminations per conception, stillbirth, bovine respiratory disease susceptibility, and others (Table S7).

Copy number variations in sperms and trio genomes

Using single sperms with deep sequencing from Sample1 ($n=8$) and Sample2 ($n=6$), as well as Sample1



trio somatic samples, we detected a total of 5,646 CNVs (ranging from 50 bp to 5 Mb), including 1,307 break end (BND), 2,779 deletion (DEL), 877 duplication (DUP), and 683 inversion (INV) events (Table 1, Table S8, and Table S9). Totally 0.27% of autosomes were covered by 6.73 Mb length of CNV (Table S10). We then focused on CNVs

(i.e., DEL and DUP), which are shown in Fig. 2 and Fig. S1. Similar to the recombination maps derived from the same sequence data (Yang et al., 2021 submitted), CNV distributions are significantly enriched in the two ends of chromosomes (Fig. 3). This result was also in line with

Table 1 Statistics of copy number variation by group

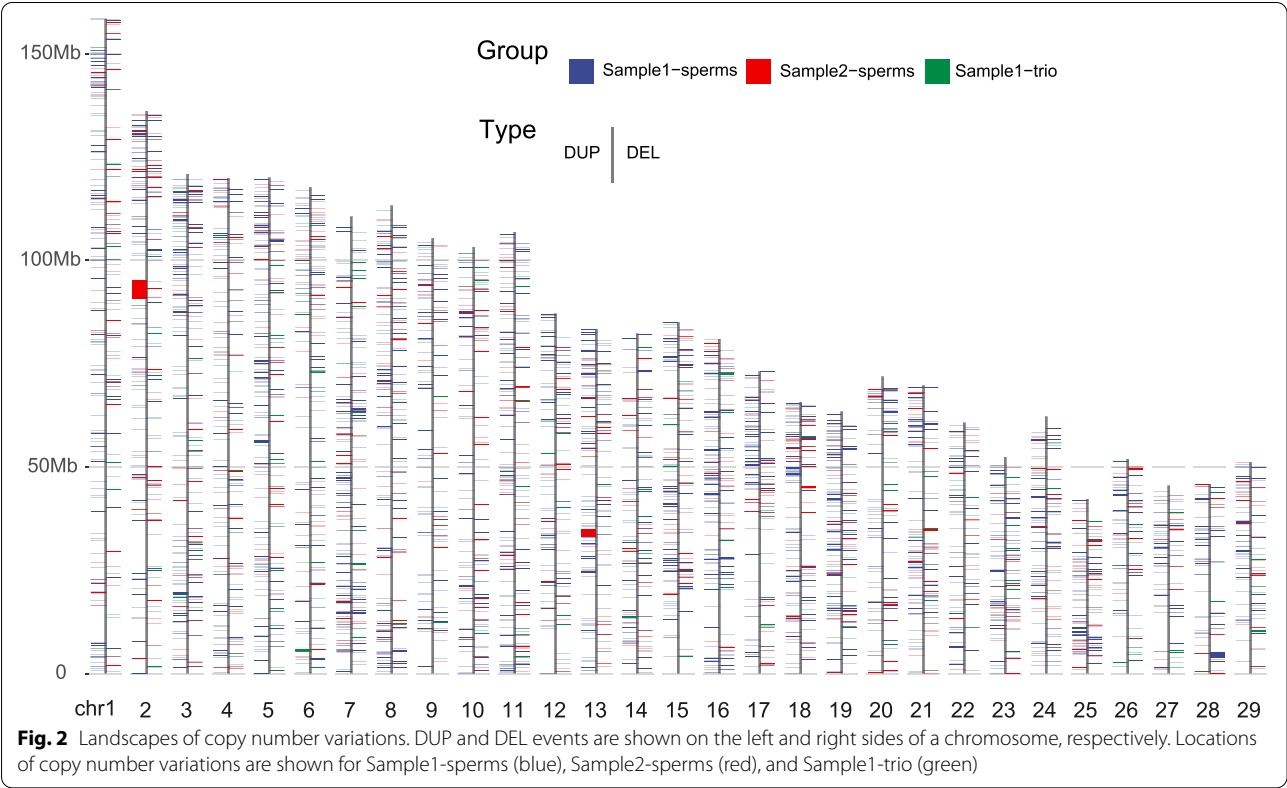
ID	Count					Length (kb)				Genome covered			
	BND	DEL	DUP	INV	Total	DEL	DUP	INV	Total	DEL	DUP	INV	Total
Total	1307	2779	877	683	5646	9724.72	16,140.42	598.91	26,464.05	0.391%	0.648%	0.024%	1.063%
Total sperms	1262	2495	859	666	5282	9048.34	14,305.07	472.94	23,826.35	0.363%	0.575%	0.019%	0.957%
Sum sample1-sperms	919	1714	732	654	4019	6892.22	6476.01	378.86	13,747.09	0.277%	0.260%	0.015%	0.552%
Sum sample2-sperms	343	781	127	12	1263	2156.12	7829.06	94.08	10,079.26	0.087%	0.314%	0.004%	0.405%
Total sample1-trio	45	284	18	17	364	676.38	1835.34	125.98	2637.70	0.027%	0.074%	0.005%	0.106%

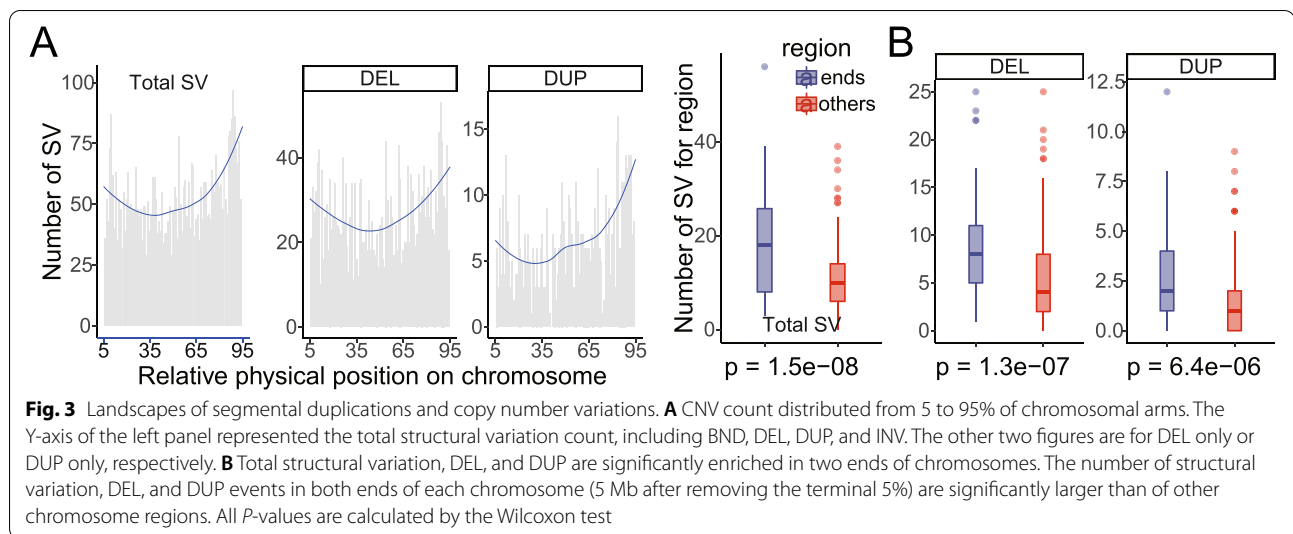
the human results from two recent large-scale CNV discovery studies [47, 48].

After removing the CNV singletons (i.e., DEL or DUP occurred only once in one sample), we obtained 433 DEL and DUP events in a total length of 1,736.6 kb. Within them, 4.16% (18 out of 433, 7.17% for length) of CNVs were detected in all three groups. Near 51.04% (221/433, 71.00% for length) were found in two groups. Among those, 10.39% SV (45/433, 15.97% for length) were shared in Sample1-sperms and Sample1-trio, and 36.49% of the total events (158/433, 32.22% for length) were shared in Sample1-sperms and Sample2-sperms (Fig. S2 and Table S10). CNVs unique to Sample1-trio accounted for only 3.46% events (15/433) (2.93% for length), while CNVs unique to Sample1-sperms accounted for 28.41% events (123/433) (16.16% for

length) (Fig. S2 and Table S10). These results indicated that a larger number of CNVs detected in sperms did not overlap with the trio CNVs. Those 433 CNVs mapped 968 genes, which were significantly enriched (adjusted *P-value* < 0.05) with four GO terms of cell migration and one KEGG pathway of pancreatic secretion (Table S11).

In addition, we analyzed the repeat content in and near 2,485 nonredundant CNV regions, similar to what we did for SegDup as described above (Table S12). We tested two flanking window sizes: 5 and 20 kb. For Sample1 sperm, Sample2 sperm, and Sample1 diploid CNVs, we consistently observed that SegDup (3.67–13.79 folds) and satellites (up to 5.64 folds) were enriched in CNVs (all *P-values* < 0.001). Within the 5 kb flanking regions, the enrichment folds of SegDups





and satellites fall to 1.74–2.91, 0.45–2.42 folds, respectively. They gradually decrease to the genome average as flanking windows around the CNV increase to 20 kb.

Discussion

Single-cell sequencing and analysis are still facing multiple grand challenges [49]. To the best of our knowledge, this is the first trial of single sperm sequencing in the livestock, and we will discuss what was achieved and what needs further improvement.

It is generally accepted that the *de novo* CNV event is infrequent. By mapping each sperm's sequencing data to the reference genome, our results focused on the germline/inherited CNVs, which are the common CNV events shared by single sperms. Using the CNV results derived from the trio bulk-cell sequencing as the ground truth, we estimated shared and unique percentages among the three groups (Sample1-sperms, Sample2-sperms, and Sample1-trio). To our disappointment, only 10.39% of CNVs detected in Sample1-sperms were shared with its family trio, while 36.49% were shared between Sample1-sperms and Sample2-sperms. Thus, it is possible that CNVs only called in single cells were less reliable. We suspect the following systematic factors may contribute to these discrepancies: 1) uneven whole genome amplification, 2) low sequence coverage, and 3) suboptimal pipelines and their parameters.

As expected, scDNAseq is limited by its DNA amount: a single sperm contains 3 pg of DNA, not enough for whole-genome sequencing. Therefore, scDNAseq template amplification and library preparation are needed. As shown previously [50], these steps could severely impact the performance of CNV detection when whole genome amplification is uneven and/or sequence

coverage is low. Additionally, the bioinformatics pipelines also influenced the performance of CNV detection. Ideally, read depth should be a better strategy given the low sequencing coverage, as compared to the pair-end and split-read approaches. As reviewed before [40], to correct for the first two factors, existing scDNAseq CNV read depth detection pipelines need to divide the genome into bins or windows first. They will then perform GC correction and mappability correction to obtain normalized reads depths (Figs. S3 and S4). Finally, they will need to remove outlier bins and outlier cells. The outlier bins often have an unusually high read count and occur near the centromere and telomere of each chromosome. The outlier cells often are low in signal-to-noise ratio or low in sequence coverage.

However, most of the existing pipelines are designed for the human genome [40], and it will take a great effort to fully customize and optimize them for livestock like cattle. In this study, CNVs were called using LUMPY [51], which was not designed for scDNA-seq data. It is also better to simultaneously apply a method to multiple samples to call germline/inherited CNVs to achieve better sensitivity and accuracy as the recently published method CHISEL did for human data [52]. Then CNV genotyping could then be performed on individual sperm cells. Our pipeline processed each sample separately using an integrated algorithm combining pair-end, split-read, and read depth. It did not specifically remove the outlier bins or the outlier cells, as no such data exists for cattle. However, our rationale for using LUMPY was that although we had a low average coverage and a low read depth for individual sperm genomes, we sampled the same genomes multiple times, through different sperms, with a total accumulating read depth of $56.99 \times$ and $43.68 \times$.

Therefore, merging reads across different sperms, i.e., pseudo bulk sequencing, should yield relatively confident results. In the future, we plan to adopt existing human pipelines to alleviate the impacts of these systematic factors on CNV calling in cattle.

During meiosis, chromosome missegregation can cause aneuploidy. Using Sperm-seq, Bell et al. sequenced 31,228 human sperm genomes from 20 men, identifying crossovers and other genomic anomalies [37]. They discovered that human sperm donors had aneuploidy rates ranging from 0.01 to 0.05 aneuploidies per gamete [37]. Due to the limited sample size and probably the signal-to-noise ratio, no aneuploidy was detected in this study.

Finally, Ebert et al. recently reported that over two-thirds of CNV detected within the human genome were associated with NAHR, mediated by repetitive sequences, such as segmental duplications and common repeat elements [41]. It was encouraging that our cattle segmental duplication and CNV flanking sequence analysis results also showed they are significantly enriched for each other and satellite repeats, despite the suboptimal data quality due to the abovementioned factors. In summary, we sequenced single sperms in cattle, performed an initial CNV detection, and found a distal chromosomal distribution pattern, which agreed with previous results derived from cattle bulk-cell sequencing or human studies. In the meantime, our results also highlighted the importance of the uniform whole genome amplification, deep sequence coverage, and dedicated software pipelines for CNV detection using scDNA-seq data.

Methods

Sample collection and whole genome amplification and sequencing

We chose two Holstein bulls with different fertility capabilities: Sample1 has a DPR (daughter pregnancy rate) PTA value of 0.0, reliability of 0.99, estimated from 6,528 daughters. In contrast, Sample2 has a DPR PTA value of -3.2, reliability of 0.99, estimated from 15,314 daughters. Somatic tissue (ear punch) samples of Holstein Sample1, together with its parent somatic tissues, were donated by Select Sires, Inc (Plain City, OH, USA). Semen samples were freshly collected by Select Sires, Inc. in its routine artificial insemination semen straw production. After receiving them under liquid nitrogen in USDA-ARS Animal Genomics and Improvement Laboratory (AGIL), we manually isolated a total of 156 sperm cells from two Holstein bulls (Sample1 with 73 sperm cells and Sample2 with 83 sperm cells). Briefly, isolated sperms were thawed in 37 °C water for 30–45 s and treated with 0.25% Trypsin–EDTA, followed by dilution with PBS + 1% BSA and

washing twice. The sperms were further diluted to a proper resolution using PBS + 1% BSA on a petri-dish, and active single sperms were picked up manually by pipetting into a reaction tube under a micromanipulator as described previously [42]. Whole-genome amplification was performed on single cells according to the manufacturer's protocol, using the Single Cell Whole Genome Amplification Kit (Yikon Genomics, Shanghai, China) developed from the Multiple Annealing and Looping Based Amplification Cycles (MALBAC) method [35]. In brief, a single sperm was initially analyzed and pre-amplified by primers supplied in the kit with 8 cycles with multiple annealing steps. PCR generated fragments with variable length at random starting positions for Illumina short-read sequencing. To evaluate the agreement rate of individual recombination from sperms and parents, we also sequenced the somatic diploid genomes of the trio, including Sample1 (Sample1-diploid) and its parents (Sample1-sire and Sample1-dam). Using their somatic ear punch tissues, we isolated their diploid genomes using a QIAGEN DNA extraction kit. DNA samples extracted from the donor and his parents' ear skin samples were then used to prepare sequencing libraries using standard Illumina protocol and sequenced on an Illumina HiSeq 2000/NextSeq 500 sequencing platform.

Identification of segmental duplications and enrichment test

We utilized software SDquest v0.1 [44] for detecting segmental duplications (SegDup, also known as low copy repeats) and constructing the breakpoint graph of these mosaic SegDups, based on the repeat masked ARS-UCD1.2 reference downloaded from ENSEMBL (ftp://ftp.ensembl.org/pub/release-102/fasta/bos_taurus/dna/). We compared the repeat content of SegDups, CNVs (DEL or DUP), or 5 kb, 20 kb flanking regions (5kbF, 20kbF). For CNVs, we combined the SegDups and repeats from UCSC Table Browser (<https://genome.ucsc.edu/cgi-bin/hgTables>). Length, Count, Average Length, Length%, and Count/Mb of repeat content for SegDups, CNVs, 5kbF, or 20kbF were based on these repeat overlapped with regions, Length% denotes the proportion of repeat length overlapped with SegDups/CNVs/5kbF/CNVs in total SegDups/CNVs/5kbF/20kbF length, Count/Mb denotes the count of repeats overlapped with SegDups/CNVs/5kbF/20kbF divided by total SegDups/CNVs/5kbF/20kbF Mb. For enrichment, ratios were defined as Average Length, Length%, and Count/Mb of repeats in SegDups/CNVs/5kbF/20kbF divided by repeats in the genome. We determined the significance of the enrichment by 1,000 times simulating the SegDups/CNVs/5kbF/20kbF

in random genome position with the same average and standard deviation length, which generated by function `createRandomRegions` from R v4.0.2 package `regioneR`. *P*-value refers to the frequency of simulated value larger than observed value divided by simulation times. The threshold was set as 0.05.

Structural variation detecting

We employed LUMPY v0.2.13 [51], which integrated read-depth, read-pair, and split-read strategies, to detect structural variations in high coverage sperms. As recommended, LUMPY was internally implemented in a pipeline `smoove` (<https://github.com/brentp/smoove>) with shorter run-time and lower false-positive rate. `smoove` was used to collect the best practices of LUMPY, such as generating empirical insert size statistics on each library in the BAM file, estimating the mean and standard deviation (SD) of the input parameters for LUMPY. From LUMPY, the four types of structural variations, including deletion (DEL), duplication (DUP), inversion (INV), and break end (BND), were reported for each sample. Due to the limitation for INV and BND detection, we focused on CNV (DEL plus DUP) in most of the analysis, after filtering away DEL and DUP with a length more than 5 Mb or short than 50 bp. For haploid sperms and diploid trio, we applied the following thresholds to filter out low-quality CNVs: the threshold of supporting read count for either paired-end event or split-read event must be more than 3/4 of the genome coverage, while the read count for the other type of split-read event must be more than 1 or paired-end event must be more than 3.

Gene annotation and enrichment analysis

We mapped regions of interest to the bovine reference gene annotation of the ARS-UCD1.2 genome from ENSEMBL using BEDtools v2.26.0 [53]. The gene features included transcripts, exons, CDS, 3'-UTR, 5'-UTR, start codon, and stop codon. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway and Gene Ontology (GO) enrichment were performed using the R (version 4.0.2) packages `org.Bt.eg.db` and `clusterProfiler`. We performed the quantitative trait loci (QTL) enrichment analysis using the Fisher exact test at `animalgenome.org` [46]. All enrichment *P*-values were also adjusted for multiple comparisons by Benjamini and Hochberg's (BH) algorithm.

Abbreviations

AGIL: Animal Genomics and Improvement Laboratory; CNV: Copy number variation; DPR: Daughter pregnancy rate; GWAS: Genome-wide association study; HMM: Hidden Markov model; INDEL: Short insertion and deletion; kb: Kilobase pairs; LD: Linkage disequilibrium; MALBAC: Multiple annealing and looping based amplification cycles; Mb: Megabase pairs; NAHR: Non-allelic

homologous recombination; PCR: Polymerase chain reaction; PRDM9: PR domain-containing 9; QC: Quality control; QTL: Quantitative trait loci; SD: Standard deviation; SE: Standard error; SegDup: Segmental duplication; SNP: Single nucleotide polymorphism.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08441-8>.

Additional file 1:

Additional file 2:

Acknowledgements

This research used resources provided by the SCINet project of the USDA ARS project number 0500-00093-001-00-D. We thank Reuben Anderson for his technical assistance. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The USDA is an equal opportunity provider and employer. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

GEL and LY conceived the study. LY, YG, LF, CB, MN, CL, LvY, ZY, and ES analyzed and interpreted data. LY, LM, and GEL wrote the manuscript. AO, CGS, JBC, LYX, LL, HPZ, BDR, and CPVT contributed tools and materials. All authors read and approved the final manuscript.

Funding

This work was supported in part by AFRI grant numbers 2016–67015-24886, 2019–67015-29321, 2020–67015-31398, and 2021–67015-33409 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome and Reproduction Programs and BARD grant number US-4997–17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund.

Availability of data and materials

The data that support the results of this research are available within the article and its Supplementary Information files. All other sequence data can be tracked in supplemental files. The single sperm sequencing data and the trio whole genome sequencing data were submitted to GEO under the accession number PRJNA691741 (<https://dataview.ncbi.nlm.nih.gov/object/PRJNA691741?reviewer=kj8n0f06eekt1uck7726jijms3>).

Declarations

Ethics approval and consent to participate

The need for ethics approval was waived as the current study didn't involve whole animals.

Consent for publication

Not applicable.

Competing interests

AO and CS are employees of Select Sires, Inc. GEL and LM serve on BMC Genomics Editorial Board. All other authors declare that they have no competing interests.

Author details

¹Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, MD 20705, USA. ²College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China. ³Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA. ⁴Select Sires Inc, 11740 U.S. 42 North, Plain City, OH 43064, USA. ⁵MRC Human Genetics Unit at the Institute of Genetics and Cancer, The University of Edinburgh, Edinburgh EH4 2XU, UK. ⁶Agricultural Research Organization (ARO), Institute of Animal Science, HaMaccabim Road, P.O.B 15159, 7528809 Volcani Center-Rishon LeTsiyon, Israel. ⁷Innovation Team of Cattle Genetic Breeding, Institute

of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China. ^aKey Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education & College of Animal Science and Technology, Huazhong Agricultural University, Wuhan 430070, China.

Received: 16 July 2021 Accepted: 15 October 2021

Published online: 17 March 2022

References

- Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK. Mapping copy number variation by population-scale genome sequencing. *Nature*. 2011;470(7332):59–65.
- Zhang F, Gu W, Hurler ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451–81.
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444–54.
- Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704–12.
- Consortium IH. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52–8.
- Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M. Global diversity, population stratification and selection of human copy-number variation. *Science*. 2015;349(6253):aab3761.
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y. An integrated map of structural variation in 2,504 human genomes. *Nature*. 2015;526(7571):75–81.
- Graubert TA, Caham P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM. A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet*. 2007;3(1):e3.
- She X, Cheng Z, Zöllner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. *Nat Genet*. 2008;40(7):909–14.
- Guryev V, Saar K, Adamovic T, Verheul M, Van Heesch SA, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet*. 2008;40(5):538–45.
- Chen W-K, Swartz JD, Rush LJ, Alvarez CE. Mapping DNA structural variation in dogs. *Genome Res*. 2009;19(3):500–9.
- Berglund J, Nevalainen EM, Molin A-M, Perloski M, André C, Zody MC, Sharpe T, Hite C, Lindblad-Toh K, Lohi H. Novel origins of copy number variation in the dog genome. *Genome Biol*. 2012;13(8):R73.
- Liu J, Zhang L, Xu L, Ren H, Lu J, Zhang X, Zhang S, Zhou X, Wei C, Zhao F. Analysis of copy number variations in the sheep genome using 50K SNP BeadChip array. *BMC Genomics*. 2013;14(1):229.
- Fontanesi L, Beretti F, Riggio V, Gómez GE, Dall'Olio S, Davoli R, Russo V, Portolano B. Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet Genome Res*. 2008;126(4):333–47.
- Fontanesi L, Martelli PL, Beretti F, Riggio V, Dall'Olio S, Colombo M, Casadio R, Russo V, Portolano B. An initial comparative map of copy number variations in the goat (*Capra hircus*) genome. *BMC Genomics*. 2010;11(1):639.
- Wang X, Nahashon S, Feaster TK, Bohannon-Stewart A, Adefoye N. An initial map of chromosomal segmental copy number variations in the chicken. *BMC Genomics*. 2010;11(1):351.
- Jia X, Chen S, Zhou H, Li D, Liu W, Yang N. Copy number variations identified in the chicken using a 60K SNP BeadChip. *Anim Genet*. 2013;44(3):276–84.
- Hou Y, Liu GE, Bickhart DM, Cardone MF, Wang K, Kim E-s, Matukumalli LK, Ventura M, Song J, VanRaden PM. Genomic characteristics of cattle copy number variations. *BMC Genomics*. 2011;12(1):127.
- Liu GE, Hou Y, Zhu B, Cardone MF, Jiang L, Cellamare A, Mitra A, Alexander LJ, Coutinho LL, Dell'Aquila ME. Analysis of copy number variations among diverse cattle breeds. *Genome Res*. 2010;20(5):693–703.
- Nicholas TJ, Baker C, Eichler EE, Akey JM. A high-resolution integrated map of copy number polymorphisms within and between breeds of the modern domesticated dog. *BMC Genomics*. 2011;12(1):414.
- Liu G, Tassell CV, Sonstegard T, Li R, Alexander L, Keele J, Matukumalli L, Smith T, Gasbarre L. Detection of germline and somatic copy number variations in cattle. *Dev Biol (Basel)*. 2008;132:231.
- Xu L, Hou Y, Bickhart DM, Song J, Van Tassell CP, Sonstegard TS, Liu GE. A genome-wide survey reveals a deletion polymorphism associated with resistance to gastrointestinal nematodes in Angus cattle. *Funct Integr Genomics*. 2014;14(2):333–9.
- Stothard P, Choi J-W, Basu U, Sumner-Thomson JM, Meng Y, Liao X, Moore SS. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. *BMC Genomics*. 2011;12(1):559.
- Seroussi E, Glick G, Shirak A, Yakobson E, Weller JI, Ezra E, Zeron Y. Analysis of copy loss and gain variations in Holstein cattle autosomes using BeadChip SNPs. *BMC Genomics*. 2010;11(1):673.
- Jiang L, Jiang J, Wang J, Ding X, Liu J, Zhang Q. Genome-wide identification of copy number variations in Chinese Holstein. *PLoS One*. 2012;7(11):e48732.
- Choi J-W, Lee K-T, Liao X, Stothard P, An H-S, Ahn S, Lee S, Lee S-Y, Moore SS, Kim T-H. Genome-wide copy number variation in Hanwoo, Black Angus, and Holstein cattle. *Mamm Genome*. 2013;24(3–4):151–63.
- Wu Y, Fan H, Jing S, Xia J, Chen Y, Zhang L, Gao X, Li J, Gao H, Ren H. A genome-wide scan for copy number variations using high-density single nucleotide polymorphism array in Simmental cattle. *Anim Genet*. 2015;46(3):289–98.
- Zhang Q, Ma Y, Wang X, Zhang Y, Zhao X. Identification of copy number variations in Qinchuan cattle using BovineHD Genotyping Beadchip array. *Mol Genet Genomics*. 2015;290(1):319–27.
- Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*. 2012;150(2):402–12.
- Lu S, Zong C, Fan W, Yang M, Li J, Chapman AR, Zhu P, Hu X, Xu L, Yan L, et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science*. 2012;338(6114):1627–30.
- Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublotte JT, Yosef N, et al. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*. 2014;510(7505):363–9.
- Smith GP. Evolution of repeated DNA sequences by unequal crossover. *Science*. 1976;191:528–35.
- Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, et al. Construction of a human cell landscape at single-cell level. *Nature*. 2020;581(7808):303–9.
- Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;36(1):89–94.
- Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*. 2012;338(6114):1622–6.
- Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet*. 2016;17(3):175–88.
- Bell AD, Mello CJ, Nemesh J, Brumbaugh SA, Wysoker A, McCarroll SA. Insights into variation in meiosis from 31,228 human sperm genomes. *Nature*. 2020;583(7815):259–64.
- McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*. 2017;168(4):613–28.
- Turajlic S, Sottoriva A, Graham T, Swanton C. Resolving genetic heterogeneity in cancer. *Nat Rev Genet*. 2019;20(7):404–16.
- Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biol*. 2020;21(1):208.
- Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372(6537):eabf7117.
- Zhou Y, Shen B, Jiang J, Padhi A, Park KE, Oswalt A, Sattler CG, Telugu BP, Chen H, Cole JB, et al. Construction of PRDM9 allele-specific recombination maps in cattle using large-scale pedigree analysis and genome-wide single sperm genomics. *DNA Res*. 2018;25(2):183–94.
- Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet*. 2003;73(4):823–34.

44. Pu L, Lin Y, Pevzner P. Detection and analysis of ancient segmental duplications in mammalian genomes. *Genome Res*. 2018;gr.228718.228117.
45. Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler EE. Analysis of recent segmental duplications in the bovine genome. *BMC Genomics*. 2009;10:571.
46. Hu ZL, Park CA, Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res*. 2019;47(D1):D701-d710.
47. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera AV, Lowther C, Gauthier LD, Wang H, et al. A structural variation reference for medical and population genetics. *Nature*. 2020;581(7809):444–51.
48. Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*. 2020;583(7814):83–9.
49. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, et al. Eleven grand challenges in single-cell data science. *Genome Biol*. 2020;21(1):31.
50. Mallory XF, Edrisi M, Navin N, Nakhleh L. Assessing the performance of methods for copy number aberration detection from single-cell DNA sequencing data. *PLoS Comput Biol*. 2020;16(7):e1008012.
51. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84.
52. Zaccaria S, Raphael BJ. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol*. 2021;39(2):207–14.
53. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



RESEARCH

Open Access



Genome-wide recombination map construction from single sperm sequencing in cattle

Liu Yang^{1,2†}, Yahui Gao^{1,3†}, Mingxun Li⁴, Ki-Eun Park³, Shuli Liu^{1,5}, Xiaolong Kang⁶, Mei Liu⁷, Adam Oswalt⁸, Lingzhao Fang⁹, Bhanu P. Telugu^{3,10}, Charles G. Sattler⁸, Cong-jun Li¹, John B. Cole¹, Eyal Seroussi¹¹, Lingyang Xu¹², Lv Yang¹³, Yang Zhou¹³, Li Li², Hongping Zhang², Benjamin D. Rosen¹, Curtis P. Van Tassel¹, Li Ma^{3*} and George E. Liu^{1*}

Abstract

Background: Meiotic recombination is one of the important phenomena contributing to gamete genome diversity. However, except for human and a few model organisms, it is not well studied in livestock, including cattle.

Results: To investigate their distributions in the cattle sperm genome, we sequenced 143 single sperms from two Holstein bulls. We mapped meiotic recombination events at high resolution based on phased heterozygous single nucleotide polymorphism (SNP). In the absence of evolutionary selection pressure in fertilization and survival, recombination events in sperm are enriched near distal chromosomal ends, revealing that such a pattern is intrinsic to the molecular mechanism of meiosis. Furthermore, we further validated these findings in single sperms with results derived from sequencing its family trio of diploid genomes and our previous studies of recombination in cattle.

Conclusions: To our knowledge, this is the first large-scale single sperm whole-genome sequencing effort in livestock, which provided useful information for future studies of recombination, genome instability, and male infertility.

Keywords: Cattle, Single sperm, Sequencing, Recombination

Background

Meiotic recombination promotes genetic diversity by reshuffling parental alleles and providing novel combinations of genes for evolutionary selection [1–5]. Recombination is also crucial for ensuring proper segregation of homologous chromosomes during meiosis [4]. Considerable variations in recombination rates between

individuals have been documented in human and other species [6–10].

Recombination hotspots are usually clustered into narrow genomic regions specified by the PR domain-containing 9 (*PRDM9*) gene in human and mouse [11–15]. *PRDM9* has driven evolutionary erosion of hotspots in *Mus musculus* through haplotype-specific initiation of meiotic recombination [16]. Since crossovers were disfavored at such hotspots, sequence divergence generated by hotspot turnover may create an impediment for recombination in hybrids, potentially leading to reduced fertility and thus, eventually, speciation [17, 18]. More recent publications investigated the rules governing DNA recombination, revealing the relationships between the distribution of crossovers,

*Correspondence: lima@umd.edu; George.Liu@usda.gov

†Liu Yang and Yahui Gao contributed equally to this work.

¹ Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, MD 20705, USA

³ Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA

Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

proteins involved in recombination, and specific factors determining whether a double-strand break becomes a crossover [19, 20].

Besides popular pedigree-based studies, there exist two other methods for measuring recombination based on sperm typing or linkage disequilibrium (LD) patterns. Single-sperm genomics and sperm typing can assess recombination in a regional or genome-wide [21, 22]. Using a single sperm isolation and sequencing approach, the Quake lab reported an average of 22.8 recombination events, 5 to 15 gene conversion events, as well as 25 to 36 de novo mutations in each human sperm [22]. Similarly, the Xie group reported aneuploidy in 4% of the cells and 26 recombination events per human sperm [23]. The Donnelly team later developed a method to sequence individual mouse sperm and applied it to mice carrying two different alleles of *PRDM9* in mammalian crossovers [20]. A new method called ReMIX was introduced to detect crossovers from gamete DNA using Illumina sequencing of 10X Genomics linked-read libraries in a single mouse and stickleback fish [24]. As a variation of Drop-seq [25], Sperm-seq is another high-throughput and low-cost approach to quantify recombination variation across the gamete genomes. Using Sperm-seq, Bell et al. sequenced 31,228 human sperm genomes from 20 men, identifying 813,122 crossovers and other genomic anomalies [26]. They discovered that crossover frequency and location, as well as other meiotic phenotypes like chromosome aneuploidy, vary across chromosomes, gametes, and human donors. The authors propose that inter-cell and inter-individual variation in meiotic chromosome compaction could partially explain this covariance.

Using large-scale cattle pedigree data, we have previously reported different recombination patterns between bulls and cows and identified several loci associated with recombination rate and hotspot usage in both sexes, including the *PRDM9* gene on chromosome 1 [27]. Similar results were also reported by other groups [28, 29]. In our second cattle study using single sperm genomics, we examined the allele pattern of *PRDM9* impacting cattle genome recombination [30]. Later, we also detected *Bos taurus-indicus* hybridization correlates with intralocus sexual-conflict effects of *PRDM9* on male and female fertility in Holstein cattle [31]. Here, we analyze 143 single sperm genomes from two Holstein bulls to derive two individualized recombination maps, identifying 4,291 crossovers. We further validated the reliability of single-sperm sequencing-based results, using the data derived from the diploid genome sequencing of one sample's family trio and our previous recombination studies. To our knowledge, this is the first large-scale single sperm whole-genome sequencing report in livestock, which

could facilitate future studies of recombination, genome instability, and male infertility.

Results

Sequencing and genotyping of haploid sperms and diploid trio

Sequencing for sperms

We chose two bulls with different fertility capabilities (See Methods). Using the MALBAC method [30], we successfully picked, amplified, and sequenced a total of 156 single sperm cells from two Holstein bulls' semen. After quality control filtering, we kept 143 sperm data (71 for Sample1 and 72 for Sample2) for downstream analyses. The sequenced sperms had an average genome coverage depth of $1.79 \times$, and 16 of them had genome coverage depth of $\sim 4 \times$, corresponding to an overall genome coverage of $\sim 11.40\%$ to $\sim 41.35\%$, respectively (Table S1). On average, we mapped 98.18% of sequencing reads from single sperms on the bovine ARS-UCD1.2 genome.

Genotyping for sperms

We used GATK to call the raw genotypes for SNPs and INDELs [32]. Each sperm generated raw calls for 15.5–43.0 million SNPs and 2.4–7.2 million INDELs (Table S2). Since sperms are haploid cells, we removed extensive heterozygous genotype calls. Only a small fraction of heterozygous raw calls was detected, with an average frequency of 2.46% for SNPs (ranging from 1.03% to 7.39%) and 2.97% for INDELs (ranging from 1.03% to 9.16%), respectively. These data indicated that most of the sperms were isolated successfully with low contamination before sequencing. After strict filtration, we kept approximately 4.29% SNPs (ranging from 0.42 to 2.68 million) and 11.21% INDELs (ranging from 0.23 to 1.04 million). Compared to our previous single sperm recombination analysis using the BovineHD SNP chip [30], our current study covered ~ 20 fold more clean SNPs, with an average of 1.12 million (Table S2).

Trio

For Samples1's family trio diploid genomes, we sequenced bulk DNA samples extracted from ear punches of Sample1, its sire Sample1-sire, and dam Sample1-dam to approximately $40 \times$, $10 \times$, and $20 \times$ genome coverage, respectively, with over 99% genome mapping rate and covering 96% genome sequence (Table S3). After QC filtering, we obtained approximately 5.61 million (62.89%) SNPs and 0.72 million (65.26%) INDELs of Sample1. Within them, 44.45% and 46.48% high-quality SNPs and INDELs were heterozygous, respectively (Table S4).

Individual recombination maps

Phasing

As described in Methods, assuming the low probability of crossovers between nearby SNPs, we phased the heterozygous genotypes of the bulls into haplotypes based on sperm linkage information. In 71 Sample1 sperms and 72 Sample2 sperms, a total of 310,271 and 307,451 autosomal heterozygous SNPs (htSNPs) were phased, and the phasing rates were 85.79% and 80.40%, respectively (Table 1, Table S5, and Table S6). To verify the phased haplotypes, we phased a total of 1,501,331 (79.81%) htSNPs from Sample1 using its family trio information. We used that as a scale plate to estimate the agreement rate of phased sperm alleles. Totally, 173,157 htSNPs for Sample1 were phased by either single sperm haploid genomes or Sample1 trio diploid genomes, and 95.22% (164,885) of them were consistent between alleles phased by both.

Crossover

With the phased autosomal htSNPs of Sample1 and Sample2, we inferred their crossovers occurred in the interval region of htSNPs using an HMM method, as previously described [30]. The 143 single sperms gave a total of 4,291 crossover events, on average $\sim 30.01 \pm 0.76$ standard error (SE) (9.12 SD) per sperm (Table S7). An average of ~ 32 Mb distance between two crossovers was observed on those chromosomes with double crossovers (Fig. S1). Approximately 80.3%, 64.6%, and 37.0% of the total crossovers can be confidently localized to intervals of 200, 100, and 30 kb, respectively (Fig. S2). The resolutions of our cattle recombination results were between the outcomes from two previous human studies, where their corresponding percentages were: 59%, 37%, and 13% [22] as well as 93%, 80%, and 45% [23] at those three interval thresholds, respectively.

When comparing the two Holstein bulls Sample1 and Sample2, we constructed individual recombination maps for all chromosomes, spanning 28.34 ± 1.12 SE (9.46 SD) Morgans in Sample1 and 31.65 ± 1.00 SE (8.52 SD) Morgans in Sample2, respectively (Fig. 1 and Table S8). Fewer crossovers were identified in some low htSNP density regions, for example, in runs of the homozygous region (ROH) in BTA 2, 3, 12, and 18 of Sample1 when compared to Sample2. The low htSNP density regions also

had large distances between htSNPs. When testing the relationship between the numbers of crossovers and the chromosome length, we did not find a strong correlation within these low htSNP density regions (ANOVA type III, P -values = 0.076). To control the ROH effects, we removed 75 regions covered by less than 50 htSNP per Mb of the genome for the two donors in all subsequent analyses (Fig. S3 and Table S9). As shown in Fig. 2A, after removing the low htSNP density regions, the number of crossovers on chromosomes increased with the chromosome length (Fig. S4). Besides, the individual recombination maps of Sample1 and Sample2 showed that most of the chromosomes are broadly similar, with differences found in chr2, chr3, and chr28 (Fig. 2B).

Hotspot

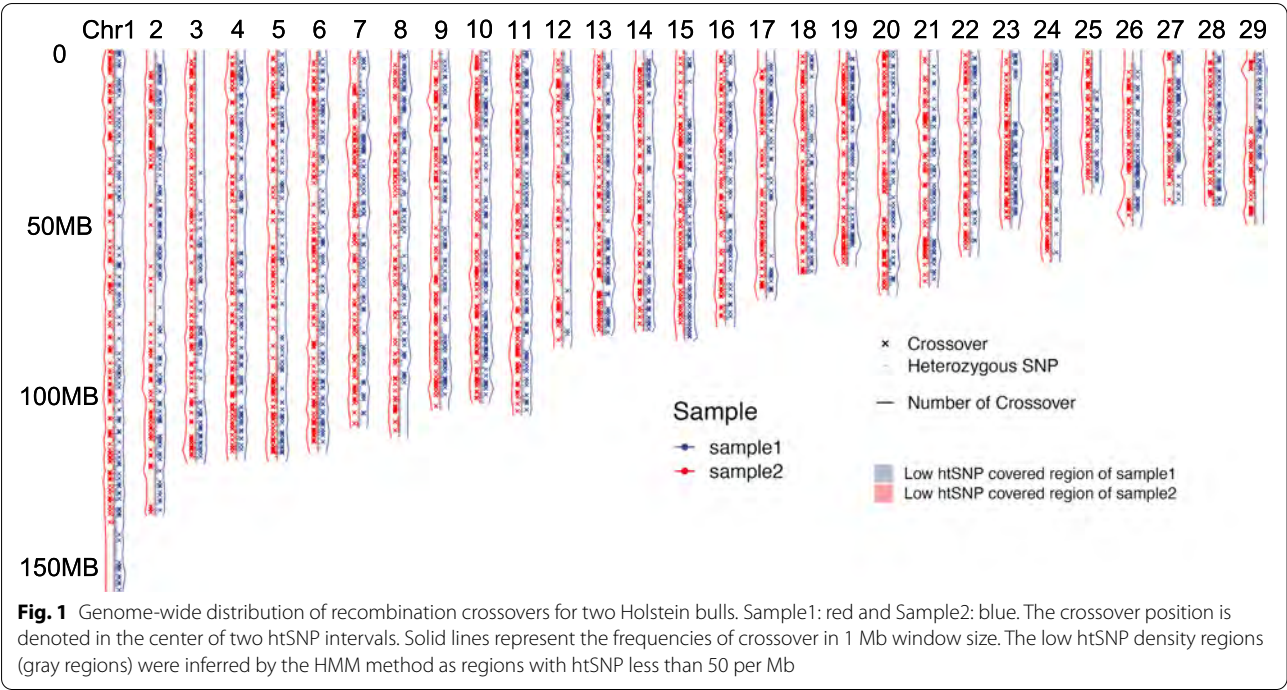
The recombination crossover locations were not uniformly distributed along the genome in these two individual bulls. We defined recombination hotspot as a short chromosomal region where crossovers occur more frequently than in other regions, as described previously [27]. In brief, we defined the recombination hotspots in these two individuals as the regions with a recombination rate of $2.5 \times$ SD greater than the mean. We detected a total of 103 (4.14% of total autosomes) hotspots in Sample1, and 41 (1.65%) hotspots in Sample2 (Table S10), with five of them, shared between the two samples (Fig. 2C). When overlapping with the bovine quantitative trait loci (QTL) database [33], these 139 hotspots were significantly enriched in 31 bovine QTL, such as non-return rate, the interval from first to last insemination, and milk-composition-related QTL (Fig. 2D and Table S11).

Compare sperm recombination maps to our earlier cattle recombination results

We also checked the consistency of recombination patterns derived from individual sperm sequencing compared to those from pedigree data [27] and individual sperm genotyping by the Illumina BovineHD BeadChip [30]. Because the pedigree data were based on SNP chips, the recombination events were usually underestimated within the first and last 5 Mb distal, i.e., terminal regions of chromosomes. After excluding these regions,

Table 1 Statistics of recombination events in sperms

Sperms	Covered htSNP	Phased SNP	Phased rate	Crossover	Morgan	SD	SE	R rate (cM/Mb)
Total	744,063	617,722	83.02%	4291	30.01	9.12	0.76	1.21
Sample1	361,651	310,271	85.79%	2012	28.34	9.46	1.12	1.14
Sample2	382,412	307,451	80.40%	2279	31.65	8.52	1.00	1.27



we converted the recombination intervals to a Mb scale, assuming 1 centimorgan or cM corresponding to 1 Mb.

Notably, the crossover hotspots were enriched in both ends of chromosomes, which corresponds to chromosomal pericentromeric and subtelomeric regions, as all bovine autosomes are acrocentric. For the same individual (Sample1), its sperm recombination maps based on sequencing or BovineHD SNP array genotyping showed a similar pattern and level, except for in the proximal regions, where sperm sequencing showed a trend of higher recombination rates (Fig. 2E). When we compared the individual sperm sequencing recombination maps (Sample1 and Sample2) to the pedigree-based population recombination map, we also detected a similar pattern (Pearson correlation coefficient of the curves between Sample1 and Sample2, Sample1 and population, and Sample2 and population are 0.677, 0.946, and 0.494 respectively, all P -values $< 2.2e-16$). But we also found that the recombination rates from single sperm sequencing were generally higher than those reported from population pedigree-based data (Fig. 2E).

Discussion

Although meiotic recombination is known to enhance genetic and phenotypic variations, it is also variable and error-prone: recombination rates vary among sperms, chromosomes, and individuals. Chromosome missegregation can cause abnormal chromosome numbers (aneuploidy), while non-allelic homologous recombination leads to over two-thirds of the structural variation detected within the human genome [34]. The purpose of this study was to probe meiotic recombination in cattle sperm.

The resolution of our cattle recombination maps is close to the previous human study [23]. The minor variances could be partially due to differences in species, platforms of whole genome amplification, quality control, and/or other factors. Given that sampling and genotype errors may potentially bias the pedigree-based results, we further confirmed our findings using the family trio diploid genome sequencing and our previous recombination study based on cattle pedigree. Our average sequencing depth is $\sim 1.79\times$, and genome coverage is from $\sim 11.40\%$

(See figure on next page.)

Fig. 2 Individual recombination maps. **A** The average number of crossovers for two samples in each chromosome. **B** Recombination maps of two samples. Accumulated relationships of the physical and genetic length of each chromosome. **C** Recombination rate per Mb in each chromosome. Red dotted lines represent thresholds of 2.5 standard deviations away from the mean genome-wide recombination rate. Chromosomes were represented in different colors. Five shared common hot spots were labeled by arrows. **D** QTL enrichment of recombination hotspots. Significance was determined by Fisher's exact test, and p -values were adjusted for multiple comparisons by the Benjamini and Hochberg's (BH) algorithm. **E** Distribution of the autosomal recombination rates over chromosomes. The curves are smoothed by the LOESS method

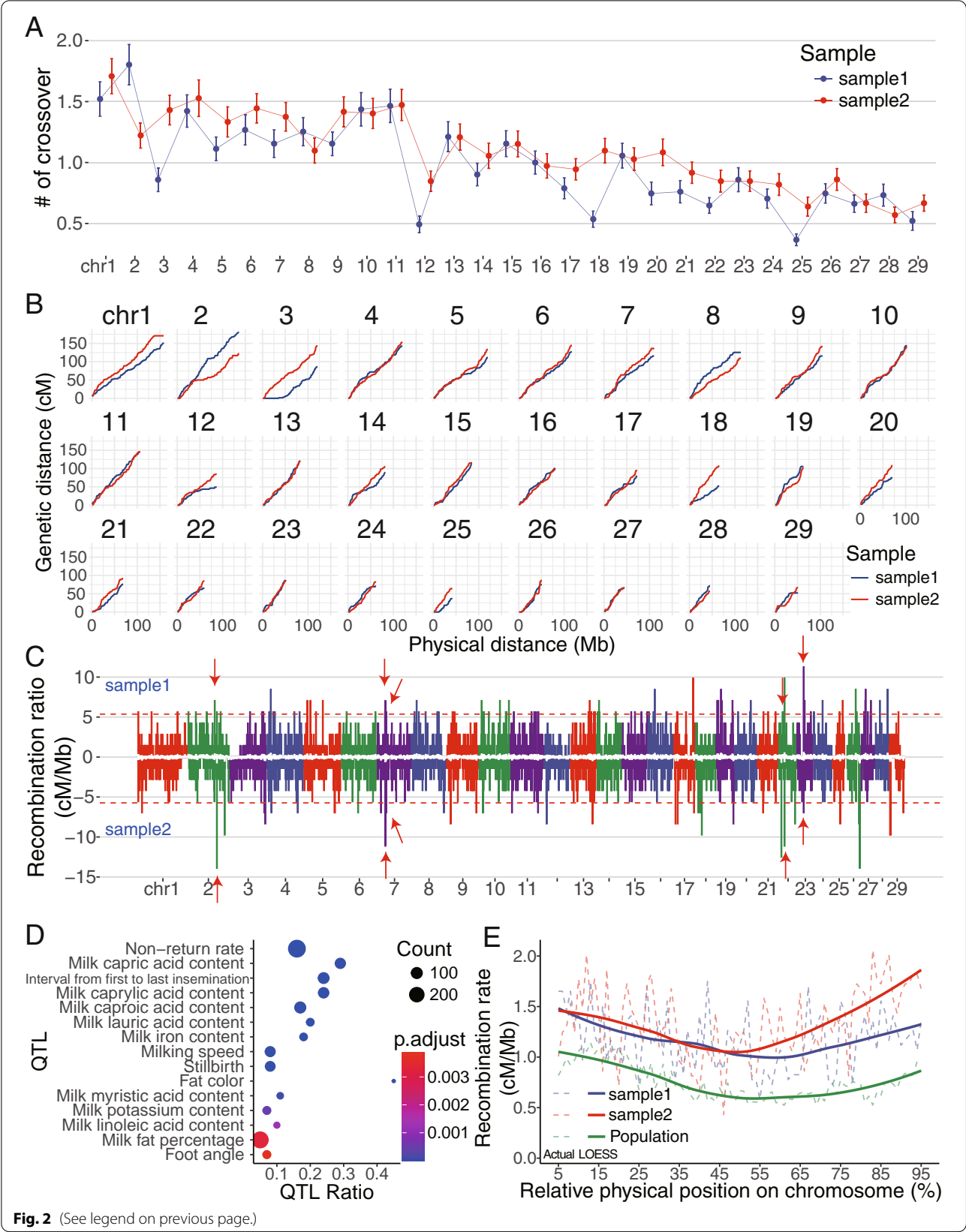


Fig. 2 (See legend on previous page.)

to ~41.35% per sperm, which are equivalent to the human study with the corresponding numbers of $\sim 1 \times$ depth and 11–44% genome coverage [23]. The Sperm-seq numbers are even lower, with $0.02 \times$ depth and 1% genome coverage [26]. Since these are typical for single sperm assays, in silico simulations or comparisons with known haplotypes were often used to verify the phasing results [23] [26]. We also sequenced the genomes from the donor's parents and used a pedigree approach to infer the phase information of the donor. We obtained 95.22% consistency, indicating the high accuracy of our approach in phasing htSNPs into chromosome-level haplotypes. In addition, the individual recombination maps of Sample1 and Sample2 showed that most of the chromosomes are broadly similar, with differences found in chr2, chr3, and chr28 (Fig. 2B). These differences also agree with previous publications, which reported that some recombination hotspots are evolving and individual-specific [35]. Interestingly, there are differences in terms of fertility traits for Sample1 and Sample2 (See Methods).

Although the genome-wide recombination distributions from these two approaches were consistent, we found the recombination rates from single sperm sequencing are generally higher than those from population pedigree-based data (Fig. 2E). These findings generally agreed with the earlier human results [22], which showed that the recombination maps from the pedigree and sperm-typing methods were largely consistent, but considerable differences were detected at a higher resolution. Because the sperms used in our study were active and viable, the differences in fitness were small between them. Therefore, different recombination patterns between sperms and live-born offspring could be caused by the selection processes during egg-sperm fertilization and embryo development till birth. Although it is intuitively unclear what factors drive such differences, based on our results and previous reports [33], we postulate the selection process between sperm-egg fertilization and embryo development to be plausible explanations. We found that a trend of higher recombination rates in the proximal regions was detected by single sperm sequencing than by the BovineHD SNP array genotyping of the same bull sperms. We partially attributed it to that sequencing could report more htSNPs than the SNP array. One limitation is that only two Holstein bulls were used in this study, so it is hard to obtain the recombination patterns within a population. The recently reported Sperm-seq will make it possible to survey more sperms in large number of samples more efficiently [26].

In conclusion, using single sperm sequencing, we investigated occurrences and distribution patterns of meiotic recombination in cattle sperm. Our results mainly agree with previous outcomes derived from population

pedigree-based data, sperm typing, and family trio diploid sequencing experiments. To our knowledge, this is the first large-scale single sperm cell sequencing report in livestock, which will further enable future studies of sperm genome instability and male infertility.

Methods

Sample collection and whole genome amplification and sequencing

We chose two Holstein bulls with different fertility capabilities: Sample1 has a DPR (daughter pregnancy rate) PTA (Predicted Transmitting Ability) value of 0.0, reliability of 0.99, estimated from 6,528 daughters. In contrast, Sample2 has a DPR PTA value of -3.2, reliability of 0.99, estimated from 15,314 daughters. Their pedigree relationship is 0.127 and the genomic relationship is 0.08, which are close to the relationship of cousins. Both are heterozygous for PRMD9 locus (allele 5/non allele 5). They were chosen based on their contrasting daughter pregnancy rates. Somatic tissue (ear punch) samples of Holstein Sample1, together with its parent somatic tissues, were donated by Select Sires, Inc (Plain City, OH, USA). Semen samples were freshly collected by Select Sires, Inc. in its routine artificial insemination semen straw production. After receiving them under liquid nitrogen in USDA-ARS Animal Genomics and Improvement Laboratory (AGIL), we manually isolated a total of 156 sperm cells from two Holstein bulls (Sample1 with 73 sperm cells and Sample2 with 83 sperm cells). Briefly, isolated sperms were thawed in 37 °C water for 30–45 s and treated with 0.25% Trypsin–EDTA, followed by dilution with PBS + 1% BSA and washing twice. The sperms were further diluted to a proper resolution using PBS + 1% BSA on a petri-dish. Active single sperms were picked up manually by pipetting into a reaction tube under a micromanipulator described previously [30]. Whole-genome amplification was performed on single cells according to the manufacturer's protocol, using the Single Cell Whole Genome Amplification Kit developed from the Multiple Annealing and Looping Based Amplification Cycles (MALBAC, Yikon Genomics, Shanghai, China) method [36]. In brief, a single sperm was initially analyzed and pre-amplified by primers supplied in the kit with 8 cycles with multiple annealing steps. PCR generated fragments with variable lengths at random starting positions for next-generation sequencing. To evaluate the agreement rate of individual recombination from sperms and parents, we also sequenced the somatic diploid genomes of the trio, including Sample1 (Sample1-diploid) and its parents (Sample1-sire and Sample1-dam). Using their somatic ear punch tissues, we isolated their diploid genomes using a QIAGEN QIAamp DNA Mini Kit protocol (QIAGEN, Valencia, CA, USA).

DNA extracted from the ear skin samples of the donor and his parents was then used for preparing sequencing libraries using standard Illumina TruSeq Library Prep Kit and sequenced on an Illumina HiSeq 2000/NextSeq 500 sequencing platform with read length of PE150 (Illumina, San Diego, CA).

Genotype calling

Paired-end sequencing reads for single sperm, and diploid samples were quality controlled by fastqc v0.11.9 and trimmed by Trimmomatic v0.39 [32]. Bwa v0.7.17 mem was used with default parameters to align clean reads against the bovine reference genome ARS-UCD1.2 (ftp://ftp.ensembl.org/pub/release-99/fasta/bos_taurus/dna/Bos_taurus.ARS-UCD1.2.dna.toplevel.fa.gz). To avoid potential PCR or sequencing optical artifacts, we marked duplicated reads that were mapped to the same location by MarkDuplicates function in GATK v4.0.8.1 [32]. FixMateInformation was also employed to ensure all mate-pair information is in sync between each read and its mate-pair. For detecting systematic errors made by the sequencing machine, Base Quality Score Recalibration (BQSR) was called for each BAM by BaseRecalibrator and ApplyBQSR with the known single nucleotide polymorphism (SNP) file from 1000 Bull Genomes Projects (<http://www.1000bullgenomes.com/>) [32]. HaplotypeCaller in GATK was used to call variants, and the parameter -ERC GVCF in CombineGVCFs was set for data combining and then performed by GenotypeGVCFs [32]. We separated SNPs and INDELs (short insertion and deletion) in a combined VCF file using the function SelectVariants, respectively.

Filtration of SNPs, INDELs, and samples

To improve the genotyping accuracy for single sperms, we applied a stringent cutoff on the raw genotyping quality score to call genotypes [32]. We removed low-quality variants with quality by depth (QD) < 2, Fisher strand (FS) > 30, strand odds ratio (SOR) > 3, root mean square of the mapping quality (MQ) < 40, and quality score (QUAL) < 40. Using the VariantFiltration function in GATK, we defined the window size as 35 to evaluate clustered SNPs and allowed three SNPs to make up a cluster. For sperm data, we kept variants with at least 2 allele support reads and removed heterozygous (0/1) SNPs or INDELs because it was potentially caused by sequencing error or sperm chromosome-scale genomic anomalies [26]. As a result, 12 sperm samples were removed as their read depth was lower than 0.5X (10 sperms) or genome coverage rate lower than 10% (2 sperms). In addition, for diploid data, we filtered those variants with allele support reads less than 1/2 genome-wide depth [32].

Inferring haplotype with sperm

We used two different genotypes—reference allele (0) and first alternate allele (1) in sperms to infer haplotypes. To avoid large numbers of unbalances between these two alleles, we only kept those sites with the minimum frequency of 30% for either allele with at least two supporting sperms. Based on sperm linkage information, we inferred haplotypes using the previously published two-stage method [23], with some modifications for our strict filtration parameters. First, we constructed a haplotype profile using a fraction (10%) of htSNPs covered by more than 20 sperm SNPs. Based on genome coordinates, we linked every two neighboring htSNPs and generated four potential combinations. As the rates of false SNP calling and recombination are low, the true links will appear much more frequently than the false links based on the frequency of neighboring htSNP pairs in all sperm data. We defined two true links that appeared eight times and two false links that occurred no more than once for a neighboring htSNP pair. If data were not satisfying these criteria, the first htSNPs would be linked to the next htSNPs until the true links appear eight times. The htSNPs satisfying these criteria were phased into one of the two haplotypes. We then imputed missing htSNPs into the haplotypes. In each sliding window of five phased htSNPs sorted by genome coordinate, those missing htSNPs were imputed recursively into either haplotype if one sperm cell had at least three confirmed phased htSNPs. To improve the phasing rate, we further imputed the remaining genotypes by borrowing information across sperm cells. We selected the top 10 sperms sorted by the genotype concordance rate with either phased haplotype. The sperm with missing htSNPs were imputed into a haplotype if two or more sperms covered this haplotype, and this haplotype had a larger number of sperm cell counts than the other haplotype. This imputation was performed for both haplotypes. After these two stages, over 80% of the htSNPs were phased into chromosome-level haplotypes for both bulls.

Phasing haplotype by Sample1 trio information

To estimate the agreement of phased haplotype of single sperms, we also sequenced the diploid genome of Sample1 and its parents. In genetics, diploid genotypes include one paternal allele and one from maternal in normal conditions, and the mutation rate is very low. Based on SNP linkage information, we phased the heterozygous genotype of Sample1 to paternal haplotype and maternal haplotype. For example, assuming the heterozygous genotype of offspring is 'AG'. Three conditions can phase 'A' into paternal haplotype and 'G' into maternal haplotype: the father's genotype is 'AA' and mother's genotype is

'GG' at this SNP; the father's is 'AG' and mother's is 'GG'; or the father's is 'AA' and mother's is 'AG'.

Inferring crossover in single sperms

The Viterbi algorithm in a Hidden Markov Model (HMM) were applied to infer the most likely states of sequence along the genome based on phased htSNPs of single sperms [20]. A crossover event occurred in the transition of a window between two htSNPs. For each chromosome of sperms, we randomly transformed a haplotype as paternal and the other one as maternal. One sample with abnormal numbers of crossovers was excluded. To avoid the genetic background, such as runs of the homozygous region (ROH) influencing the comparison of individual recombination patterns, we applied the HMM method for excluding the low htSNP density region with htSNP less than 50 per Mb across sperms of two samples.

Abbreviations

AGIL: Animal Genomics and Improvement Laboratory; DPR: Daughter pregnancy rate; GWAS: Genome-wide association study; HMM: Hidden Markov model; htSNP: Heterozygous SNP; INDEL: Short insertion and deletion; kb: Kilobase pairs; LD: Linkage disequilibrium; MALBAC: Multiple annealing and looping based amplification cycles; Mb: Megabase pairs; PRDM9: PR domain-containing 9; QC: Quality control; QTL: Quantitative trait loci; SD: Standard deviation; SE: Standard error; SNP: Single nucleotide polymorphism.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-022-08415-w>.

Additional file 1.

Additional file 2.

Acknowledgements

We thank Reuben Anderson for his technical assistance. We thank US dairy producers for providing phenotypic, genomic, and pedigree data through the Council on Dairy Cattle Breeding (Bowie, MD) under ARS-USDA Material Transfer Research Agreement 58-8042-8-007. We also thank the Cooperative Dairy DNA Repository (Columbia, MO) for providing the data used in this study. Access to 1000 Bull Genomes Project data was provided under ARS-USDA Data Transfer Agreement 15443. International genetic evaluations were calculated by the International Bull Evaluation Service (Interbull; Uppsala, Sweden). Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. The USDA is an equal opportunity provider and employer.

Authors' contributions

GEL and LM conceived the study. LY, YG, MiL, SL, XK, MeL, LF, CL, LvY, ZY, and ES analyzed and interpreted data. LY, LM, and GEL wrote the manuscript. KP, AO, BPT, CGS, JBC, LYX, LL, HPZ, BDR, and CPVT contributed tools and materials. All authors read and approved the final manuscript.

Funding

This work was supported in part by AFRI grant numbers 2016–67015-24886, 2019–67015-29321, 2020–67015-31398, and 2021–67015-33409 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome and Reproduction Programs and BARD grant number US-4997–17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund. JBC and GEL were also supported by appropriated projects 1265–31000-096–00,

"Improving Genetic Predictions in Dairy Animals Using Phenotypic and Genomic Information", and 8042–31000-104–00, "Enhancing Genetic Merit of Ruminants Through Genome Selection and Analysis", of the Agricultural Research Service of the United States Department of Agriculture, respectively. This research used resources provided by the SCINet project of the USDA ARS project number 0500–00093-001–00-D. The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The data that support the results of this research are available within the article and its Supplementary Information files. All other sequence data can be tracked in supplemental files. The single sperm sequencing data were submitted to GEO under the accession number PRJNA691741 (<https://dataview.ncbi.nlm.nih.gov/object/PRJNA691741?reviewer=kj8n0f06eekt1uck7726jjjms3>).

Declarations

Ethics approval and consent to participate

The need for ethics approval was waived as the current study didn't involve whole animals.

Consent for publication

Not applicable.

Competing interests

AO and CS are employees of Select Sires, Inc. All other authors declare that they have no competing interests.

Author details

¹Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural Research Center, Agricultural Research Service, USDA, Beltsville, MD 20705, USA. ²College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, China. ³Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA. ⁴College of Animal Science and Technology, Yangzhou University, Yangzhou 225009, China. ⁵College of Animal Science and Technology, China Agricultural University, Beijing 100193, China. ⁶College of Agriculture, Ningxia University, Yinchuan 750021, China. ⁷Animal Nutritional Genome and Germplasm Innovation Research Center, College of Animal Science and Technology, Hunan Agricultural University, Changsha 410128, China. ⁸Select Sires Inc, 11740 U.S. 42 North, Plain City, OH 43064, USA. ⁹The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, UK. ¹⁰Division of Animal Sciences, University of Missouri, Columbia, MO 65201, USA. ¹¹Agricultural Research Organization (ARO), Volcani Center, Institute of Animal Science, P.O.B 15159, HaMaccabim Road, 7528809 Rishon LeTsiyon, Israel. ¹²Innovation Team of Cattle Genetic Breeding, Institute of Animal Sciences, Chinese Academy of Agricultural Sciences, Beijing 100193, China. ¹³Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education & College of Animal Science and Technology, Huazhong Agricultural University, Wuhan 430070, China.

Received: 29 October 2021 Accepted: 24 February 2022

Published online: 05 March 2022

References

- Barton NH, Charlesworth B. Why sex and recombination? *Science*. 1998;281(5385):1986–90.
- Stumpf MP, McVean GA. Estimating recombination rates from population-genetic data. *Nat Rev Genet*. 2003;4(12):959–68.
- Kauppi L, Jeffreys AJ, Keeney S. Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet*. 2004;5(6):413–24.
- Coop G, Przeworski M. An evolutionary view of human recombination. *Nat Rev Genet*. 2006;8(1):23–34.
- Paigen K, Petkov P. Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*. 2010;11(3):221–33.
- Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT. Fine-scale

- recombination rate differences between sexes, populations and individuals. *Nature*. 2010;467(7319):1099–103.
7. Shifman S, Bell JT, Copley RR, Taylor MS, Williams RW, Mott R, Flint J. A high-resolution single nucleotide polymorphism genetic map of the mouse genome. *PLoS biology*. 2006;4(12):e395.
 8. Hunter CM, Huang W, Mackay TF, Singh ND. The genetic architecture of natural variation in recombination rate in *Drosophila melanogaster*. *PLoS Genet*. 2016;12(4):e1005951.
 9. Nachman MW, Payseur BA. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Phil Trans R Soc B*. 2012;367(1587):409–21.
 10. Balcova M, Faltusova B, Gergelits V, Bhattacharyya T, Mihola O, Trachtulec Z, Knopf C, Fotopulosova V, Chvatalova I, Gregorova S. Hybrid sterility locus on chromosome X controls meiotic recombination rate in mouse. *PLoS genetics*. 2016;12(4):e1005906.
 11. Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. *Science*. 2010;327(5967):835–835.
 12. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, De Massy B. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*. 2010;327(5967):836–40.
 13. Berg IL, Rita N, Lam KWG, Shriparna S, Linda OH, May CA, Jeffreys AJ. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet*. 2010;42(10):859–63.
 14. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*. 2010;327(5967):876–9.
 15. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD. Recombination initiation maps of individual human genomes. *Science*. 2014;346(6211):1256442.
 16. Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, Petkov PM, Paigen K. PRDM9 drives evolutionary erosion of hotspots in *Mus musculus* through haplotype-specific initiation of meiotic recombination. *PLoS genetics*. 2015;11(1):e1004916.
 17. Payseur BA. Genetic Links between Recombination and Speciation. *PLoS Genet*. 2016;12(6):e1006066.
 18. Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R. Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature*. 2016;530(7589):171–6.
 19. Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nat Commun*. 2019;10(1):3900.
 20. Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, Bowden R, Donnelly P. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science*. 2019;363(6433):eaau8861.
 21. Hubert R, MacDonald M, Gusella J, Arnheim N. High resolution localization of recombination hot spots using sperm typing. *Nat Genet*. 1994;7(3):420–4.
 22. Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*. 2012;150(2):402–12.
 23. Lu S, Zong C, Fan W, Yang M, Li J, Chapman AR, Zhu P, Hu X, Xu L, Yan L, et al. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science*. 2012;338(6114):1627–30.
 24. Dréau A, Venu V, Avdievich E, Gaspar L, Jones FC. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun*. 2019;10(1):4309.
 25. Macosko Evan Z, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas Allison R, Kamitaki N, Martersteck Emily M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. 2015;161(5):1202–14.
 26. Bell AD, Mello CJ, Nemesh J, Brumbaugh SA, Wysoker A, McCarroll SA. Insights into variation in meiosis from 31,228 human sperm genomes. *Nature*. 2020;583(7815):259–64.
 27. Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE, et al. Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS Genet*. 2015;11(1):e1005387.
 28. Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M. Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS Genet*. 2012;8(7):e1002854.
 29. Kadri NK, Harland C, Faux P, Cambisano N, Karim L, Coppieters W, Fritz S, Mullaart E, Baurain D, Boichard D, et al. Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle. *Genome Res*. 2016;26(10):1323–32.
 30. Zhou Y, Shen B, Jiang J, Padhi A, Park KE, Oswalt A, Sattler CG, Telugu BP, Chen H, Cole JB, et al. Construction of PRDM9 allele-specific recombination maps in cattle using large-scale pedigree analysis and genome-wide single sperm genomics. *DNA Res*. 2018;25(2):183–94.
 31. Seroussi E, Shirak A, Gershoni M, Ezra E, de Abreu Santos DJ, Ma L, Liu GE. Bos taurus-indicus hybridization correlates with intralocus sexual-conflict effects of PRDM9 on male and female fertility in Holstein cattle. *BMC Genet*. 2019;20(1):71.
 32. Yang L, Gao Y, Boschiero C, Li L, Zhang H, Ma L, Liu GE. Insights from Initial Variant Detection by Sequencing Single Sperm in Cattle. *Dairy*. 2021;2(4):649–57.
 33. Hu ZL, Park CA, Reedy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res*. 2019;47(D1):D701–d710.
 34. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*. 2021;372(6537):eabf7117.
 35. Jeffreys AJ, Neumann R. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*. 2002;31(3):267–71.
 36. Zong C, Lu S, Chapman AR, Xie XS. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*. 2012;338(6114):1622–6.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

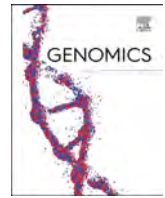
Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions





Functional annotation of regulatory elements in cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning

Yahui Gao^{a,b}, Shuli Liu^{a,c}, Ransom L. Baldwin VI^a, Erin E. Connor^d, John B. Cole^a, Li Ma^b, Lingzhao Fang^{e,*}, Cong-jun Li^{a,*}, George E. Liu^{a,*}

^a Animal Genomics and Improvement Laboratory, BARC, Agricultural Research Service, USDA, Beltsville, MD 20705, USA

^b Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA

^c College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

^d Department of Animal and Food Sciences, University of Delaware, Newark, DE 19716, USA

^e MRC Human Genetics Unit at the Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom

ARTICLE INFO

Keywords:

Cattle genome
Functional annotation
Rumen development
Chromatin state
Cell interaction
Butyrate

ABSTRACT

We profiled landscapes of bovine regulatory elements and explored dynamic changes of chromatin states in rumen development during weaning. The regulatory elements (15 chromatin states) and their coordinated activities in cattle were defined through genome-wide profiling of four histone modifications, CTCF-binding, DNA accessibility, DNA methylation, and transcriptome in rumen epithelial tissues. Each chromatin state presented specific enrichment for sequence ontology, methylation, trait-associated variants, transcription, gene expression-associated variants, selection signatures, and evolutionarily conserved elements. During weaning, weak enhancers and flanking active transcriptional start sites (TSS) were the most dynamic chromatin states and occurred in tandem with significant variations in gene expression and DNA methylation, significantly associated with stature, production, and reproduction economic traits. By comparing with *in vitro* cultured epithelial cells and *in vivo* rumen tissues, we showed the commonness and uniqueness of these results, especially the roles of cell interactions and mitochondrial activities in tissue development.

1. Introduction

Mapping chromatin accessibility and epigenomic marks have developed as a robust process to annotate genomes, identify putative regulatory elements, and study their changing activity across different cell types, developmental stages, and complex phenotypes [1–4]. The functional annotation of genomes has been well investigated in diverse

tissues and cell types in human and model organisms. However, we are still inadequate in livestock genomes' functional annotation, encumbering our interpretation of complex trait variation, domestication, and adaptive evolution. Exploring the global regulatory elements of genomes in livestock enlightens basic biology and enhances genomic improvement programs [5–7]. To generate normal tissue maps of functional elements in livestock genomes, an international collaborative

Abbreviations: ATAC, Assay for Transposase-Accessible Chromatin; AW, after weaning; BW, before weaning; BivFlnk, flanking bivalent TSS/enhancer; BT, butyrate treatment; CO, control; DEG, differentially expressed gene; EnhA, active enhancer; EnhAATAC, active enhancer with ATAC; EnhWk, weak active enhancer; EnhPois, poised enhancer; EnhPoisATAC, poised enhancer with ATAC; eQTL, expression quantitative traits loci; FAANG, Functional Annotation of Animal Genomes; GERP, Genomic Evolutionary Rate Profiling; GWAS, Genome-Wide Association study; HDAC, histone deacetylase; MAD, median absolute deviation; MDBK, Madin-Darby Bovine Kidney Epithelial Cell; MDS, multi-dimensional scaling; Quies, quiescent; REPC, Rumen Epithelial Primary Cell; ReprWkCTCF, weak repressive with CTCF; ReprPC, repressive Polycomb; SCS, somatic cell score; TES, transcriptional end site; TPM, Transcripts Per Kilobase Million; TSS, transcriptional start site; TssA, active TSS; TssAATACCTCF, active TSS with ATAC and CTCF; TssAFlnk, Flanking active TSS; TxFlnk, Transcribed at gene 5' and 3'; WGBS, whole-genome bisulfite sequencing.

* Corresponding authors at: Animal Genomics and Improvement Laboratory, USDA-ARS, Building 306, Room 216, BARC-East, Beltsville, MD 20705, USA.

E-mail addresses: Ransom.Baldwin@usda.gov (R.L. Baldwin VI), eeconnor@udel.edu (E.E. Connor), lima@umd.edu (L. Ma), Lingzhao.fang@igmm.ed.ac.uk (L. Fang), Congjun.Li@usda.gov (C.-j. Li), George.liu@usda.gov (G.E. Liu).

<https://doi.org/10.1016/j.ygeno.2022.110296>

Received 6 May 2021; Received in revised form 20 December 2021; Accepted 1 February 2022

Available online 8 February 2022

0888-7543/© 2022 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

effort, the Functional Annotation of Animal Genomes (FAANG) project, was initiated in 2015 [8].

Cattle, one of the critical species of ruminant animals, have had significant roles in several industries worldwide for centuries. Ruminants are herbivorous mammals that can utilize plant-based feed to obtain nutrients, principally through microbial fermentation. Fermentation is essential to break down complex carbohydrates, such as cellulose, to produce short-chain volatile fatty acids as the utilizable nutrition elements. The rumen is the primary site of microbial fermentation. Cattle begin life as simple stomach animals yet spend most of their lives as ruminants whose digestion depends largely on fermentation [9]. The changes from one digestive method to another require rumen development. Rumen development begins when calves start eating solid feeds that enter the rumen during weaning. Proper rumen development is essential for successfully transitioning from a milk-based diet to a diet of grain and forages with significant economic consequences.

Because rumen development is a crucial feature regulating early solid feed intake of weanlings, growth performance, and cattle feed efficiency, the study of rumen development has attracted ample attention [9–11]. Due to its unique characterization, rumen development also presents a genuine prospect and a model to study development and adaptive evolution in general. However, the genomic activities controlling ruminal morphological and physiological transformations arising during this critical time in rumen development remain incompletely characterized. In particular, system-wide analysis of the underlying regulatory dynamics of the chromatin states is lacking.

Cell adhesion and communication are essential for tissue development and differentiation [12]. Cell type, number, and spacing are also critical to tissue structure and function. Adhesive molecules help maintain communication among cells and the extracellular matrix to preserve proper tissue architecture. Within tissues, adhesive molecules, like integrins, allow cells to maintain contact with one another and structures in the extracellular matrix. Integrins link the actin cytoskeleton of a cell to various external structures, controlling cell shape and motility [13]. Cell-to-cell junctions are crucial for surfaces like the skin, intestines, and airways. For example, within gastrointestinal tracts, the side surfaces of epithelial cells are tightly linked to neighboring cells, forming a sheet that acts as a barrier. Through integrins, each cell's basal end connects to a specialized layer of extracellular matrix - the basal lamina. The adhesive transmembrane proteins interact with similar proteins on adjacent cells and the intracellular cytoskeleton through these junctions. For instance, adaptor complexes bind adherens junctions to cytoskeletal actin [14], and other adaptor complexes bind desmosomes to intermediate filaments [15]. These junctional complexes provide cells and tissues with mechanical support, and they also recruit intracellular signaling molecules to relay positional information to the nucleus. The lateral surfaces of epithelial cells also contain several other specialized junctions, including tight and gap junctions [16]. Gap junctions permit small molecules and ions to move across, thus providing metabolic and electrical coupling between cells [13]. Apoptosis is an essential aspect of development. Cell signaling also plays an essential role in the balance between cell growth and death [17].

In a previous report [18], we conducted the first attempt to establish the global map of regulatory elements (15 chromatin states) in cattle using an *in vitro* cell culture system. We defined the coordinated activities of the cattle genome's regulatory elements by mapping chromatin accessibility and epigenomic marks in rumen epithelial primary cells (REPC) and an established cell line (Madin-Darby Bovine Kidney Epithelial Cells, MDBK cell line). We also explored the dynamics of chromatin states in rumen epithelial cells *in vitro* induced by butyrate, one of the short-chain volatile fatty acids produced by rumen fermentation and a key regulator for rumen development [19,20].

To further refine the global landscape of genomic regulatory elements and explore the regulatory dynamics of chromatin states in rumen development *in vivo* during weaning, we report here a system-wide

analysis of the underlying regulatory dynamics of the chromatin states of rumen epithelial tissue from calves before and after weaning. We profiled genome-wide data sets in parallel at high resolution for four histone modifications (H3K4me3, H3K4me1, H3K27ac, and H3K27me3), DNA accessibility (Assay for Transposase-Accessible Chromatin using sequencing - ATAC-seq), and CTCF-binding sites. Additionally, we profiled the RNA-transcriptome of rumen epithelial tissue and DNA methylation by whole-genome bisulfite sequencing (WGBS) from rumen tissues to explore changes in gene expression and DNA methylation. By integrating epigenomic marks with other genome-wide data sets, including sequence ontology, gene expression, DNA methylation, transcription factors, evolutionary conservation elements, regulatory motif instances, genome-wide association study (GWAS) in large scale signals of 45 complex traits, cattle QTLdb, expression quantitative trait loci (eQTLs), and selection signatures in cattle, we were able to systematically and functionally define and characterize 15 chromatin states in cattle rumen epithelial tissue. We demonstrated that active transcription start sites (TssAs) are a hotspot for transcription regulatory factors and that highly expressed genes require a complex regulatory mechanism to ensure their proper function. We also demonstrated that weaning-induced dynamics in chromatin states, gene expression, and DNA methylation are closely correlated to rumen development, and cell interactions are vital to maintaining rumen tissue architecture, function, and development. By comparing evolutionary conservation regulatory elements between humans and cattle, we explored the role of functional annotation for understanding adaptive evolution.

2. Results and discussion

2.1. Characteristics of histone modification, DNA methylation, and transcriptomic data

We created the first global *in vivo* landscape of regulatory elements in cattle and explored the dynamics of chromatin states in rumen tissue development before and after weaning (BW and AW in Fig. 1a). Previously, we reported similar *in vitro* efforts [18], using REPC (CO) and their artificial inductions by butyrate treatment (BT). We produced 14 new genome-wide data sets of four histone modifications (*i.e.*, H3K4me3, H3K4me1, H3K27ac, and H3K27me3) at high resolution, ATAC-seq for DNA accessibility, and CTCF-binding sites (Fig. S1a), producing a total of 704,477,066 clean paired-end reads with an average uniquely mapping rate of 69.22%. Furthermore, we outlined six RNA-seq data sets and two WGBS data sets from the same rumen tissues to explore changes in gene expression and DNA methylation before and after weaning (Fig. S1a), producing a total of 115,288,544 (the average uniquely mapped rate of 93.32%) and 696,471,452 (43.45%) clean paired-end reads, respectively. Details of summary statistics for all 22 new data sets, and the control sample, are described in Table S1.

For all 28 epigenomic data sets, as shown in Fig. S1b, we obtained a total of 1,187,532 peaks with an average of 42,412 (ranging from 15,098 for H3K27me3 in BT to 98,962 for ATAC in CO). Overall, we obtained more peaks from the primary cells (*i.e.*, REPC) than rumen tissues. The transcription and epigenome profile at the tissue level reflects the integration of data from all cell types within the sample, which may cause differences in the sensitivity of measuring epigenomic markers in actual tissues and cells due to cellular heterogeneity, as our previous study showed a similar pattern for three other distinct histone modifications (H3K27ac, H3K9ac, and H3K9me3) [18]. The corresponding genome peak coverage in each sample had an average of 2.07% (ranging from 0.20% for ATAC in rumen tissue before weaning to 11.87% for H3K27me3 in REPC following BT). Also, we observed several notable high-level features of the data series. As expected, the landscape of histone modifications varies between tissues and cells, particularly for marks of activity such as H3K27ac and H3K27me3 (Fig. S1c). Within tissues or cells, chromatin landscapes changed

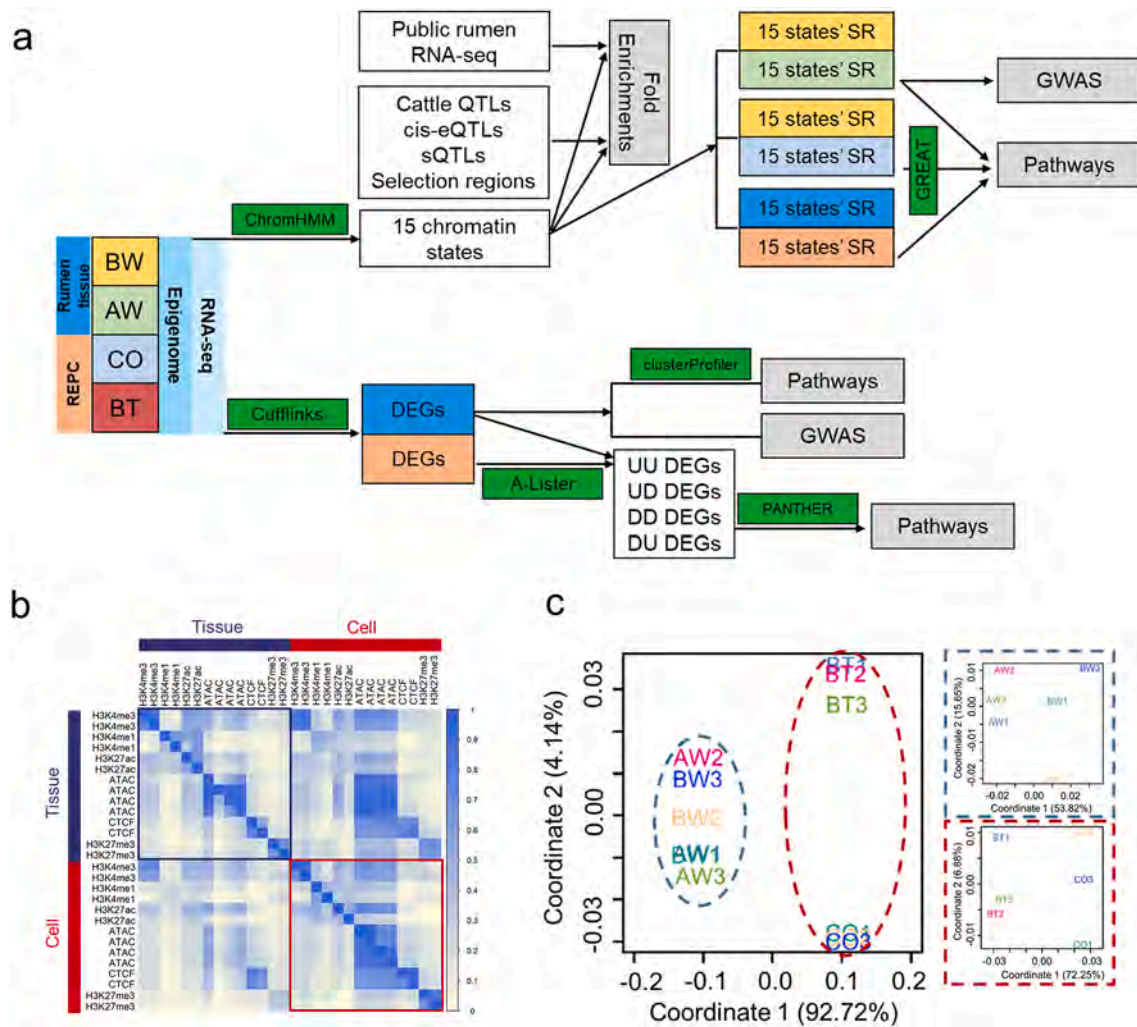


Fig. 1. Overview of the profiling dataset. (a) The workflow used in this study. SR, specific regions. (b) Correlations among epigenomic data sets across the rumen tissue and REPC. (c) Sample similarity clustering based on gene expression values for all genes. MDS clustering was performed using distance = 1 – Spearman correlation.

progressively across stages (Fig. S1c). These developmental dynamics are likely to reflect two underlying biological processes: changes in the epigenetic landscape within tissues or cells as they undergo differentiation, and weaning and butyrate treatment could induce relatively different changes. DNA methylation landscapes showed similar patterns with minor differences in BW and AW rumen tissues (Fig. S1d). Meta-gene plot showed the typical patterns of histone modification enrichment at all genes (Fig. S1e), suggesting a similar trend among marks. We also observed that ATAC was associated with CTCF and active histone modifications (e.g., H3K4me1, H3K4me3, and H3K27ac) in both rumen tissues and REPC (Fig. 1b, Fig. S1f), demonstrating the rumen tissues and primary cells shared epigenomic modification similarities in general.

On the other hand, RNA-seq triplicate results also showed a similar pattern among BW, AW, CO, and BT samples, in terms of gene expressions (Transcripts Per Kilobase Million - TPM). When we clustered them using the multi-dimensional scaling (MDS) plot based on TPM values (Fig. 1c), the primary separation was between rumen tissues and REPC. When we clustered rumen tissues and REPC separately, the primary separation was due to weaning or butyrate treatment, respectively, suggesting that either of them is the most crucial determinant during the developmental differentiation process.

2.2. Characterization and systematic definition of 15 chromatin states in cattle rumen

We used ChromHMM v1.20 [21] to define 15 states jointly from the 28 cattle rumen epigenomes, all of which were completed with six chromatin marks (Fig. 2a-d). Three of our predicted states were proximal to active TSSs (TssA, TssAATACCTCF, and TssAFlnk, approximately 1.04% of the entire genome); one state was associated with actively transcribed genes (TxFlnk, 0.4%); six states were enhancer-related (EnhA, EnhAATAC, EnhWk, EnhPois, EnhPoisATAC, and EnhWkCTCF-FATAC; 5.01%); one bivalent state often was located near inactive TSS or Enh (BivFlnk, 0.47%); one state was repressive (ReprPC, 3.40%), and another state was quiescent (Quies, 87.08%). The remaining 2.58% of the genome was assigned to ATAC or ReprWkCTCF. The first four states were distinguished by a high occurrence of H3K4me3 in high enrichments near promoter regions (± 1 kb around TSS), protein-coding regions, zinc finger genes, transcription factors [22], and expressed genes (TPM ≥ 0.1), but not repressed genes (TPM < 0.1) (Fig. 2e). TssA also displays a symptomatically high enrichment for CpG islands, parallel to a low level of DNA methylation (Fig. 2f), increasing the expression of nearby genes and validating the well-known adverse correlation between promoter methylation and gene expression [23]. Meanwhile, TssAFlnk and TxFlnk exhibited high methylation levels, again consistent with high DNA methylation levels of gene bodies being positively

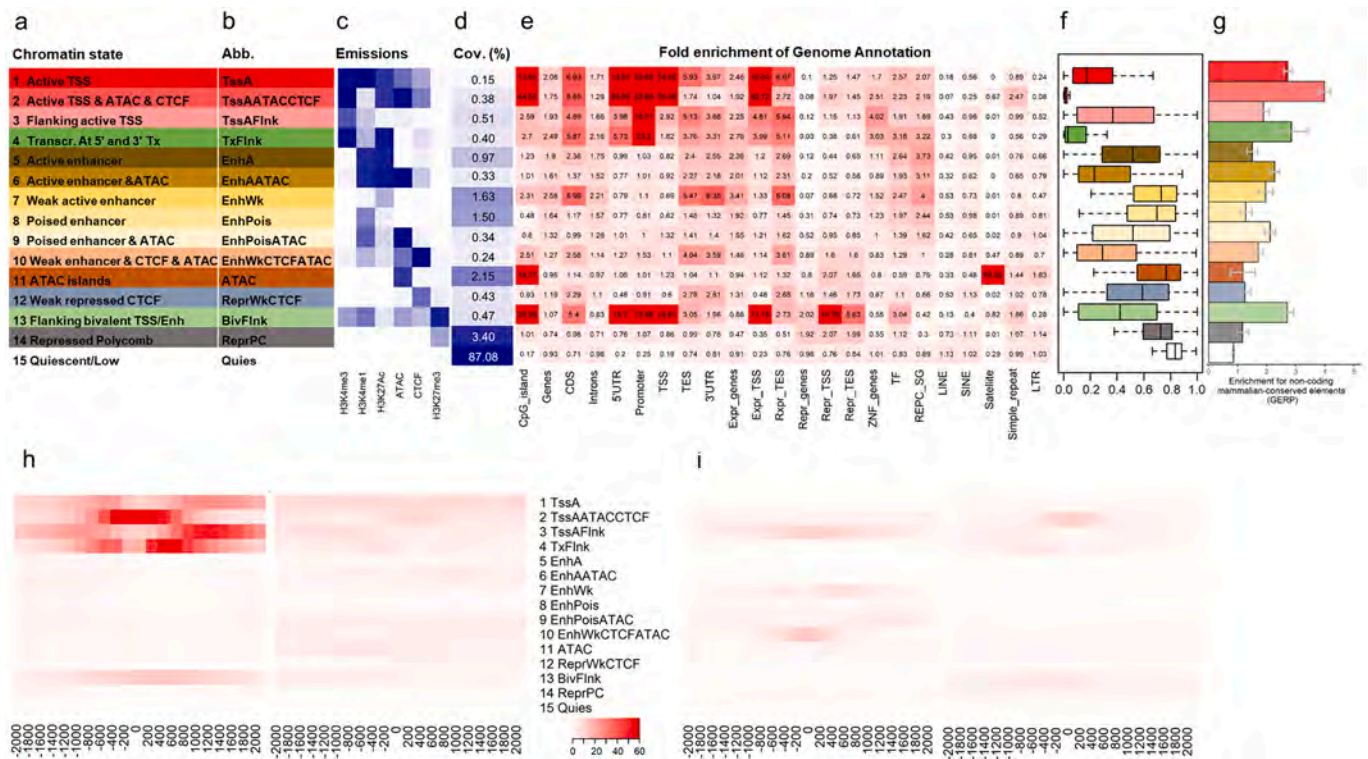


Fig. 2. Definition and characteristics of 15 chromatin states in rumen tissue. (a) (b) Definitions and abbreviations of 15 chromatin states, respectively. (c) Emission probabilities of individual epigenomic marks for each chromatin state. (d) Genomic coverages of chromatin states in BW rumen tissue. (e) Enrichments of chromatin states for diverse genomic annotations, including CpG islands, gene contents (promoters were defined as ± 1 kb around TSS), expressed genes (TPM ≥ 0.1), repressive genes (TPM < 0.1), transcription factors (TF), and common repeats. (f) DNA methylation across 15 chromatin states in rumen tissue. (g) Fold enrichments for noncoding mammalian conserved elements as measured by Genomic Evolutionary Rate Profiling (GERP). (h) (i) Enrichments of chromatin states around ± 2 kb of TSS and TES of expressed genes and repressive genes, respectively.

correlated to gene expression [23]. We further identified that TssA had the utmost enrichment for non-exonic mammalian conserved elements (Fig. 2g). By further assessing gene TSS and transcriptional end site (TES), we perceived that the first three states had high-level enrichment in the vicinity (± 2 kb) of TSS and TES for expressed genes but not for repressed genes (Fig. 2h, i). While TssAFlnk and TxFlnk bordered around TSS of expressed genes, TssA centered at TSS of expressed genes (Fig. 2h, i). The transition parameters (signaling the proximal genomic locations) among chromatin states discovered from ChromHMM indicated that the first three states were more probable to shift among one another than to other states, while TssAFlnk was more likely to shift to the quiescent state than TssA and TxFlnk (Fig. S2).

2.3. Functional characteristics of 15 chromatin states

By examining the 108,274 QTLs for 317 complex traits in cattle QTLdb (release 42, Aug. 27, 2020) [24], we verified that active promoters/transcripts (chromatin states 1–3), followed by BivFlnk, showed the highest enrichment for all these QTLs compared to the other 11 chromatin states (Fig. 3a). As previous studies revealed that the majority of eQTLs were conserved across tissues [25,26], we then overlapped chromatin states with rumen eQTLs identified by our cattle GTEx effort [27] and revealed that active promoters/transcripts (chromatin states 1–3) had the highest enrichment for rumen eQTLs among all 15 chromatin states (Fig. 3b). We also established that active promoters/transcripts had the highest enrichment for rumen sQTLs (Fig. 3c) and selection signatures that were detected in the previous study [28] (Fig. 3d, Fig. S3a), revealing that active promoters and transcripts are more likely correlated with positive selection.

By overlapping chromatin states with eQTLs from diverse tissues, we validated that chromatin states associated with TssA, TssAATACCTCF,

and TssAFlnk (chromatin states 1–3) were highly over-represented for all the tissues (Fig. 3e), suggesting that the commonly expressed genes are generally conserved across tissues and under similar regulatory networks. Alternatively, eQTLs from blood/immune tissues and the salivary gland were associated with BivFlnk. For sQTLs, compared to other tissues, the salivary gland, and skin fibroblast were more enriched within multiple promoter/enhancer regions (Fig. 3e), while Leukocyte was more enriched in BivFlnk and ReprPC regions. We speculate that these may represent some underlining cell types shared between these tissues and rumen, but the conjecture warrants future investigation.

Our large-scale GWAS signal enrichment analysis discovered that active promoters and transcripts (*i.e.*, TssA, TssAATACCTCF, and TxFlnk) were the highest enriched chromatin states within 45 complex traits of economic significance in the US Holstein population (Fig. 3f), in line with the findings in cattle QTLdb (Fig. 3a). Interestingly, enhancer-associated regions, which were likely to be tissue-specific, were particularly enriched for body type traits (particularly for stature) and somatic cell score (SCS, a mastitis incidence indicator), suggesting the potential roles of rumen tissue in growth and innate immune responses (Fig. 3f). The motif enrichment analysis revealed that the tested motifs were significantly (adjusted $P < 0.01$) enriched in EnhWkCTCFATAC, mainly including motif families of zinc finger transcription factors (Table S2). This observation demonstrates that enhancers are a hotspot for transcription regulatory factors and imply that highly expressed genes also involve a complex regulatory mechanism to ensure their proper function. We also found that ReprPC was mainly enriched in the bHLH and Zf families (Fig. 3g).

We downloaded and analyzed additional 293 public rumen RNA-seq data sets to explore relationships between chromatin states and rumen gene expression. After log transforming gene TPM values, we calculated the average (Mean) and median absolute deviation (MAD, which can be

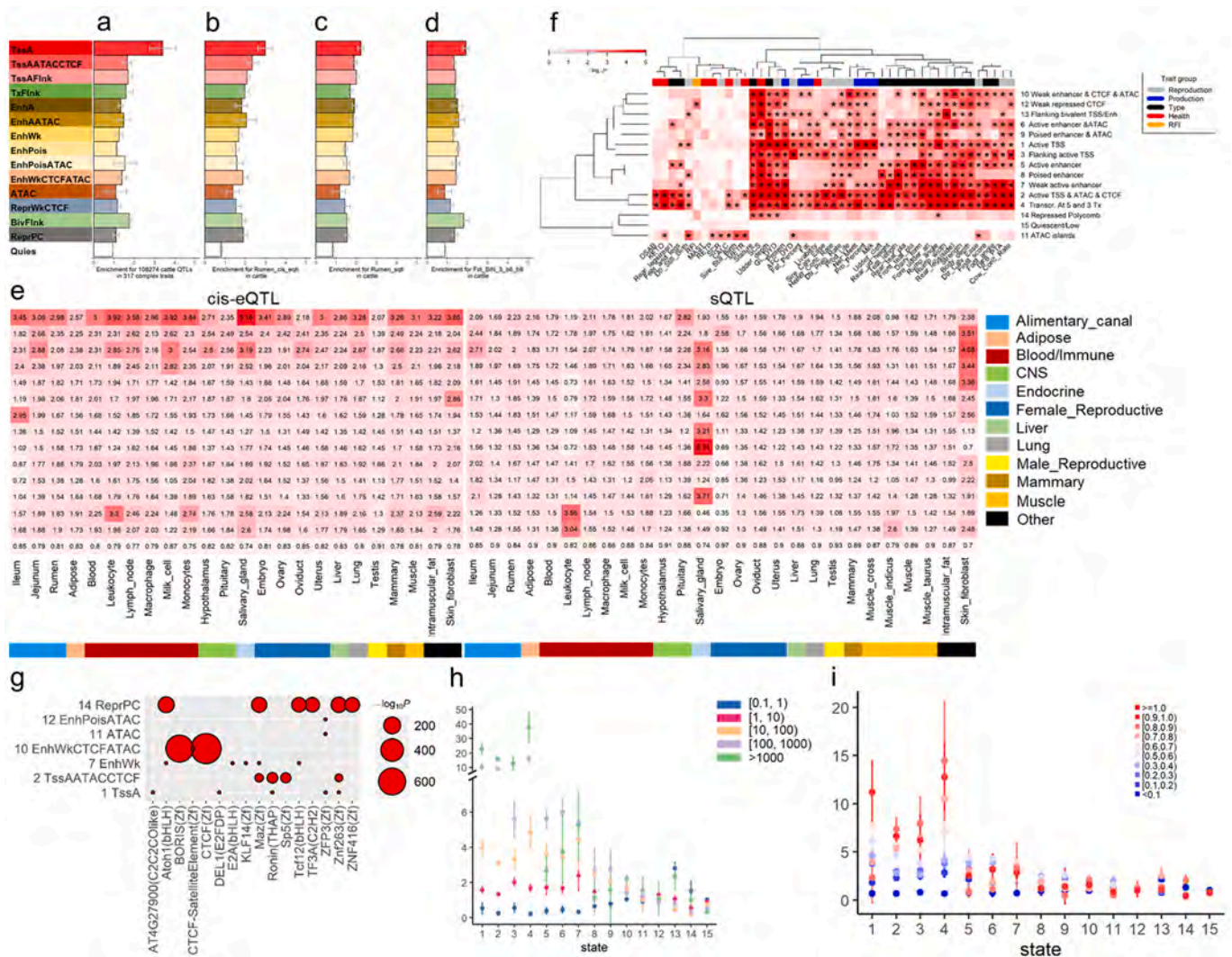


Fig. 3. Functional characteristics of 15 chromatin states. (a) Fold enrichments for 108,274 QTLs (length < 10 kb) of 317 complex traits in cattle QTLdb. (b) Fold enrichments for rumen cis-eQTLs in cattle. (c) Fold enrichments for rumen sQTL in cattle. (d) Fold enrichments for selected regions. (e) Enrichments of chromatin states for cis-eQTL and sQTL of diverse tissues. These diverse tissues can be grouped into 12 broad tissues shown in the different color bars. (f) GWAS signal enrichment of 45 complex traits in the US Holstein population. (g) The top enriched motifs among the seven chromatin states (TssA, TssAATACCTCF, EnhWk, EnhWkCTCFATAC, ATAC, EnhPoisATAC and ReprPC). (h) The average fold enrichment of expressed genes with 5 expression levels ([0.1, 1], [1,10], [10,100], [100,1000], > 1000) in 15 chromatin states. (i) The average fold enrichment of genes with 10 MAD levels in 15 chromatin states.

used to measure each gene's inter-individual variability) values for each gene, as described before [29]. The correlation coefficient between these two variables is 0.24 (Fig. S3b). Our other study [30] showed that inter-individual variable genes were significantly engaged in tissue-relevant functions, while consistent genes were significantly involved in essential biological functions, such as system processes and stimulus detection. We classified genes into five categories according to their average gene expression level and MAD and then made enrichment estimates with 15 chromatin states. We found that with increasing gene expression/variability, the fold enrichment also increased, and the key states showing growing trends were from promoter and enhancer groups (Fig. 3h and i). The low MAD genes and low Mean genes were rarely enriched for any states (Fig. S3c), indicating many promoters' activities for ensuring gene expression and differentiation. We also explored the functions of top MAD and Mean genes and found the most significant terms for both mainly were related to mitochondrial energy metabolisms, reflecting stark changes in metabolism and energy production in rumen epithelium, which must occur when transitioning from glucose use to short chain fatty acid use by epithelium during the weaning process (Fig. S3d, Table S3). Transcriptomic reprogramming is required

to induce developmental changes in ruminal epithelium and functional transitions during weaning [31]. These enrichment patterns and dynamic chromatin state changes *in vivo* during weaning further ratify that promoters and enhancers dictate and safeguard gene expression and differentiation in rumen development.

2.4. Weaning-induced dynamics in chromatin states, gene expression, and DNA methylation

In this assessment, we tried to find the regulatory element changes during the weaning process *in vivo*. After weaning, we observed the greatest changes in chromatin state for the ReprWkCTCF, TssA, and EnhA states, which showed ~1–3% increases in their overall proportion of regions as compared to BW (Fig. 4a). We grouped 15 chromatin states into six broad groups based on the functional regions they affect, including promoter, enhancer, bivalent, heterochromatin, ATAC, and CTCF, and then evaluated their functional impact using the GREAT tools [32]. In promoter regions, we found that BW was enriched for cell-substrate junction assembly, focal adhesion assembly, and mitochondrial outer membrane translocase complex, while AW was for many

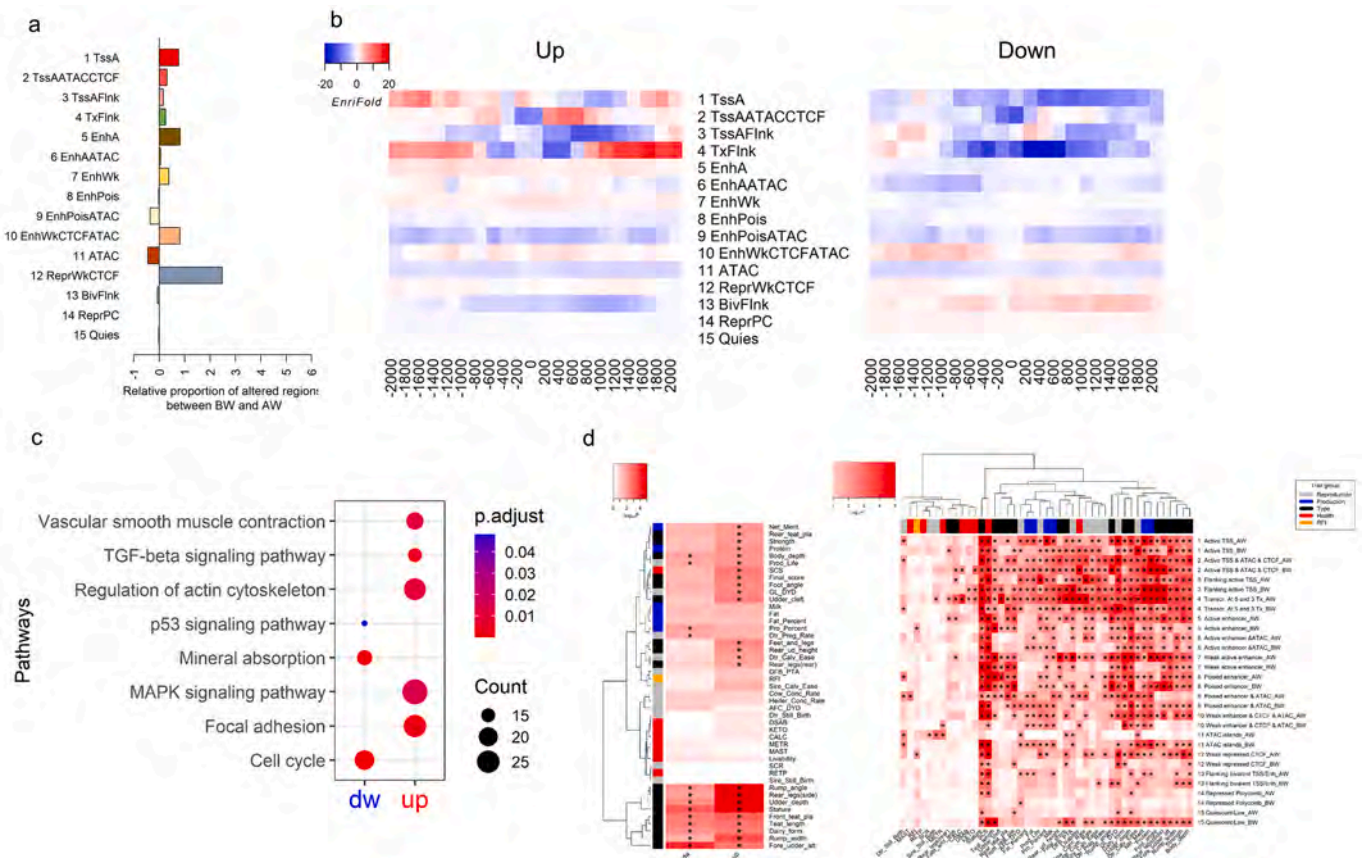


Fig. 4. The dynamics after weaning in chromatin states, gene expression, and their associated traits. (a) The relative proportion of changed regions between BW and AW in rumen tissue. (b) Changes of enrichment folds of upregulated (left) and downregulated (right) DEGs for 15 chromatin states in BW and AW, respectively. (c) Significantly enriched pathways for up and downregulated DEGs. (d) GWAS signal enrichment of DEGs (left) and specific regions (right) for 45 complex traits in the US Holstein population.

development-related processes. In enhancer regions, we found that BW was enriched for the following BP terms (substrate adhesion-dependent cell spreading, regulation of cell migration involved in sprouting angiogenesis, negative regulation of intrinsic apoptotic signaling pathway, cell-substrate junction assembly, and cellular response to vascular endothelial growth factor stimulus), MF terms (collagen binding, extracellular matrix binding, and transforming growth factor-beta binding).

On the other hand, AW was enriched for the BP terms (cell-matrix adhesion, positive regulation of adherens junction organization; CC terms of ruffle membrane, podosome, beta-catenin destruction complex) and MF terms (cadherin binding involved in cell-cell adhesion). In ATAC chromatin open regions, for BW, we found that enriched BP terms like cell-substrate junction assembly, adherens junction assembly, and regulation of membrane depolarization, while for AW, there were many general terms. In CTCF regions, we found BW-enriched terms (cell-substrate junction assembly, epithelial cell-cell adhesion, and basal lamina) and AW-enriched terms (regulation of apoptotic process involved in morphogenesis). In bivalent regions, we found that BW was enriched for pattern specification process, skeletal system morphogenesis, cell fate commitment, dorsal/ventral pattern formation, and extracellular matrix binding, at the same time, AW was for cornification, positive regulation of morphogenesis of epithelium, vascular smooth muscle cell differentiation, morphogenesis of an epithelial sheet, smooth muscle cell differentiation, keratin filament, and delta-catenin binding. In heterochromatin regions, we saw that many general developmental GO terms were enriched (Table S4).

From RNA-seq, we detected 2193 differentially expressed genes (DEGs) between BW and AW group, including 1143 up- and 1050

downregulated DEGs, respectively (Table S5). Remarkably, we observed that TSS of upregulated DEGs (± 2 kb) attained enrichments for TssA and TxFlnk, while eluding enrichment for BivFlnk and ReprPC in AW, demonstrating that a portion of BivFlnk likely transitioned into active promoters and thus led to increased transcriptions of the corresponding genes post-weaning (Fig. 4b). Histone modifications changes (for example, increase in H3K4me3 and decrease in H3K27me3) were associated with these transitions. The TSS of downregulated DEGs reduced TssA, TssAFlnk, and TxFlnk enrichments and promoted BivFlnk and ReprPC enrichments, probably explaining the concomitant reduction in their gene expression (Fig. 4b). These findings reveal the imperative interplay between chromatin state and gene expression in rumen tissue during weaning. Similar results were also observed in our previous REPC *in vitro* experiments [18]. Functional enrichment analysis further demonstrated that upregulated DEGs participated in the MAPK signaling pathway, focal adhesion, regulation of actin cytoskeleton, vascular smooth muscle contraction, and TGF-beta signaling pathway (Fig. 4c, Table S6).

GWAS signal enrichment analysis demonstrated that both down- and upregulated genes were significantly associated with the stature traits in dairy cattle (Fig. 4d). Interestingly, the upregulated genes were also associated with other economic traits in dairy cattle, like milk production and SCS (Fig. 4d). Also, the GWAS analysis based on specific regions from AW- and BW-specific regions revealed that active promoters and transcripts were the top enriched chromatin states across trait groups, followed by enhancers (Fig. 4d). We found that most states were also associated with stature traits and SCS. Generally, the chromatin states of AW were significantly more enriched than BW's corresponding states (Fig. 4d). These results suggested the potential roles of rumen tissue in

growth and innate immune responses.

2.5. Comparison of chromatin state, gene expression changes in the rumen *in vivo* and REPC *in vitro*

2.5.1. RNA-seq DEGs shared or unique in *in vivo* and *in vitro*

To uncover the development processes of *in vivo* rumen and *in vitro* REPC, we first detected DEGs for either of their processes and then compared DEGs between them (Table S7). We obtained four groups of DEGs (Table S8): (1) DEGs were downregulated in both processes (DD); (2) DEGs were downregulated only *in vivo* but upregulated *in vitro* (DU); (3) DEGs were upregulated *in vivo* but downregulated *in vitro* (UD); (4) DEGs were upregulated in both processes (UU). In group DD, it is interesting to note that multiple CC terms were related to adhesion, adherens junction, anchoring junction, cell junction, and organelle. Other enriched terms were related to rRNA, ribosome, unfold protein, RNA processing, and metabolic process. On the other hand, the enriched terms in group UU were related to metabolic processes, oxidoreductase, and hydrolase activity. In group DU, we detected extracellular matrix (GO: 0031012) and growth factor binding (GO: 0019838), which were significantly enriched only *in vivo*. Lastly, in group UD, many processes were involved, including cell division, chromosome segregation, microtubule cytoskeleton, mitochondrial electron transport, carbohydrate derivative metabolic process, inorganic cation transmembrane transport, and phosphorus metabolic process, suggesting there might be certain degrees of subtle differences *in vitro*, as compared to the *in vivo* conditions (Table S9).

2.5.2. Regulatory elements shared or unique in *in vivo* and *in vitro*

When comparing the relative proportion changes of the same chromatin states induced by weaning and butyrate, we found that most chromatin states had the exact changing directions (Fig. 5a), suggesting that weaning and butyrate may impact similar pathways. To uncover the dynamic pattern of *in vivo* rumen and *in vitro* REPC, we used GREAT [32] to compare the relative proportion of changed regions and genome coverage differences *in vivo* and *in vitro* induced by weaning and butyrate, respectively. Besides the previously described BW: AW comparison (Comparison 1), we performed three additional comparisons: (2) CO: BT; (3) CO: BW; and (4) BT: AW.

In Comparison 2, for CO's Enhancer regions, we found epithelial to mesenchymal transition, collagen fibril organization, cell-substrate junction assembly, substrate adhesion-dependent cell spreading,

collagen binding, extracellular matrix binding, and Wnt-activated receptor activity (Table S10a). We also found a cellular response to epidermal growth factor stimulus and an intermediate filament-based process for BT's promoter regions (Table S10b).

In Comparison 3, we took a close look at *in vivo* and *in vitro* samples, *i.e.*, before weaning rumen tissue (BW) and control REPC (CO). We observed primarily similar trends in the relative proportion of 15 chromatin states between BW and CO, with minor differences for TxFlnk and ATAC (~2%). In promoter regions of both BW and CO, we found that multiple terms were enriched like cell cycle (Table S11a, b). For BW in enhancer regions, we found related terms like basal plasma membrane and response to fluid shear stress (Table S11b). While for CO, we found cell-substrate junction assembly and regulation of transforming growth factor β receptor signaling pathway (Table S11a). Additionally, for CO's MF terms, we detected integrin-binding, extracellular matrix binding, laminin-binding, and fibronectin-binding, suggesting that *in vitro* systems and the intact rumen tissue have specific differences (Table S11a). Intriguingly, in ATAC chromatin open regions, for BW but not for CO, we found multiple enrichments of extracellular interaction terms, like focal adhesion assembly, cell-substrate junction assembly, adherens junction assembly, cell junction assembly, and cell-matrix adhesion (Table S11b). Lastly, we did not find clear patterns for CTCF regions, heterochromatin regions, and bivalent regions, which may be related to their broad genome coverage, leading to their associations with many development processes (Table S11a, b).

In Comparison 4, we aimed to identify the differences between the ending patterns between the two processes. In promoter regions, for BT, we found cell-substrate junction assembly, adherens junction assembly, and hemidesmosome (Table S12a); while for AW, we detected cellular response to topologically incorrect protein, intrinsic apoptotic signaling pathway in response to DNA damage, miRNA mediated inhibition of translation, heterochromatin organization, focal adhesion assembly, and histone deacetylase activity (Table S12b). We found related terms like cell-substrate junction assembly and actin filament bundle organization for BT in enhancer regions (Table S12a). While for AW, besides many development terms, the enriched terms also included establishing epithelial cell polarity, regulation of lamellipodium organization, basal plasma membrane, basal part of the cell, and podosome (Table S12b). For AW in open chromatin regions, we found embryonic digestive tract development, morphogenesis of an epithelial bud, lamellipodium organization, and Wnt-activated receptor activity (Table S12b).

The resulting chromatin state maps allow the visualization of

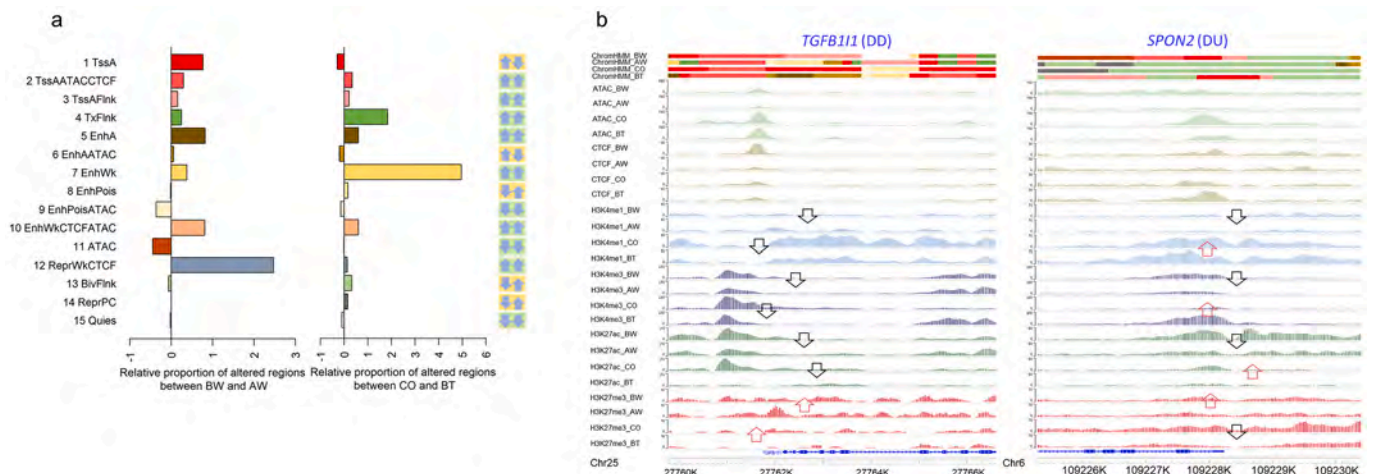


Fig. 5. Comparison of *in vivo* rumen and *in vitro* REPC. (a) Comparison of the relative proportion of changed regions between the weaning effect in rumen tissue and butyrate effect in REPC in 15 chromatin states. Upward arrow, relative proportion increasing; Downward arrow, relative proportion decreasing. Orange rectangle, consistent trends; Green rectangle, inconsistent trends. (b) WashU Epigenome Browser [74] view of *TGFBI1* and *SPON2* showing the 15 chromatin states across the whole genome (top 4 rows) and peaks distribution of ATAC, CTCF, H3K4me1, H3K4me3, H3K27ac, and H3K27me3 (the remaining rows). Chromatin states are colored as in Fig. 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

multiple functional predictions within a genomic interval. For example, we used *TGFB111* (Transforming Growth Factor Beta 1 Induced Transcript 1) and *SPON2* (Spondin 2) to show signals from several chromatin states across CO, BT, BW, and AW (Fig. 5b). *TGFB111* was a DD gene, which had decreased expression in both *in vivo* and *in vitro* processes. Its chromatin state around TSS changed from TssAATACCTCF in BW to EnhWk in AW. At the same time, it changed from TssA in CO to EnhA in BT. (Fig. 5b). The protein encoded by *TGFB111* can function as a molecular adapter coordinating multiple protein-protein interactions at the focal adhesion complex and nucleus [33]. It links various intracellular signaling modules to plasma membrane receptors and regulates the Wnt and TGF β signaling pathways [34]. The TGF β superfamily is critical in wound healing and repair. It must be activated by release from the extracellular matrix, where it is bound by latent TGF β -binding proteins and active proteases, such as the matrix metalloproteinase [35]. TGF β has been shown to inhibit the proliferation of keratinocytes [36,37]. This pathway involves many cellular processes in both the adult and the developing embryo, including cell growth, cell differentiation, apoptosis, cellular homeostasis, and other cellular functions. We previously reported that TGF β 1 is an essential transcriptional regulator of gene expression networks related to specific diets using the same calf rumen epithelium samples during weaning [38]. Our rediscovery of the same TGF β pathways further confirmed that these cytokines and their related proteins are likely involved in regulating the growth and differentiation of the rumen epithelium.

On the contrary, *SPON2* is a DU gene encoding Spondin 2. It was downregulated *in vivo* but upregulated *in vitro*. Its chromatin state around TSS changed from TssA/TssAATACCTCF in BW to BivFlnk in AW. Meanwhile, it changed from BivFlnk in CO to TssA in BT. *SPON2* is a cell adhesion protein that promotes adhesion [39]. It is also crucial to initiate the innate immune response and facilitate a unique pattern recognition in the extracellular matrix for microbial pathogens, such as lipopolysaccharide. Thus, it can bind directly to bacteria and their components and functions as an opsonin for macrophage phagocytosis of bacteria [40].

When checking the underlining histones codes, we observed histone codes changed dynamically in the expected directions (as labeled in red/up and black/down arrows in Fig. 5b), including H3K4me1 (primed enhancers), H3K4me3 (transcriptionally active promoters), H3K27ac (distinguishes active enhancers from poised enhancers), and H3K27me3 (found in facultatively repressed genes). From the chromatin state assignments near these two genes' TSS, we found that the TssAATACCTCF and EnhWk were prevalent at well-characterized regulators of rumen tissue, while another two states, TssA and EnhA, were found mainly in REPC.

Taking together, we found similar regulatory genomic elements in general, but with certain degrees of distinction, involved in weaning and butyrate-induced genomic activities. Also, we found that the most distinct regions within the chromatin states between rumen tissue and REPC were promoters, enhancers, BivFlnk, and open ATAC chromatin regions (Fig. S4) implies their conserved but also versatile functions in weaning and response to butyrate. This similarity and distinction of the genomic regulatory elements between rumen tissues and primary rumen epithelial cells may suggest that intact tissues preserve local *in vivo* cell-to-cell interactions, which is critical for their normal development. In contrast, this microenvironment is absent in primary cells, especially after cell separations, digestions, or multiple rounds of cell division *in vitro*. Those results indicated that animal models in functional genomics studies could confirm and improve the study results on primary cells or cell lines, while using primary cells or cell lines allows the researchers to elude complications in using animal models, such as availability, cost, and ethics.

This study created the first *in vivo* global map of regulatory elements (15 unique chromatin states) in cattle. We defined their coordinated activities through genome-wide profiling for four specific histone modifications, CTCF-binding sites, DNA accessibility, and DNA

methylation. Functional annotations of the genome in the intact tissue describe a significant diversity of genomic functions determined by distinct chromatin states and reveal that most of them are consistent. We identified significant correlations of chromatin states with gene expression and DNA methylation. We demonstrated the importance of comprehensive functional annotation to facilitate the enriched interpretation of the genetic basis underlining complex trait variation, eQTLs, positive selection, and adaptive evolution in cattle. Our data further suggested that most defined chromatin states were generally consistent across tissues and primary cell types. However, there are significant differences between the *in vivo* and *in vitro* processes, suggesting the importance of cell adhesions and communications as well as the impacts of the microenvironment.

Overall, our data indicated that epigenomic landscapes and chromatin states in both rumen tissues and primary rumen epithelial cells could change dynamically induced by butyrate or weaning, resulting in specific gene expression changes and influencing rumen development. We illustrated that the up- and down-regulated genes induced by butyrate treatment and weaning process exhibited a distinctive alteration in chromatin states and altered biological functions. It has been generally recognized that histone modifications play an essential role in controlling gene expression. Butyrate, a native histone deacetylase (HDAC) inhibitor, stimulates histone post-translational modifications and, thus, regulates cell growth, apoptosis, and cell differentiation in many types of cancer [43]. There is an abundance of information on butyrate. As an HDAC inhibitor, butyrate plays the role of aberrant histone acetylation in tumorigenesis and the potential for cancer chemoprevention and therapy [43–46]. There is little information about normal rumen development and butyrate's biological impacts therein. Outlining the extent to which the epigenomic landscape and chromatin states are modified by normal rumen development and butyrate-induced histone post-translational modification is critical to understanding how these processes function at the mechanistic level. By comparing normal and butyrate-induced dynamic variation of chromatin states concomitantly with changes in transcription activities observed in REPC and intact tissue, we established correlations among cell interactions, nutritional elements, histone modifications, chromatin states, and other genomic activities like transcription regulation, DNA methylation, positive selection, and others. Indeed, future studies with additional epigenomic marks and tissues/cell samples are required for a more inclusive functional annotation of the cattle genome and corroboration of the essential steps of rumen development.

3. Methods

3.1. Tissue collection and next-generation sequencing

3.1.1. Rumen epithelial tissue collection

The Beltsville Area Animal Care approved animal care and tissue isolation work (Committee Protocol Number 07–025). Animals and tissue collection were fully described in our previous report [8]. Briefly, two Holstein bull calves were chosen: one calf (before weaning) was fed with milk replacer only (MRO - Cornerstone 22:20, Purina Mills, St. Louis, MO, USA; 22.0% crude protein, 20.0% crude fat, 0.15% crude fiber, 0.75 to 1.25% Ca, 0.70% P, 66,000 IU/kg vitamin A, 11,000 IU/kg vitamin D3, and 220 IU/kg vitamin E) for two weeks; while the other (after weaning) was fed with MRO for six weeks, followed by a combination of milk replacer and grain-based commercial calf starter for four weeks. Calves were euthanized by captive bolt followed by exsanguination at day 14 or day 70 to represent development at two stages of weaning on a grain concentrate diet. The methods for rumen epithelial tissue collection were described previously [11,47,48]. Briefly, rumen epithelial tissue was collected from Holstein bull calves at the slaughter. Rumen epithelial tissue was collected from the anterior portion of the ventral sac of the rumen beneath the reticulum and below the rumen fluid layer at slaughter. The epithelial layer of the rumen tissue was

separated manually from the muscular layer. After rinsed in tap water to remove residual feed particles, samples were further rinsed in ice-cold saline and snap-frozen in liquid nitrogen before moved to -80 °C for future use.

3.1.2. ChIP-seq and ATAC-seq

In the present study, the ATAC-seq and ChIP-seq of H3K27ac (antibody Cat No. 30133, Active Motif, Inc), H3K27m3 (antibody Cat No. 39155, Active Motif, Inc), H3K4m1 (antibody Cat No. 39297, Active Motif, Inc), H3K4m3 (antibody Cat No. 39159, Active Motif, Inc), and CTCF (antibody Cat No. 61311, Active Motif, Inc) in rumen tissues were performed by using HiSeq 2500 (Illumina, Inc. San Diego, CA, USA) at Active Motif, Inc. (Carlsbad, CA, USA). The histone modifications of REPC were reported in our earlier publication [18]. The input ChIP DNA prepared for sequencing libraries were recovered from a conventional ChIP procedure and quantified using the QuantiFluor fluorometer (Promega, Madison, WI, USA). Agilent Bioanalyzer 2100 (Agilent; Palo Alto, CA, USA) was used to verify DNA integrity. Using an Illumina sample prep kit following the manufacturer's instructions (Illumina, Inc., San Diego, CA, USA), the DNA was then processed (end repair, adaptor ligation, and size selection). After final validation, DNA libraries were sequenced at 75-nt per sequence read, using an Illumina HiSeq 2500 platform.

3.1.3. RNA sequencing

RNA extraction was performed, following the procedure reported previously [49]. Total RNA from six rumen samples (in addition to two rumen samples used in ChIP-seq, ATAC-seq, and CTCF seq, four more rumen samples were included for RNA-seq replicates) with three replicates for each condition was extracted using Trizol reagent (Invitrogen, Carlsbad, CA, USA) followed by DNase digestion and Qiagen RNeasy column purification (Qiagen, Valencia, CA, USA). High-quality RNA (RNA integrity number [RIN]: 9.0, quality-controlled (QC) using Agilent's Bioanalyzer 2100) was processed using an Illumina TruSeq RNA sample prep kit, abiding by the manufacturer's instruction (Illumina, Inc.). After QC, individual RNA-seq libraries were pooled with their respective sample-specific 6-bp adaptors and paired-end sequenced at 150 bp/sequence reads (PE150) using an Illumina HiSeq 2500 platform.

3.1.4. Whole-genome bisulfite sequencing (WGBS)

All experiments were carried out following published procedures [50,51]. Briefly, DNA from rumen tissues was extracted using phenol/chloroform. DNA (100 ng) was bisulfite-converted and subjected to library preparation using the Pico Methyl-Seq™ Library Prep Kit (Zymo Research, Irvin, CA, USA) following the instructions of the supplier. High-sensitivity DNA chips were used to assess libraries for quality on the Agilent Bioanalyzer and quantified with a Qubit fluorometer (ThermoFisher Scientific, Waltham, MA, USA). Libraries were sequenced on an Illumina HiSeq2500 (150-bp paired-end sequencing).

3.2. Bioinformatics and data analysis

3.2.1. Chromatin states

After removing the raw reads that failed Illumina's quality control, we generated 318,737,324 and 385,739,742 clean paired-end reads for four ATAC-seq data sets and 10 ChIP-seq data sets, respectively. Simultaneously, we generated a total of 43,160,815 paired-end clean reads as the random background input. We next aligned clean reads to the cattle reference genome (UMD3.1.1 [52]) using the BWA aligner with default settings [53]. We only retained the uniquely mapping reads aligned with less than two mismatches and filtered out multiply-mapping reads. We employed MACS2.1.1 for peak-calling with default parameter settings by surveying for substantial enrichment in the studied samples compared to the input data file (*i.e.*, random background) [54]. The details of the remaining 14 marks were reported in our previous study [18]. To illustrate all the marks' characteristic

enrichment patterns, we plotted metagene profiles with deepTools plotProfile [55]. We calculated peak correlations among all 28 epigenomic samples. Briefly, we computed the correlation of sample A with sample B as the number of peaks in A overlapped with B, divided by the total number of peaks in A, while the correlation of B with A as the number of peaks in B overlapped with A, divided by the total number of peaks in sample B.

We utilized ChromHMM v.1.20 [21], which is based on a multivariate Hidden Markov Model, to capture the significant combinatorial interactions between different chromatin marks in their spatial context (chromatin states). ChromHMM trunks the genome into non-overlapping bins and assigns each bin to one of the 15 chromatin states. We defined the chromatin states, as described in our earlier publication [18]. Briefly, we ran ChromHMM on the 28 epigenomes at the default 200-bp resolution, using the histone ChIP-seq BED files and the relevant control files for each dataset. We eventually defined 15 chromatin states using the processed data described above on the 12 marks. This method could endow an impartial and systematic chromatin state discovery along the whole genome [21,56]. We computed the enrichment fold of each state for each external annotation (*e.g.*, CpG islands) as $(C/A)/(B/D)$, where A stands the number of bases in the chromatin state, B stands the number of bases in the external annotation, C stands the number of bases overlapped between state and the external annotation, and D stands the number of bases in the genome. We calculated the significance of enrichment using the Fisher's-Exact Test.

3.2.2. RNA-seq and WGBS

Before any processing, we did QC and trimming by employing FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and Trim-Galore (version 0.4.1) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) for all six RNA-seq (three biological replicates in each condition) and two WGBS data sets. Generally, we removed adapters and reads with low quality ($Q < 20$) or shorter than 20 bp.

For RNA-seq, to quantify gene expression and conduct differentially expression analysis, we used STAR aligner [57], and Cufflinks software tools [58], and only the uniquely mapped reads were used. The TPM value of each gene was used as the normalized expression level. We defined DEGs as Bonferroni-corrected P -value less than 0.05 and \log_2 (fold-change) greater than 2. In addition, we explored gene expression similarity between tissues and across samples by performing hierarchical clustering using TPM. Distance between samples was estimated using the formula $\text{distance} = 1 - \text{correlation}$, where Spearman's correlation coefficient defines correlation. MDS was performed to represent the distances among samples in a parsimonious way. We used the isoMDS [59] function from R (version 3.6.3), with the distance being defined for the MDS analysis.

For WGBS, all clean data were mapped to the cattle reference genome (UMD 3.1.1) using bowtie2 [60]. We then applied Bismark software [61] with default settings to map clean reads to the reference genome (UMD3.1.1) and extracted methylcytosine information using the *bismark_methylation_extractor* (`--ignore-r2 6`) function after de-duplicating duplicated reads. We used the *symmetric-cpgs* program implemented in MethPipe [62] to merge those symmetric CpG pairs and computed the methylation level by genomic region of interest by *roi-methstat* function.

3.2.3. GWAS signal enrichment analysis

We previously reported details of the single-marker GWAS and fine-mapping analyses for 18 body type, six reproduction, and 12 production traits from 27,214 U.S. Holstein bulls [63], eight health traits from 11,880–24,699 bulls, and for one feed efficiency trait from 3,947 Holstein cows [63–66]. Because these 45 complex traits being studied here are highly polygenic, we applied a sum-based marker-set test, implemented by the R package for Quantitative Genetic and Genomic analyses (QGG package; <http://pscoerensen.github.io/qgg/>), for GWAS signal

enrichment analyses across all 15 chromatin states and DEGs between AW vs. BW. We added 20-kb windows around gene regions to include the potential *cis*-regulatory variants. Previous studies showed that this approach had at least equal power compared to other commonly used GWAS signal enrichment methods in humans, *Drosophila melanogaster*, and livestock, especially for highly polygenic traits [6,67–70]. Briefly, we calculated the summary statistics for each genomic feature (e.g., a chromatin state or a list of DEGs):

$$T_{\text{sum}} = \sum_{i=1}^{m_f} b_i^2,$$

where T_{sum} is the summary statistics for each genomic feature, b is the SNP effect in the single-marker GWAS, and m_f is the number of SNPs overlapping a tested genomic feature. We controlled marker-set sizes and linkage disequilibrium patterns among markers by applying the following genotype cyclical permutation strategy [65,71]. Briefly, we first ordered marker effects (i.e., b^2) using their chromosome positions (i.e., $b_1^2, b_2^2, \dots, b_{m-1}^2, b_m^2$). We then randomly selected one marker (i.e., b_k^2) from this vector as the first place, and shifted the remaining ones to new positions while retaining their original orders (i.e., $b_k^2, b_{k+1}^2, \dots, b_{m-1}^2, b_m^2, b_1^2, \dots, b_{k-1}^2$) to maintain LD patterns among markers. We calculated a new summary statistic for the genomic feature using their original chromosome locations. To obtain an empirical P -value for the genomic feature, we repeated this permutation procedure 10,000 times and employed a one-tailed test of the proportion of random summary statistics greater than that observed.

3.2.4. Tissue enrichment analysis for DEGs and other downstream bioinformatics analyses

For tissue/cell type-specific genes, we chose the top 5% of genes specifically expressed in a tissue/cell type as the corresponding tissue/cell-type-specific genes. We filtered and compared lists of DEGs across two pairwise comparisons, using A-Lister [72]. We then employed the PANTHER Classification System (version 15.0) to perform GO enrichment analysis. We used HOMER (<http://homer.ucsd.edu/homer/motif/>) to conduct the motif enrichment analysis for chromatin states considering the whole genome as background. We adjusted P -values for multiple testing using the FDR method. We used GREAT (v4.0.4) [32] to explore the function of our interesting genomic regions and then exported the significant GO terms ($P < 0.05$) after multiple comparison corrections. We kept its significant results, using the criteria (BinomFdrQ ≤ 0.05 , RegionFoldEnrich ≥ 2 , and HyperFdrQ ≤ 0.05).

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2022.110296>.

Ethics approval and consent to participate

All animal procedures were conducted under the approval of the Beltsville Agricultural Research Center (BARC) Institutional Animal Care Protocol Number 07–025.

Consent for publication

Not applicable.

Availability of data and materials

All RNA sequencing data were submitted to NCBI, SRA database (SUB3040669, BioProject ID: PRJNA658627). All other newly generated sequencing data were submitted to NCBI, SRA database (SUB8420017, BioProject ID: PRJNA672996). The reference genome and gene annotation files (including all the sequence ontology, orthologues genes among mammals, and evolutionarily conserved regions) of UMD3.1.1 were downloaded from Ensembl v94 [73]. The Cattle QTLdb (release 42, Aug. 27, 2020) was obtained from [24]. The selection signatures in

cattle were obtained from [28]. The eQTLs in cattle were obtained from Liu et al., 2020 (submitted). All scripts and source codes can be found in <https://github.com/YahGao/15-Chromatin-States>.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by AFRI grant numbers 2013–67015–20951, 2016–67015–24886, 2019–67015–29321, and 2020–67015–02848 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome and Reproduction Programs and BARD grant number US-4997-17 from the US-Israel Binational Agricultural Research and Development (BARD) Fund. G.E. Liu was supported by appropriated project 8042–31000–001-00-D, “Enhancing Genetic Merit of Ruminants Through Improved Genome Assembly, Annotation, and Selection” of the Agricultural Research Service of the United States Department of Agriculture. E.E. Connor, R.L. Baldwin, and C-J Li were supported by appropriated project 8042–31310–078-00-D, “Improving Feed Efficiency and Environmental Sustainability of Dairy Cattle through Genomics and Novel Technologies.” J.B. Cole was supported by appropriated project 8042–31000–002-00-D, “Improving Dairy Animals by Increasing Accuracy of Genomic Prediction, Evaluating New Traits, and Redefining Selection Goals.”

Authors' contributions

CJL, GEL and LF conceived and designed the experiments. EEC, RLB, GEL, and CJL collected samples and/or generated data. YG, LF, SL, and LM performed computational and statistical analyses. YG, LF, EEC, RLB, LM, GEL, and CJL wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

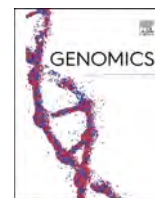
We thank Reuben Anderson, Mary Bowman, Donald Carbaugh, Christina Clover, Cecelia Niland, and Sara McQueeney for technical assistance and sample collection. We thank the Council on Dairy Cattle Breeding for genotype, phenotype, and pedigree data, Interbull for global trait evaluations, and the anonymous reviewers for many helpful comments. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture (USDA). The USDA is an equal opportunity provider and employer.

References

- [1] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shores, L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, et al., Mapping and analysis of chromatin state dynamics in nine human cell types, *Nature* 473 (2011) 43–49.
- [2] P.V. Kharchenko, A.A. Alekseyenko, Y.B. Schwartz, A. Minoda, N.C. Riddle, J. Ernst, P.J. Sabo, E. Larschan, A.A. Gorchakov, T. Gu, et al., Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*, *Nature* 471 (2011) 480–485.
- [3] O. Ram, A. Goren, I. Amit, N. Shores, N. Yosef, J. Ernst, M. Kellis, M. Gymrek, R. Issner, M. Coyne, et al., Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells, *Cell* 147 (2011) 1628–1639.
- [4] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M.J. Ziller, et al., Integrative analysis of 111 reference human epigenomes, *Nature* 518 (2015) 317–330.
- [5] I.M. MacLeod, P.J. Bowman, C.J. Vander Jagt, M. Haile-Mariam, K.E. Kemper, A. J. Chamberlain, C. Schrooten, B.J. Hayes, M.E. Goddard, Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits, *BMC Genomics* 17 (2016) 144.
- [6] L. Fang, G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M.S. Lund, P. Sorensen, Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic

- transcriptomic regions responsive to intra-mammary infection, *Genet. Sel. Evol.* 49 (2017) 44.
- [7] L. Fang, G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M.S. Lund, P. Sorensen, Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds, *BMC Genomics* 18 (2017) 604.
 - [8] L. Andersson, A.L. Archibald, C.D. Bottema, R. Brauning, S.C. Burgess, D.W. Burt, E. Casas, H.H. Cheng, L. Clarke, C. Couldrey, et al., Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project, *Genome Biol.* 16 (2015) 57.
 - [9] N. Malmuthuge, G. Liang, L.L. Guan, Regulation of rumen development in neonatal ruminants through microbial metagenomes and host transcriptomes, *Genome Biol.* 20 (2019) 172.
 - [10] M.A. Khan, A. Bach, D.M. Weary, M.A.G. von Keyserlingk, Invited review: transitioning from milk to solid feed in dairy heifers, *J. Dairy Sci.* 99 (2016) 885–902.
 - [11] E.E. Connor, Baldwin RLT, C.J. Li, R.W. Li, H. Chung, Gene expression in bovine rumen epithelium during weaning identifies molecular regulators of rumen development and growth, *Funct. Integr. Genom.* 13 (2013) 133–142.
 - [12] B.M. Gumbiner, Cell adhesion: the molecular basis of tissue architecture and morphogenesis, *Cell* 84 (1996) 345–357.
 - [13] D.A. Goodenough, D.L. Paul, Gap junctions, *Cold Spring Harb. Perspect. Biol.* 1 (2009), a002576.
 - [14] T.J. Harris, U. Tepass, Adherens junctions: from molecules to morphogenesis, *Nat. Rev. Mol. Cell Biol.* 11 (2010) 502–514.
 - [15] J.L. Johnson, N.A. Najor, K.J. Green, Desmosomes: regulators of cellular signaling and adhesion in epidermal health and disease, *Cold Spring Harb. Perspect. Med.* 4 (2014), a015297.
 - [16] E. Steed, M.S. Balda, K. Matter, Dynamics and functions of tight junctions, *Trends Cell Biol.* 20 (2010) 142–149.
 - [17] M.T. Santini, G. Rainaldi, P.L. Indovina, Apoptosis, cell adhesion and the extracellular matrix in the three-dimensional growth of multicellular tumor spheroids, *Crit. Rev. Oncol. Hematol.* 36 (2000) 75–87.
 - [18] L. Fang, S. Liu, X. Kang, S. Lin, B. Li, E.E. Connor, Baldwin RLT, A. Tenesa, L. Ma, et al., Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations, *BMC Biol.* 17 (68) (2019).
 - [19] C.J. Li, R.W. Li, Y.H. Wang, T.H. Elsasser, Pathway analysis identifies perturbation of genetic networks induced by butyrate in a bovine kidney epithelial cell line, *Funct. Integr. Genom.* 7 (2007) 193–205.
 - [20] L. Liu, D. Sun, S. Mao, W. Zhu, J. Liu, Infusion of sodium butyrate promotes rumen papillae growth and enhances expression of genes related to rumen epithelial VFA uptake and metabolism in neonatal twin lambs, *J. Anim. Sci.* 97 (2019) 909–921.
 - [21] J. Ernst, M. Kellis, ChromHMM: automating chromatin-state discovery and characterization, *Nat. Methods* 9 (2012) 215–216.
 - [22] M.M. de Souza, A. Zerlotini, L. Geistlinger, P.C. Tizioto, J.F. Taylor, M.I.P. Rocha, W.J.S. Diniz, L.L. Coutinho, L.C.A. Regitano, A comprehensive manually-curated compendium of bovine transcription factors, *Sci. Rep.* 8 (2018) 13747.
 - [23] James R. Wagner, Stephan Busche, Bing Ge, Tony Kwan, Tomi Pastinen, M. Blanchette, The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts, *Genome Biol.* 15 (2014) R37.
 - [24] Z.L. Hu, C.A. Park, X.L. Wu, J.M. Reecy, Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era, *Nucleic Acids Res.* 41 (2013) D871–D879.
 - [25] GTEx C, The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans, *Science* 348 (2015) 648–660.
 - [26] A.C. Bouwman, H.D. Daetwyler, A.J. Chamberlain, C.H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, et al., Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals, *Nat. Genet.* 50 (2018) 362–367.
 - [27] S. Liu, Y. Gao, O. Canela-Xandri, S. Wang, Y. Yu, W. Cai, B. Li, E. Pairro-Castineira, K. D'Mellow, K. Rawlik, et al., A Comprehensive Catalogue of Regulatory Variants in the Cattle Transcriptome, 2020 submitted.
 - [28] N. Chen, W. Fu, J. Zhao, J. Shen, Q. Chen, Z. Zheng, H. Chen, T.S. Sonstegard, C. Lei, Y. Jiang, BGVD: an integrated database for bovine sequencing variations and selective signatures, *Genomics Proteomics Bioinform.* 18 (2) (2020) 186–193, <https://doi.org/10.1016/j.gpb.2019.03.007>.
 - [29] M. Teng, M.I. Love, C.A. Davis, S. Djebali, A. Dobin, B.R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, et al., A benchmark for RNA-seq quantification pipelines, *Genome Biol.* 17 (2016) 74.
 - [30] Y. Yao, S. Liu, C. Xia, Y. Gao, Z. Pan, O. Canela-Xandri, S. Wang, B. Li, J. Li, G. Cai, et al., Comparative Transcriptome in Large-Scale Human and Cattle Populations, 2020 submitted.
 - [31] R.L.V. Baldwin, M. Liu, E.E. Connor, T.G. Ramsay, G.E. Liu, C.J. Li, Transcriptional Reprogramming in Rumen Epithelium during the Transition of Pre-Ruminant to Ruminant in Cattle, 2020 submitted.
 - [32] C.Y. McLean, D. Bristor, M. Hiller, S.L. Clarke, B.T. Schaar, C.B. Lowe, A. M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions, *Nat. Biotechnol.* 28 (2010) 495–501.
 - [33] N. Fujimoto, S. Yeh, H.Y. Kang, S. Inui, H.C. Chang, A. Mizokami, C. Chang, Cloning and characterization of androgen receptor coactivator, ARA55, in human prostate, *J. Biol. Chem.* 274 (1999) 8316–8321.
 - [34] M. Matsuya, H. Sasaki, H. Aoto, T. Mitaka, K. Nagura, T. Ohba, M. Ishino, S. Takahashi, R. Suzuki, T. Sasaki, Cell adhesion kinase beta forms a complex with a new member, Hic-5, of proteins localized at focal adhesions, *J. Biol. Chem.* 273 (1998) 1003–1014.
 - [35] O. Tatti, P. Vehviläinen, K. Lehti, J. Keski-Oja, MT1-MMP releases latent TGF-beta1 from endothelial cell extracellular matrix via proteolytic processing of LTBP-1, *Exp. Cell Res.* 314 (2008) 2501–2514.
 - [36] M. Smidt, I. Kirsch, L. Ratner, Deletion of Alu sequences in the fifth c-sis intron in individuals with meningiomas, *J. Clin. Invest.* 86 (1990) 1151–1157.
 - [37] J. Kalucka, A. Ettinger, K. Franke, S. Mamlouk, R.P. Singh, K. Farhat, A. Muschter, S. Olbrich, G. Breier, D.M. Katschinski, et al., Loss of epithelial hypoxia-inducible factor prolyl hydroxylase 2 accelerates skin wound healing in mice, *Mol. Cell. Biol.* 33 (2013) 3426–3438.
 - [38] E.E. Connor, Baldwin RLT, M.P. Walker, S.E. Ellis, C. Li, S. Kahl, H. Chung, R.W. Li, Transcriptional regulators transforming growth factor-beta1 and estrogen-related receptor-alpha identified as putative mediators of calf rumen epithelial tissue development and function during weaning, *J. Dairy Sci.* 97 (2014) 4193–4207.
 - [39] R. Manda, T. Kohno, Y. Matsuno, S. Takenoshita, H. Kuwano, J. Yokota, Identification of genes (SPON2 and C20orf2) differentially expressed between cancerous and noncancerous lung cells by mRNA differential display, *Genomics* 61 (1999) 5–14.
 - [40] Y. Li, C. Cao, W. Jia, L. Yu, M. Mo, Q. Wang, Y. Huang, J.M. Lim, M. Ishihara, L. Wells, et al., Structure of the F-spondin domain of mindin, an integrin ligand and pattern recognition molecule, *EMBO J.* 28 (2009) 286–297.
 - [43] A. Ahmad, A. Mrkvicova, M. Chmelarova, E. Peterova, R. Havelek, I. Baranova, P. Kazimirova, E. Rudolf, M. Rezacova, The effect of sodium butyrate and cisplatin on expression of EMT markers, *PLoS One* 14 (2019).
 - [44] R.H. Dashwood, M.C. Myzak, E. Ho, Dietary HDAC inhibitors: time to rethink weak ligands in cancer chemoprevention? *Carcinogenesis* 27 (2006) 344–349.
 - [45] M.C. Melinda, H.R. Roderick, Histone deacetylases as targets for dietary cancer preventive agents: lessons learned with butyrate, diallyl disulfide, and sulforaphane, *Curr. Drug Targets* 7 (2006) 443–452.
 - [46] M.C. Myzak, E. Ho, R.H. Dashwood, Dietary agents as histone deacetylase inhibitors, *Mol. Carcinog.* 45 (2006) 443–446.
 - [47] R.L. Baldwin, The proliferative actions of insulin, insulin-like growth factor-I, epithelial growth factor, butyrate and propionate on ruminal epithelial cells in vitro, *Small Ruminant Res.* 32 (1999) 261–268.
 - [48] S. Lin, L. Fang, X. Kang, S. Liu, M. Liu, E.E. Connor, R.L. Baldwin, G. Liu, C.J. Li, Establishment and transcriptomic analyses of a cattle rumen epithelial primary cells (REPC) culture by bulk and single-cell RNA sequencing to elucidate interactions of butyrate and rumen development, *Heliyon* 6 (2020), e04112.
 - [49] L. Fang, J. Jiang, B. Li, Y. Zhou, E. Freebern, P.M. Vanraden, J.B. Cole, G.E. Liu, L. Ma, Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle, *Commun. Biol.* 2 (2019) 100.
 - [50] S. Gravina, X. Dong, B. Yu, J. Vijg, Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome, *Genome Biol.* 17 (2016) 150.
 - [51] B. Yu, X. Dong, S. Gravina, O. Kartal, T. Schimmel, J. Cohen, D. Tortoriello, R. Zody, R.D. Hawkins, J. Vijg, Genome-wide, single-cell DNA methylomics reveals increased non-CpG methylation during human oocyte maturation, *Stem Cell Rep.* 9 (2017) 397–407.
 - [52] A.V. Zimin, A.L. Delcher, L. Florea, D.R. Kelley, M.C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C.P. Van Tassel, T.S. Sonstegard, et al., A whole-genome assembly of the domestic cow, *Bos taurus*, *Genome Biol.* 10 (2009) R42.
 - [53] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, *Bioinformatics* 25 (2009) 1754–1760.
 - [54] Y. Zhang, T. Liu, C.A. Meyer, J. Eeckhoutte, D.S. Johnson, B.E. Bernstein, C. Nusbaum, R.M. Myers, M. Brown, W. Li, X.S. Liu, Model-based analysis of ChIP-Seq (MACS), *Genome Biol.* 9 (2008) R137.
 - [55] F. Ramírez, D.P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A.S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis, *Nucleic Acids Res.* 44 (2016) W160–W165.
 - [56] J. Ernst, M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome, *Nat. Biotechnol.* 28 (2010) 817–825.
 - [57] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics* 29 (2013) 15–21.
 - [58] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S. L. Salzberg, J.L. Rinn, L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks, *Nat. Protoc.* 7 (2012) 562–578.
 - [59] W.N. Venables, B.D. Ripley, *Modern Applied Statistics with S*, Springer, Fourth edition, edn, 2002.
 - [60] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol.* 10 (2009) R25.
 - [61] F. Krueger, S.R. Andrews, Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications, *Bioinformatics* 27 (2011) 1571–1572.
 - [62] Q. Song, B. Decato, E.E. Hong, M. Zhou, F. Fang, J. Qu, T. Garvin, M. Kessler, J. Zhou, A.D. Smith, A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics, *PLoS One* 8 (2013), e81148.
 - [63] J. Jiang, J.B. Cole, E. Freebern, Y. Da, P.M. Vanraden, L. Ma, Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls, *Commun. Biol.* 2 (2019) 212.
 - [64] B. Li, L. Fang, D.J. Null, J.L. Hutchison, E.E. Connor, P.M. Vanraden, M. J. VandeHaar, R.J. Tempelman, K.A. Weigel, J.B. Cole, High-density genome-wide association study for residual feed intake in Holstein dairy cattle, *J. Dairy Sci.* 102 (2019) 11067–11080.

- [65] L. Fang, W. Cai, S. Liu, O. Canela-Xandri, Y. Gao, J. Jiang, K. Rawlik, B. Li, S. G. Schroeder, B.D. Rosen, et al., Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle, *Genome Res.* 30 (2020) 790–801.
- [66] E. Freebern, D.J.A. Santos, L. Fang, J. Jiang, K.L. Parker Gaddis, G.E. Liu, P. M. VanRaden, C. Maltecca, J.B. Cole, L. Ma, GWAS and fine-mapping of livability and six disease traits in Holstein cattle, *BMC Genomics* 21 (2020) 41.
- [67] L. Fang, P. Sorensen, G. Sahana, F. Panitz, G. Su, S. Zhang, Y. Yu, B. Li, L. Ma, G. Liu, et al., MicroRNA-guided prioritization of genome-wide association signals reveals the importance of microRNA-target gene networks for complex traits in cattle, *Sci. Rep.* 8 (2018) 9345.
- [68] P.D. Rohde, D. Demontis, B.C. Cuyabano, Genomic Medicine for Schizophrenia G, A.D. Borglum, P. Sorensen, Covariance association test (CVAT) identifies genetic markers associated with schizophrenia in functionally associated biological processes, *Genetics* 203 (2016) 1901–1913.
- [69] P. Sarup, J. Jensen, T. Ostensen, M. Henryon, P. Sorensen, Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs, *BMC Genet.* 17 (2016) 11.
- [70] I.F. Sorensen, S.M. Edwards, P.D. Rohde, P. Sorensen, Multiple trait covariance association test identifies gene ontology categories associated with chill coma recovery time in *Drosophila melanogaster*, *Sci. Rep.* 7 (2017) 2413.
- [71] P.D. Rohde, I. Fourie Sorensen, P. Sorensen, qgg: an R package for large-scale quantitative genetic analyses, *Bioinformatics* 36 (2019) 2614–2615.
- [72] S.A. Listopad, T.M. Norden-Krichmar, A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons, *BMC Bioinform.* 20 (2019) 595.
- [73] S.E. Hunt, W. McLaren, L. Gil, A. Thormann, H. Schuilenburg, D. Sheppard, A. Parton, I.M. Armean, S.J. Trevanion, P. Flicek, F. Cunningham, Ensembl variation resources, Database (Oxford) bay119 (2018), <https://doi.org/10.1093/database/bay119>.
- [74] D. Li, S. Hsu, D. Purushotham, R.L. Sears, T. Wang, WashU epigenome browser update 2019, *Nucleic Acids Res.* 47 (2019) W158–w165.



Original Article

Single-cell transcriptomic analyses of dairy cattle ruminal epithelial cells during weaning

Yahui Gao^{a,b}, Lingzhao Fang^c, Ransom L. Baldwin VI^a, Erin E. Connor^d, John B. Cole^a, Curtis P. Van Tassell^a, Li Ma^b, Cong-jun Li^{a,*}, George E. Liu^{a,*}

^a Animal Genomics and Improvement Laboratory, BARC, USDA-ARS, Beltsville, MD 20705, USA

^b Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA

^c MRC Human Genetics Unit at the Medical Research Council Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, United Kingdom

^d Department of Animal and Food Sciences, University of Delaware, Newark, DE 19716, USA

ARTICLE INFO

Keywords:

Cattle
Ruminal epithelial cell
Single-cell RNA-seq

ABSTRACT

Using the 10× Genomics Chromium Controller, we obtained scRNA-seq data of 5064 and 1372 individual cells from two Holstein calf ruminal epithelial tissues before and after weaning, respectively. We detected six distinct cell clusters, designated their cell types, and reported their marker genes. We then examined these clusters' underlining cell types and relationships by performing cell cycle, pseudotime trajectory, regulatory network, weighted gene co-expression network and gene ontology analyses. By integrating these cell marker genes with Holstein GWAS signals, we found they were enriched for animal production and body conformation traits. Finally, we confirmed their cell identities by comparing them with human and mouse stomach epithelial cells. This study presents an initial effort to implement single-cell transcriptomic analysis in cattle, and demonstrates ruminal tissue epithelial cell types and their developments during weaning, opening the door for new discoveries about tissue/cell type roles in complex traits at single-cell resolution.

1. Background

Rumen development is a critical process necessary for the digestion of solid feed (concentrates and roughage) and optimal growth performance in weaned cattle. Moreover, calf health during the weaning phase, specifically digestive health, has a central role in lifelong feed efficiency and methane emissions [1,2]. The neonatal rumen is undeveloped at birth, exhibiting rudimentary papillae without the high degree of keratinization, which is characteristic of the mature organ. The rumen's physical and metabolic development is incomplete at birth and largely remains so until further development is triggered by short-chain fatty acids (SCFA) resulting from bacterial fermentation of solid feed-stuffs. After establishing a viable ruminal fermentation, the maturation process proceeds [3] resulting in the epithelial layer increasing in

surface area to support absorption and metabolic differentiation to use SCFA as the primary energy substrate. Rumen epithelium serves as both a protective barrier from the digestive luminal environment and a metabolically important tissue for whole-animal energy metabolism [4–6]. A clear understanding of regulatory control of epithelial cell proliferation and differentiation and nutrient-gene interactions is crucial for the optimization of management strategies to support healthy ruminal development. The stratified squamous epithelium absorbs SCFA, which provides up to 70% of the energetic needs of mature animals, and serves as the primary producer of ketones in fed animals [7]. The ruminal epithelium consists of four strata: stratum basale, stratum spinosum, stratum granulosum, and stratum corneum [8]. The stratum basale is the layer of cells immediately adjacent to the basal lamina. These cells contain fully functional mitochondria and other organelles.

Abbreviations: AW, after weaning; BW, before weaning; FAANG, Functional Annotation of Animal Genome project; FDR, False Discovery Rate; GO, Gene Ontology; SCFA, short-chain fatty acids; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site; UMI, unique molecular identifier.

* Corresponding authors at: Animal Genomics and Improvement Laboratory, USDA-ARS, Building 306, Room 111, BARC-East, Beltsville, MD 20705, USA.

E-mail addresses: gyhalvin@gmail.com (Y. Gao), Lingzhao.fang@igmm.ed.ac.uk (L. Fang), Ransom.Baldwin@usda.gov (R.L. Baldwin), eeconnor@udel.edu (E.E. Connor), John.B.Cole@gmail.com (J.B. Cole), curt.vantassell@usda.gov (C.P. Van Tassell), lima@umd.edu (L. Ma), Congjun.Li@usda.gov (C.-j. Li), George.Liu@usda.gov (G.E. Liu).

<https://doi.org/10.1016/j.ygeno.2021.04.039>

Received 9 October 2020; Received in revised form 20 March 2021; Accepted 27 April 2021

Available online 29 April 2021

0888-7543/Published by Elsevier Inc.

The intermediate cell layers are the stratum spinosum and stratum granulosum, which are not distinctively separated [5]. RNA-sequencing has been used to identify the molecular mechanisms involved in rumen development, and several functional studies have been reported in the last decade [9–13]. However, those studies were performed using RNA isolated from whole tissues that include a composite of differentiated cell types. Therefore, they were limited to only measuring whole tissues and providing an average expression profile for all constituent cells [14].

Because of the unique and variable physical composition of the ruminal epithelium, it has been difficult to investigate the effects of differing ruminal environments on all aspects of rumen epithelial functions. The use of isolated ruminal epithelial cells provides several advantages in the study of ruminal metabolism and development [15]. As a crucial and high-value tool to study the development of rumen, we established a stable rumen epithelial primary cell (REPC) culture, explored its transcriptomic profile, and identified the direct effects of butyrate on gene expression in these cells. Correlated gene networks elucidated the putative roles and mechanisms of butyrate action in rumen epithelial development [2,16]. However, the transcriptome profiles are unique to individual cell type, developmental stage, health status, and biological function [17]. In our previous study, transcriptomic profiling of a total of 18 single cells and the clustering of the differentially expressed transcripts showed high divergence and variation in gene expression among the REPC [2]. Single-cell transcriptome complexity and single-cell transcriptome variation have also been reported in different cell types, such as gonadal and stem cell populations [18–20]. It is expected that individual cell phenotypes such as cell type, size, ultrastructure, and stage of the cell cycle could directly control cell-to-cell transcriptome variability. Therefore, large-scale sampling and single-cell transcriptome sequencing are necessary to identify cell type from tissues, evaluate dynamic cellular transitions with complex cell compositions, and illustrate impacts of cell-to-cell interactions among hundreds- to tens-of-thousands of cells.

Breakthroughs in the development of single-cell RNA-seq (scRNA-seq) technologies provide an avenue for dissecting tissue heterogeneity and understanding cell identity, fate, and function. High-throughput single-cell transcriptomes offer an unbiased approach to understanding gene expression variations between seemingly identical cells [21]. Han et al. used scRNA-seq to determine the cell-type composition of all major human organs and constructed a schematic representation of the human cell landscape (HCL) [22]. Their ‘single-cell HCL analysis’ pipeline helped to define cell identity. It was used to perform a single-cell comparative analysis of landscapes from humans and mice to identify conserved genetic networks. Several studies reported scRNA-seq analyses in the human small intestinal epithelium [23], in the human esophagus, stomach, and small and large intestines [24], as well as in the murine gastric organoid [25]. Despite all those developments, cell type profiles of cattle rumen epithelium at a single-cell resolution are lacking.

Many state-of-art analysis tools are available to process scRNA-seq data. For example, Seurat 3.0 [26] is an R package designed for quality control, integration, and scRNA-seq data analysis. Based on highly variable genes (HVG), Seurat performs clustering on cells using the Louvain algorithm [27–29]. Another analytical challenge is the interpretation of clusters and the assignment of cell types. SingleR 1.2.4 is an automatic annotation method that labels new cells from a test dataset based on similarity to the reference cell types [30]. Within SingleR, seven reference atlases are available, including the Human Primary Cell Atlas dataset [31] and the Blueprint and Encode dataset [32,33]. Gene regulatory network (GRN) inference can also reveal regulatory interactions and help identify the role of single cells. SCENIC (Single-Cell rEgulatory Network Inference and Clustering) can construct GRN from scRNA-seq data and infer transcription factor (TF) activity using a statistical method (AUCell 1.8.0) [34]. Trajectory inference methods interpret single-cell data as a snapshot of a continuous process. Monocle 2 [35] can order single cells in pseudotime to represent a biological process such as cell differentiation, according to an individual cell’s

asynchronous progression, using advanced machine learning techniques (such as Reversed Graph Embedding).

In this report, using the 10× Genomics Chromium Controller, we obtained transcriptomic profiling of 5064 and 1372 cells from ruminal epithelial cells of Holstein calves during weaning. We detected thousands of candidate marker genes among different cell clusters. We then examined these clusters’ underlying cell types and relationships by performing cell cycle, pseudotime trajectory, regulatory network, weighted gene co-expression network, and gene ontology (GO) analyses. This study provides an initial example for bovine single-cell analysis and opens the door for discoveries about tissue/cell type roles in complex traits at single-cell resolution.

2. Results

2.1. Data generation and quality assessment

We used the 10× Genomics Chromium platform [36] to generate single-cell transcriptomes for two rumen tissues during weaning, one before weaning (BW) and another after weaning (AW) from two animals. In total, we sequenced 7479 single cells, with approximately 180,000 reads per cell (Table S1). After quality filtering and integration, we obtained 6436 single cells, which corresponded to a median of 37,000 unique molecular identifiers per cell, and more than 15,000 total genes detected in the whole population. Overall, 79% of all single cells (5064) belonged to BW, while the remaining 21% (1372) to AW.

2.2. Seurat cell cluster analyses

Using the Seurat v3.0 R package [26], we performed a community detection-based clustering to groups of cells according to their gene expression profiles. After visualizing the Uniform Manifold Approximation and Projection (UMAP) plots, we found the single-cell transcriptomes of two studied samples were largely similar (Fig. 1A), indicating a high degree of reproducibility. In total, we obtained 6 distinct clusters and named them Clusters C0, C1, C2, C3, C4, and C5 (Fig. 1B & C).

Additionally, we attempted to assign the cell types to the 6 Seurat clusters with SingleR [30], using the human cell reference datasets - Blueprint [37] and Encode [33]. In total, we obtained 12 cell types of the 6436 individual cell transcriptomes from the two rumen tissues (Fig. S1A). The cell count of different cell types ranged from 1 to 5634 (Table S2). After removing cell types with fewer than 12 cells, the three main cell types remaining were epithelial cells, keratinocytes, and mesangial cells (Fig. S1B). The majority of cells in C0, C1, C2, and C3 (Table 1) as epithelial cells. Compared to other clusters, the percentages of keratinocytes and mesangial cells were the highest in C4 and C5, respectively. When comparing the BW versus AW samples, based on the relative cell proportions, we noticed that C1 decreased 9.06%, C0 increased 5.15%, and C3 decreased 5.57%, while C4 increased 8.50% post-weaning. On the other hand, C2 decreased barely at 0.16% and C5 increased slightly at 1.14%, respectively (Fig. 1D).

To compare the performances of bulk and scRNA-seq, we calculated the gene expression correlations between bulk and scRNA-seq. We retrieved RNA-seq data of rumen bulk samples both before and after weaning, as we reported before [2]. As shown in Fig. S1C, the overall correlations were more than 0.65, no matter using all cells’ gene expression (Left panel) or the clustered cells’ gene expression within each cluster (Right panel), which suggested that scRNA-seq and bulk RNA-seq results were generally consistent. However, for a heterogeneous tissue, conventional bulk RNA-seq approaches have difficulties in accurately revealing cell-type-specific changes in gene expression, particularly for rare cell types. On the other hand, since scRNA-seq can capture the gene expression at both cell and bulk tissue levels, we used it to study individual cell types by analyzing alterations in gene transcription at the single-cell level.

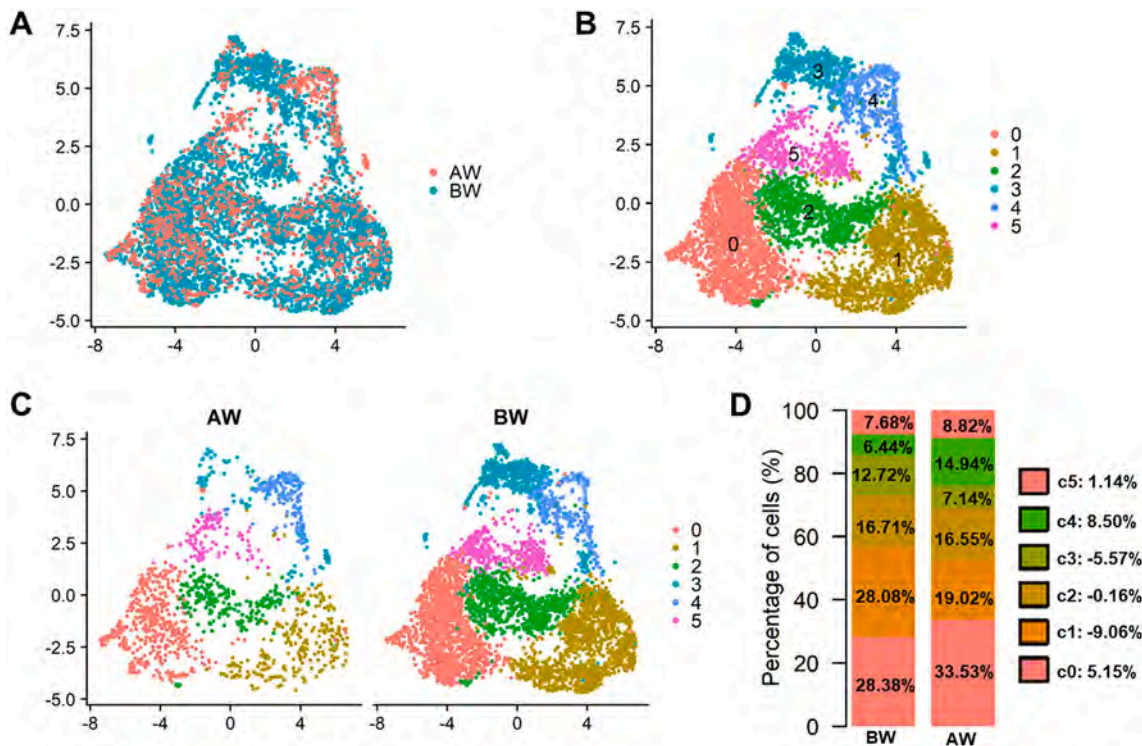


Fig. 1. Cluster analysis of single-cell transcriptomes from two calf rumen tissues. (A) UMAP projection plot showing dimensional reduction of the distribution of 6436 individual cell transcriptomes from two rumen tissues (green = before weaning; red = after weaning); (B) UMAP projection plot showing six major clusters of the 6436 individual cell transcriptomes; (C) UMAP projection plot showing annotation by before and after weaning rumen tissues. (D) The percentage of cell types across the pre-and post-weaning rumen tissues. The cell types were annotated based on (C). The numbers in the legend indicate the differences between the pre-and post-weaning rumen tissues within each cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Cell components and cell-cycle index for each cluster.

Cluster	Epithelial cells		Keratinocytes		Muscular cells		Others		All	Dividing cells
	Count	%	Count	%	Count	%	Count	%	Count	%
0	1740	91.72	27	1.42	130	6.85	0	0	1897	1.42
1	1593	94.65	17	1.01	71	4.22	2	0.12	1683	96.14
2	999	93.1	10	0.93	62	5.78	2	0.19	1073	20.32
3	614	82.75	27	3.64	87	11.73	14	1.89	742	21.29
4	337	63.47	113	21.28	45	8.47	36	6.78	531	41.81
5	351	68.82	43	8.43	102	20	14	2.75	510	32.75
Total	5634		237		497		68 ^a		6436	

^a Others included 35 Fibroblasts, 14 Myocytes, 7 Erythrocytes, 6 MEP and 1 Astrocyte, 1 Chondrocyte, 1 CLP, and 1 Endothelial cell. In each cluster, these other cell counts were fewer than 12.

2.3. Cell cycle analysis and SCENIC results for the rumen tissues

To explore the proliferation status of Seurat cell clusters, we performed the cell cycle analysis to calculate their cell cycle indices, using sets of 43 G1/S and 55 G2/M genes [38] (Table S3). The expression profiles of cell cycle-related genes revealed that the overall cell cycle indices were 1%, 96%, 20%, 21%, 42%, and 33% for Clusters C0 to C5, respectively (Fig. 2A and Table 1). These results suggested that C1 cells were actively dividing (96.14%), whereas C0 cells were not actively proliferating (1.42%). The cell cycle indices for C2 and C3 were around 20%, while those for C4 and C5 were 41.81% and 32.75%, respectively. Additionally, the average cell cycle indices were 75%, 59% for all BW or AW cells, respectively (Fig. S2A). Within each cell cluster, we also compared cell cycle indices between BW and AW cells. We found that cell cycle indices stayed similar for C0 and C1; increased for C3 (18% to 30%), but decreased for C2 (22% to 14%), C4 (46% to 12%), and C5 (55% to 18%) between BW and AW cells, respectively (Fig. S2B).

Furthermore, we performed another Seurat cell clustering after removing these cell cycle genes. When we compared the Seurat results with vs. without the cell cycle genes (Fig. S1D and Fig. S1E), we found that the global distribution patterns of BW and AW cells, as well as those of their corresponding cell clusters, were generally similar.

As important regulators of gene expression, transcription factors (TF) are very useful for identifying cell types. Thus, we performed the SCENIC analysis [34] to identify regulators and gene regulatory networks. Briefly, SCENIC infers co-expression modules between TF and candidate target genes using machine learning regression techniques (e.g., random forest or gradient boosting machines), which are pruned based on the enrichment of the TF motif around the TSS of the potential target genes, resulting in regulons. Based on the AUCell algorithm, SCENIC calculates the activity of each regulon in single-cell transcriptomes to obtain the corresponding area under the curve (AUC) scores, which are used to rank the cells for a given regulon and determine a threshold for active or inactive expression. Through this analysis, we identified 30 active

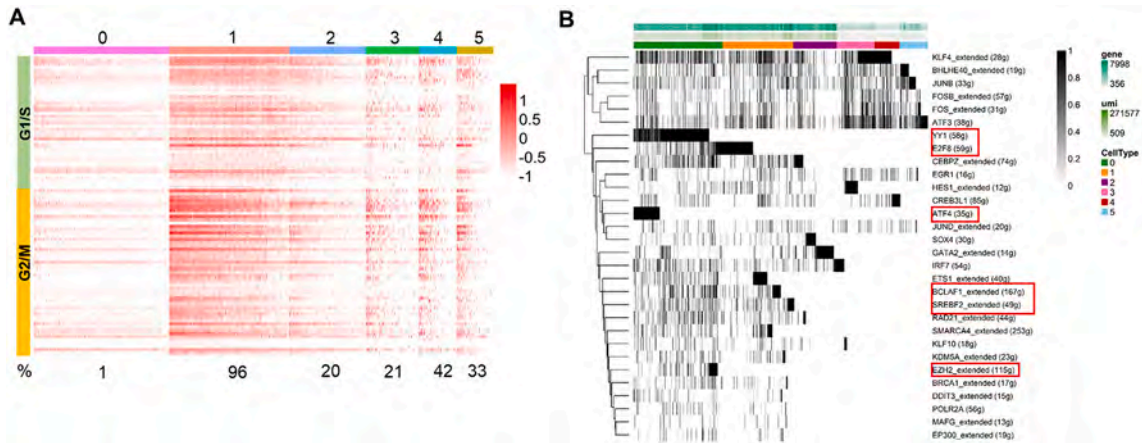


Fig. 2. Cell-cycle analysis and SCENIC results on the rumen tissues. (A) Heatmap showing expression levels of cell-cycle-related genes in each Seurat cluster. Cells were ordered according to the average expression level of cell-cycle-related genes within each cell. The color key from white to red indicated expression levels from low to high. The cell-cycle index of each cell type is shown at the bottom of the heatmap. (B) SCENIC binary regulon activity matrix shows all correlated regulons active in at least 1% of all regulons. Each column represents a single cell, and cluster labels correspond to those used in the UMAP plot. Representative transcription factors are highlighted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

regulons in the rumen (Fig. 2B). The count range of target genes of these regulons was between 12 and 252 (Table S4). SCENIC analysis revealed several important transcriptional regulators modulating cell type-specific gene regulatory networks. For Cluster C0, we identified its specific TF, including ATF4, EZH2, and YY1. For clusters combining C1 and C0, we detected E2F8, ETS1_extended, BCLAF1_extended, SREBP2_extended, SMARCA4_extended, KDM5A_extended, BRCA1_extended, DDIT3_extended, POLR2A, MAFG_extended, and EP300_extended. For Clusters C0, C1, and C2, especially for C2, we discovered CEBPZ_extended, SOX4, and GATA2_extended.

2.4. Marker gene expression for rumen cell clusters

To profile gene expression patterns of the different clusters identified by Seurat above (Table S5), we analyzed the expression of the top 10 marker genes in each cluster, as compared to all other clusters. Heatmap analysis revealed distinct signatures from each cluster (Fig. S3A). Of note, some of these top marker genes were highly expressed in only one

cluster, such as *UHRF1* (Ubiquitin Like With PHD And Ring Finger Domains 1) in C1 and *ACTA2* (Actin Alpha 2, Smooth Muscle) in C2, whereas other markers were conserved across two or more clusters, such as *FTH1* (Ferritin Heavy Chain 1), *MT-ND3* (Mitochondrially Encoded NADH: Ubiquinone Oxidoreductase Core Subunit 3), *MT-COX1* (official name *MT-CO1*: mitochondrially encoded cytochrome c oxidase I), and *RPLP1* (Large Ribosomal Subunit Protein P1 in Figs. S3B and S3C). Among these marker genes, there were some specific genes related to the cell cycle, such as *MKI67*, *HMMR*, *EZH2*, and *BRCA2*. We also detected epithelial cell marker genes, including *TGFβ1*, *TGFβ2*, and *TGFβR2*. Moreover, we obtained distinct sets of keratins for these cell clusters, some of which were up- or down-regulated in a specific cluster (Fig. 3).

2.5. Pseudotime analysis

In order to estimate the lineage relationships among the Seurat clusters and better understand the development states of all cells, we conducted a pseudotime analysis to infer the cell trajectories using

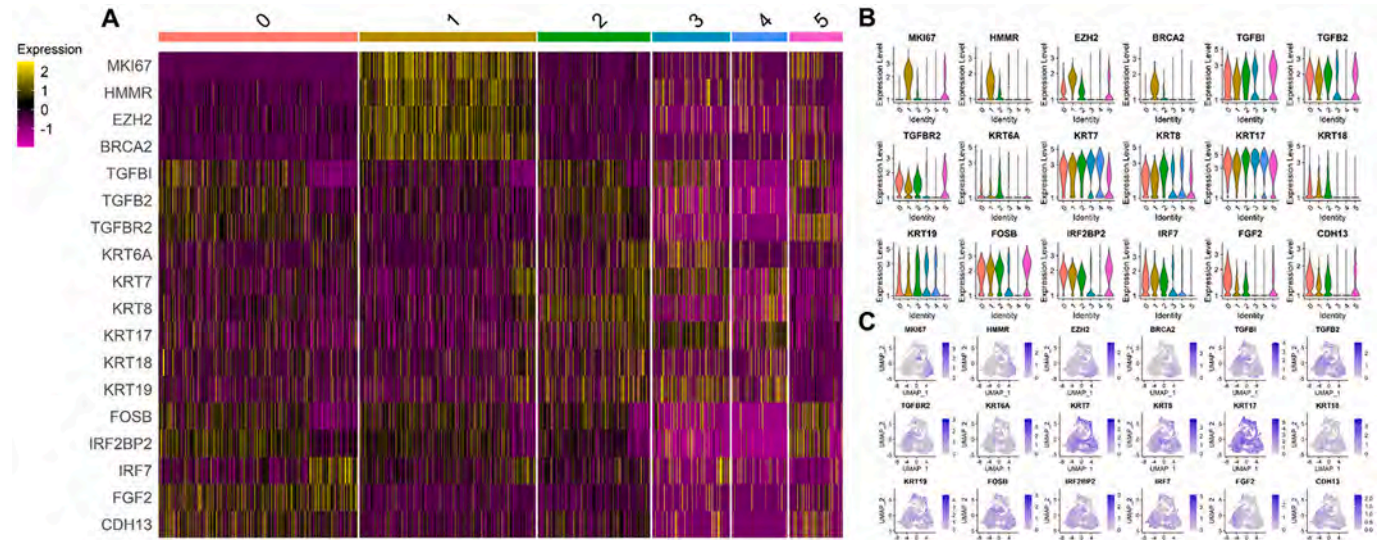


Fig. 3. Characterization of differential gene expressions for rumen tissues. (A) Gene expression heatmap of the 10 most differentially expressed genes in each cluster compared to all other clusters. Genes are represented in rows and cell clusters in columns. (B) Violin plots of gene expression. Expression in each cell is shown along with the probability density of gene expression, denoted by the shape of the plot. (C) UMAP projection plots showing transcript accumulation for cell marker genes in individual cells. Color intensity indicates the relative transcript level for the indicated gene in each cell.

Monocle 2 [35]. Following a “developmental/transitional” path according to their transcriptomic similarity, we identified one major and long-trajectory branch and one minor and short-trajectory branch, with cells ordered in an arrangement from proximal to distal distribution (Fig. 4A). Combining with the pseudotime values (Table S6), we observed that the long-trajectory tree rooted from C1, sprouted into C0, C3, and C4 (C1 → C0 → C3 → C4), while the short trajectory tree rooted from C2, sprouted into C5 (C2 → C5). The long path appeared to agree with our definitions on the Seurat clusters previously, i.e., from proliferating epithelial cells (C1) to resting epithelial cells (C0), to differentiated epithelial cells (C3), finally to keratinized epithelial cells (C4). The short path from C2 to C5 seemed to correspond to vascular smooth muscle development. C4 and C5 with the highest pseudotime scores might represent the terminal developmental states for either of these two paths, respectively.

2.6. Co-expression analyses

To systematically investigate the genetic program dynamics, we performed weighted gene co-expression network analysis (WGCNA) [39] using 2000 marker genes derived by Seurat. WGCNA identified 6 gene modules (Fig. 5A), each containing gene sets that tend to be co-expressed (Table S7). We then performed gene ontology (GO) analyses for genes in each module to investigate their biological functions (Fig. 5B, Table S8). To assign co-expressed gene functions to Seurat clusters, we generated a correlation heatmap in Fig. 5C. For example, the blue module genes were enriched for cell cycle and division, and the blue module was significantly represented in all cell clusters from C1 to C5, except for C0, likely corresponding to C0's resting nature. The brown module genes were enriched for epithelial cell proliferation, and the negative regulation of the developmental process and metabolic process. The blue module is significantly associated with all other Clusters, except for C4, indicating C4's terminal differentiation. The turquoise module genes were enriched for multiple GO terms, including epithelial cell mobility and differentiation, cell-cell junction and adhesion, cell division and death, and blood vessel development. This module was more correlated with all other clusters (correlation coefficients of 0.63–0.98) than with C1 (0.49), suggesting C1's dividing but undifferentiated states. The yellow module genes were enriched for extracellular matrix organization and this module was strongly associated with Clusters C3 and C4, probably related to their absorption and protection

mechanisms. Especially for C4, the yellow module was most correlated, suggesting its keratinized epithelial cells could provide skin-like function inside rumen and further support our C4 assignment.

2.7. Trait-relevant cell clusters

Using a permutation-based marker-set test approach (Methods), we tested the enrichment of 45 GWAS signals within marker genes of distinct clusters (Table S5) reported by Seurat (FDR < 0.05) (Fig. 6A). Production and body conformation traits were significantly associated with all clusters, especially C5, C0, and C4, reflecting the important functions of these cell types related to SCFA absorption and tissue development. In addition, health traits, such as SCS (somatic cell score) were associated with all clusters except C1 and C2, suggesting that the differentiated and terminal cell types have a role in tissue integrity and immunity. This might reflect that rumen plays a role in the regulation of immunity. Moreover, based on the marker genes reported by edgeR (Table S9) between cell clusters across the BW and AW rumen samples, we also detected similar results (Fig. 6B).

2.8. Cross-species comparison

To support our cattle rumen results using human data, we downloaded the scRNA-seq dataset of the human stomach from GSE134355 [22] and performed Seurat clustering analysis [26]. Plotting the single-cell transcriptomes via UMAP projection yielded largely overlapping distributions of cells from 2 cattle and 7 human samples (Fig. S5A), validating our scRNA-seq data generation, processing, and cell type assignment. In total, we identified 13 distinct clusters (Fig. S5A). Using SingleR [30], we obtained the 29,926 individual cell transcriptomes of 42 cell types from the nine samples (Fig. S5B). The UMAP plot distribution reflected that the main cell types were adipocytes, epithelial cells, keratinocytes, mesangial cells, monocytes, plasma cells, and skeletal muscle (Fig. S5B). Within them, we could validate epithelial cells, keratinocytes, and mesangial cells identified in our cattle rumen samples.

3. Discussion

A previous human study showed that stratified squamous epithelia of internal organs are generally similar to skin, although they make

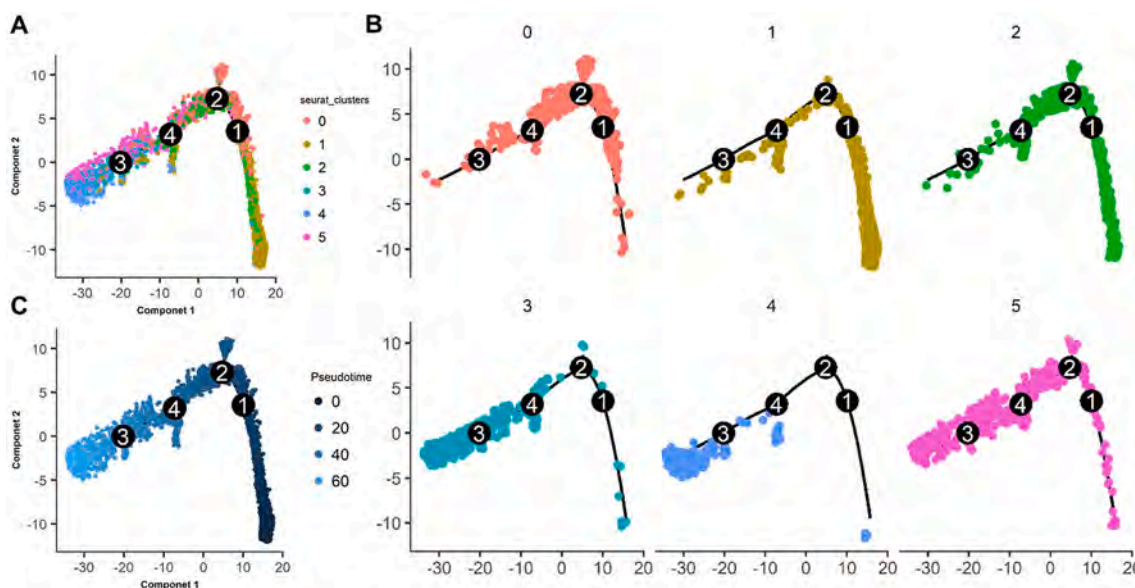


Fig. 4. Pseudotime analysis using Monocle 2 for cell transcriptomes. Solid black lines indicate the main diameter path of the minimum spanning tree (MST) and provide the backbone of Monocle's pseudotime ordering of the cells. Each dot represents an individual cell colored by cluster (A), or pseudotime (B), or pseudotime (C).

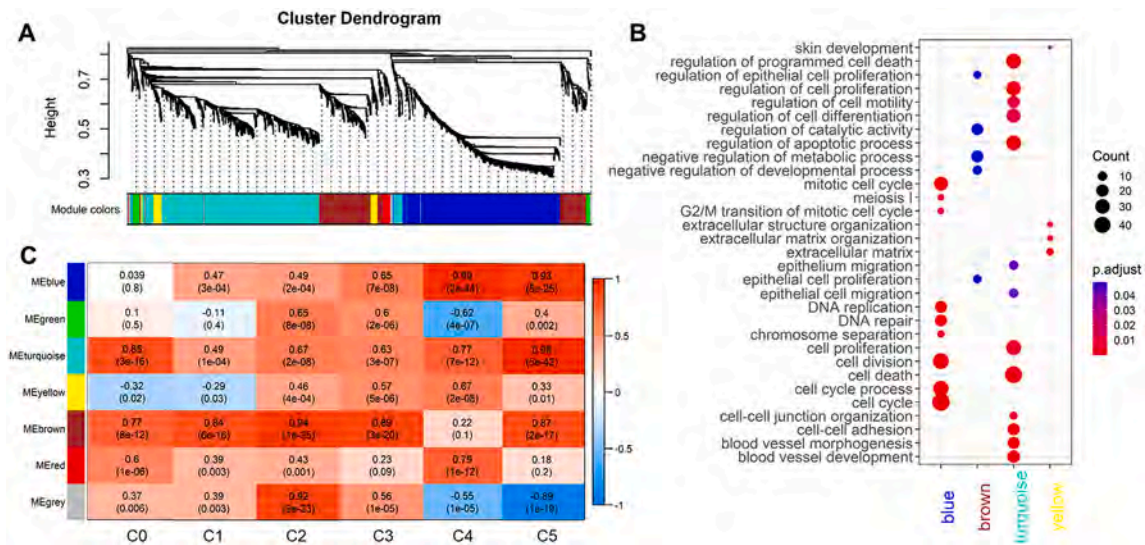


Fig. 5. WGCNA suggested genetic networks. (A) Dendrogram showing the gene co-expression network constructed using WGCNA. The color bar labeled as “Module colors” beneath the dendrogram represents the module assignment of each gene. (B) Significantly enriched GO terms based on genes within each module. (C) The relationship between Modules and Seurat clusters. The upper numbers within each grid are the correlation between each module and Seurat cluster. The numbers in brackets represent the *P*-values.

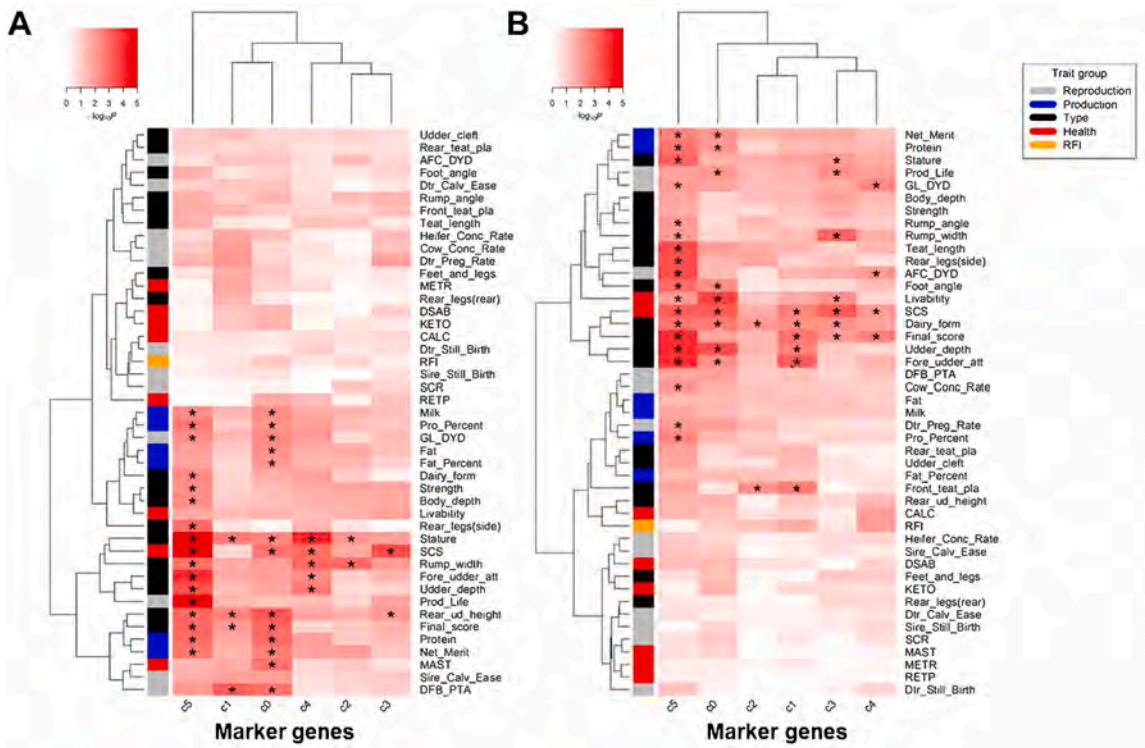


Fig. 6. Associations of cell clusters with complex traits based on GWAS signal enrichment analyses using marker genes among cell clusters (A) and between pre-and post-weaning (top 5%) (B). “*” denotes *FDR* < 0.05.

different types of keratins [40]. An early study reported influences of extracellular matrix components on the growth and differentiation of ruminal epithelial cells in primary culture [41]. Xiang et al. studied the response to and regulation of the layers of the full rumen wall to different diets fed to sheep, using bulk RNA-s [11]. They identified clusters of genes characteristic of cell proliferation and differentiation, as well as metabolism-specific genes within the epithelium network. The expressions of cell-cycle and metabolic genes were positively correlated with dry matter intake, ruminal SCFA concentrations, and methane

production. They reported that the TF expression patterns and their targets, ruminal epithelium, mimic proliferating and differentiating skin, suggesting conservation of regulatory networks. From an evolutionary perspective, Pan et al. reported that a positively selected and ruminant-specific gene, *WDR66*, regulates the expression of occludin, which then tightens the intercellular space and controls epithelial permeability [42]. They speculated that when junction structure (desmosome) between keratinocytes of the ruminal epithelium becomes loose, the enlarged intercellular space with its copious blood supply

enables nutrient absorption across the ruminal epithelium. However, all the above studies were performed with bulk samples. Cell-type profiles for cattle rumen epithelial cells at a single-cell resolution are necessary to compare the undeveloped ruminal epithelium, consisting primarily of strata basal and spinosum cell types, and fully differentiated ruminal epithelium, comprising four strata with highly variable cell content [15].

Using the 10× Genomics Chromium Controller, we did scRNA-seq on Holstein ruminal epithelial cells during weaning. To our knowledge, this represents the first reported single-cell transcriptomic analysis in cattle. Our study was successful in generating rumen single-cell transcriptomes, revealing major and some novel cell types. In the current study, we identified 6 distinct cell clusters and tentatively assigned their cell types using Human Cell Atlas/Blueprint reference cell datasets (Fig. 1 and Table 1). For mesangial cell assignment, because we used human reference cell types to assign cattle cells, these designations may be biased towards human structure and function. Even in humans, mesangial cells are believed to share the same origin as vascular smooth muscle cells and are sometimes considered to be a type of specialized vascular smooth muscle cell [43]. Therefore, it might be more helpful to rename those cells as vascular smooth muscle cells despite their mesangial cell assignments. We also detected thousands of marker genes among cell clusters during weaning (Table S1). We then performed GO and KEGG-based gene enrichment and cell cycle analyses (Fig. S4 and Fig. 2A). In the co-expression analyses (Fig. 5), we obtained 6 distinct modules (Fig. 5A) and significantly enriched GO terms based on genes within each module (Fig. 5B). We then assigned co-expressed gene functions to specific cell clusters (Fig. 5C). When we integrated these marker genes with Holstein GWAS signals, we observed all clusters, especially C5 and C0, were enriched for animal production and body type traits. Additionally, we also substantiated the cattle cell identities by comparing them with the human and mouse stomach epithelial cells.

We found that Cluster C1 contained 94.65% epithelial cells and 96.14% cells were dividing. C1-specific genes were enriched for cell cycle, chromosome segregation, cytoskeleton, DNA replication, nuclear division, etc. (Fig. S4 C1). C0 contained 91.72% epithelial cells, but only 1.42% cells were dividing, and C0-specific genes were enriched for RNA binding, localization, and degradation, cytosolic ribosome, regulation of peptidase activity, and metabolic processes (Fig. S4 C0). C2 contained 93.10% epithelial cells, 0.93% keratinized epithelial cells, 5.79% vascular muscle cells, and 20.32% cells were dividing. C2-specific genes were enriched for cell differentiation processes, including myofibril, smooth muscle proliferation and contraction, extracellular matrix-receptor interaction, regeneration, and cell cycle (Fig. S4 C2). C3 contained 82.75% epithelial cells, 3.64% keratinized epithelial cells, 11.73% vascular muscle cells, and 21.29% cells were dividing. C3-specific genes were enriched for cell aging, positive regulation of establishment of protein localization to the telomere, insulin-like growth factor binding, response to cytokine, and interaction with symbiont (Fig. S4 C3). C4 contained 63.47% epithelial cells, 22.28% keratinized epithelial cells, 8.47% of vascular muscle cells, and 41.81% cells were dividing. Notably, C4-specific genes were enriched for skin development (keratinization), response to hypoxia, extracellular matrix organization, apoptotic process, cell cycle and death, as well as cell adhesion and migration (Fig. S4 C4). C5 contained 68.82% epithelial cells, 8.43% keratinized epithelial cells, 20.00% vascular muscle cells, and 32.75% cells were dividing. C5-specific genes were enriched for mitochondrial respiratory chain complex I assembly and peptide and protein metabolic process (Fig. S4 C5).

To estimate the effects of the cell cycle genes on cell clustering, we performed Seurat cell clustering with or without these cell cycle genes. Our cell clustering results showed that there were no significant distribution differences for BW or AW samples, or for six rumen cell types during the weaning process (Fig. S1 D and E). These results indicated that the cell clustering and type assignment mainly reflected physiological differences among these cell types (cell differentiation), rather

than the effects of the cell cycle statuses (cell division) under our analysis conditions. Based on these results and existing literature, we proposed the following model for cattle rumen epithelial development (Fig. S6). We designate cell types to Seurat clusters along 2 lineages in the following networks, which can better reflect the temporal and spatial distributions for the ruminal epithelium layer. Lineage 1 is C1 → C0 → C3 → C4, including proliferating epithelial cells (C1), resting poised epithelial cells (C0), differentiated epithelial cells (C3), and keratinized epithelial cells (C4). Lineage 2 is C2 → C5, including vascular muscle precursor cells (C2) and vascular muscle cells (C5). As C2 appeared at roughly the same time as C1, and both were earlier than C0, it is less likely that C2 was derived from C0. However, the relationship between C1 and C2 was not clear, even though the overwhelming majority of their cells were the same type (epithelial cells) at 94.65% and 93.10%, respectively. Thus, we speculated that C1 and C2 could be derived from the same epithelial stem cells, which linked Lineages 1 and 2 together.

With the above model in mind, we checked cell-type-specific marker genes and TF. For cell cycle-related TF, we could readily identify them from Cluster C1's marker genes and SCENIC results, including MKI67 (ranked as No. 8 by its *P*-value), HMMR (No. 54), and EZH2 (No. 70). These three TFs were the same as reported by Xiang et al. in sheep rumen feed efficiency research [11]. They also reported the cell cycle regulator, BRCA1, was present in sheep, while we found its close relative, BRCA2, in cattle.

For epithelial cell marker genes, we detected transforming growth factor-beta receptors or ligands, such as TGFβ1, TGFβ2, and TGFβR2. Their expressions were decreased in C4, while expressions of TGFβ1 and TGFβR2 were increased in C5. The TGFβ1 (Transforming growth factor, beta-induced) protein contains the common peptide motif (arginylglycylaspartic acid - RGD), which binds to type I, II, and IV collagens. Previous studies reported that the TGFβ1 protein is secreted, induced by TGFβ, and associated with normal skin and adhesion of dermal fibroblasts [44] or keratinocytes [45]. Bond et al. reported their first discovery of TGFβ1 in rumen epithelium, possibly modulating cell adhesion [46]. The TGFβ superfamily is critical in wound healing and repair. It must be activated by release from the extracellular matrix where it is bound by latent TGFβ-binding proteins and active proteases, such as the matrix metalloproteinase [47]. TGFβ has been shown to inhibit the proliferation of keratinocytes [48,49]. Additionally, in humans, ligands TGFβ1, TGFβ2, and TGFβ3 all function through the same receptor signaling systems. This pathway is involved in many cellular processes in both the adult organism and the developing embryo, including cell growth, cell differentiation, apoptosis, cellular homeostasis, and other cellular functions. We previously reported that TGFβ1 is an important transcriptional regulator of gene expression networks related to certain diets using the same calf rumen epithelium samples during weaning [50]. Our rediscovery of the same TGFβ pathways from the scRNA-seq assay, further confirmed that these cytokines and their related proteins are likely involved in regulating the growth and differentiation of the rumen epithelium. In C4 and C5, their expression repressions and inductions may correspond to the different cell specializations for keratinized epithelial cells and vascular smooth muscle cells, respectively. Further characterization of TGFβ pathway gene expression and distribution within the extracellular matrix and among the layers of the rumen epithelium during proliferation and differentiation is needed to better understand its function in the ruminal mucosa related to these physiological processes.

Interestingly, we also obtained distinct sets of keratins from the marker genes among these cell clusters, such as C1: down-regulation of KRT17; C2: up-regulation of KRT8, KRT17; C3: up-regulation of KRT17, but down-regulation of KRT6A; and C5: down-regulation of KRT7, KRT19, KRT8, KRT18, and KRT17. For example, KRT8 is well-known to be expressed in epithelial cells of the human gastrointestinal tract (including stomach, colon, small intestine, gall bladder, liver, and pancreas) and mammary gland ducts [51]. Additionally, we found *FOSB*

(FosB proto-oncogene, AP-1 transcription factor subunit) was down-regulated in C4 but up-regulated in C5, which was reported to play a role in epithelial proliferation and differentiation [11]. We also detected that interferon *IRF2BP2*, which is known to be involved in immunity, was decreased in C4, and *IRF7* was also decreased in C5. At the same time, we also obtained distinct sets of genes from Cluster marker genes, like *FGF2* (Fibroblast Growth Factor 2, which is related to *FGF7* – a potent epithelial cell-specific growth factor). *FGF2* gene expression was decreased in C1 but increased in C0. *FGF2*'s other close relatives include *FGF6*, which has been shown to be associated with cattle traits, like body depth, rump width, sire calving ease, stature, and others [52]. *CDH13* (cadherin 13, related to the cell adhesion protein E-Cadherin) was down-regulated in C4.

Within the marker genes reported among cell clusters, the *CENPF* gene encodes a protein that associates with the centromere-kinetochore complex and it may play a role in chromosome segregation during mitosis [53] and *NPM1* encoded a protein that is involved in several cellular processes, including centrosome duplication and cell proliferation [54]. We also discovered important TF modulating cell-type-specific gene regulatory networks. For Cluster C0, we identified its specific TF, including ATF4, EZH2, and YY1. *EZH2* encodes a member of the Polycomb-group (PcG) family, which maintains the transcriptionally repressive state. YY1 is a ubiquitously distributed transcription factor belonging to the GLI-Kruppel class of zinc finger proteins, which can activate or repress the promoter [55]. For clusters combining C1 and C0, we detected BCLAF1, BRCA1, SMARCA4, EP300, and other TF. *BRCA1* encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor [56]. *BCLAF1* encodes a nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. As a member of the BRCA1-associated genome surveillance complex (BASC), the gene product plays a role in transcription, DNA repair of double-stranded breaks, and recombination [57]. Both SMARCA4 and EP300 regulate transcription via chromatin remodeling and are important in cell proliferation and differentiation [58,59]. For Clusters C0, C1, and C2, especially for C2, we discovered CEBPZ, SOX4, and GATA2, all of which are important transcriptional regulators for cell growth and differentiation [60–62].

Conclusions: In summary, this study provides an initial example for bovine single-cell analysis and opens the door for new discoveries about tissue/cell type roles in complex traits at single-cell resolution. We provided the first cell type profiles for cattle rumen epithelial cells at a single-cell resolution. We characterized their cell cycle, component, relative timing, and regulatory networks, as well as co-expression and gene function patterns. With our proposed cell lineage development model, we reported 6 cell types identified across their temporal and spatial distributions, which appear to be correlated with the rumen epithelium's underlying layers, structures, and functions. This rumen cell development model will need to be further tested and improved by more replicates and functional validations. For example, spatial transcriptomics data will be needed to locate the relative position for each cell cluster over the development stages. More future experiments are warranted to investigate the mechanisms, which regulate the commitment of epithelial stem cells to differentiate in distinct lineages.

4. Methods

4.1. Sample collection

Animals and tissue collection were fully described in our previous report [8]. Briefly, two Holstein bull calves were chosen: one calf (pre-weaning) was fed with milk replacer only (MRO - Cornerstone 22:20, Purina Mills, St. Louis, MO, USA; 22.0% crude protein, 20.0% crude fat, 0.15% crude fiber, 0.75 to 1.25% Ca, 0.70% P, 66,000 IU/kg vitamin A, 11,000 IU/kg vitamin D3, and 220 IU/kg vitamin E) for two weeks; while the other (post-weaning) was fed with MRO for six weeks,

followed by a combination of milk replacer and grain-based commercial calf starter for four weeks. Calves were euthanized by captive bolt followed by exsanguination at day 14 or day 70 to represent development at two stages of weaning on a grain concentrate diet. Rumen epithelial tissue was collected from the anterior portion of the ventral sac of the rumen beneath the reticulum and below the rumen fluid layer at slaughter. The epithelial layer of the rumen tissue was separated manually from the muscular layer. After rinsing with tap water to remove residual feed particles, samples were further rinsed in ice-cold physiological saline, and subsamples of epithelial tissues (approximately 600 mg) were fixed in RNAlater (Life Technologies, Grand Island, NY, USA) RNA stabilization solution according to the manufacturer's instructions and stored at -80°C until use.

4.2. Single-cell isolation and RNA-seq library preparation and sequencing

Rumen tissue samples from the one pre-weaned and one weaned calf were collected and processed by a commercial service provider, Singulomics (New York, NY, USA), for scRNA-seq analysis. Library preparation was performed according to instructions by using the 10× Genomics Chromium single-cell controller. The libraries were then pooled and sequenced on a HiSeq4000 (Illumina, San Diego, CA, USA).

4.3. Generation of single-cell transcriptomes

We first processed 10× Genomics raw data by the Cell Ranger Single-Cell Software Suite (release 3.1.0), including using Cell Ranger *mkfastq* to demultiplex raw base-call files into FASTQ files followed by the use of Cell Ranger *count* to perform alignment, filtering, barcode counting, and UMI counting. The raw reads were aligned to the ARS-USD1.2 cattle reference genome [63] by Cell Ranger *pipeline* using default parameters. The output summary of the two samples is shown in Supplemental Table 1. All downstream single-cell analyses were performed using the Seurat 3.0 [26] R package v3.6.3 unless explicitly mentioned.

4.4. Quality control, dimension reduction, and cell clustering

Overall, 5064 and 1372 cells passed the following quality control thresholds: all genes expressed in fewer than 3 cells were removed; the number of genes expressed per cell >200 as low and <8000 as high cut-off; UMI counts less than 200; the percent of mitochondrial-DNA derived gene-expression $<30\%$. We used the LogNormalize method of the “Normalization” function to calculate the expression value of genes. We then restricted the corrected expression matrix to the subsets of highly variable genes (HVG), and centered and scaled values before performing dimension-reduction and clustering on them. We selected 2000 genes as HVG using the “FindVariableFeatures” function with default parameters. We then used the “RunPCA” function to perform the principal components analysis (PCA) on the single-cell expression matrix with genes restricted to HVG. The number of significant principal components was determined using a permutation test implemented by the “JackStraw” function. Within all the PC, the top 10 PC were used for clustering and Uniform Manifold Approximation and Projection (UMAP) analysis. To find clusters, we used the weighted Shared Nearest Neighbor (SNN) graph-based clustering method implemented by the “FindNeighbors” function and then utilized the “FindClusters” function to conduct the cell-clustering analysis by embedding cells into a graph structure in PCA space. Based on the number of cells in our study, we set the parameter resolution to 0.24. Visualization of the cells was performed using the UMAP algorithm as implemented by the Seurat “RunUMAP” function.

4.5. Assigning cell types to single-cell clusters

Two methods were utilized to assign the cell clusters identified by Seurat. First, canonical cell-type marker genes that are conserved across

conditions were identified using the “FindConservedMarkers” function with default parameters. Marker genes with significant specificity to each cluster were annotated with their known functions. Additionally, raw expression data for the filtered cells were used for cell type assignment using SingleR [30] with default parameters using the Blueprint [37] and Encode [33] human cell atlases. To compare scRNA-seq gene expression levels with bulk RNA-seq data, we calculated average gene expression values across all cells or cells within a cluster. Pearson correlation coefficients were calculated between snRNA-seq and bulk RNA-seq values for all genes expressed in both data sets.

4.6. Pseudotime trajectory analysis

For trajectory analysis, we used Monocle 2 [35] to order cells in pseudotime based on their transcriptional similarities, with UMI counts modeled using a negative binomial distribution. First, we integrated the preprocessed Seurat objects into Monocle 2, utilizing the “new-CellDataSet” function. We then determined the differentially expressed genes or marker genes that were identified using the “differentialGeneTest” function. We next reduced the dimensionality of the data to two dimensions using the discriminative dimensionality reduction with trees (DDRTree) method implemented in the “reduceDimension” function. Finally, after pseudotime calculations were made for each cell, we projected clusters derived from the Seurat object onto the minimum spanning tree upon cell order using the “plot_cell_trajectory” function.

4.7. Cell-cycle analysis

Sets of 43 G1/S and 55 G2/M genes [38] were used in the cell-cycle analysis. To calculate the ratio of actively proliferating cells of each feature, such as different clusters and different weaning stages, we first calculated the total expression levels of all 98 cell-cycle genes in every single cell, and only cells with mean expression levels higher than the average values of all clusters were regarded as actively proliferating.

4.8. Single-cell regulatory network inference and clustering (SCENIC) analysis

We conducted SCENIC analysis on cells after filtering for each major cell type using the R package SCENIC v1.1.2 [34], which is a computational workflow that predicts TF activities from scRNA-seq data. Instead of interrogating predefined regulons, individual regulons are constructed from the scRNA-seq data. Regions for TF searching were restricted to a 10 kb distance centered on the transcriptional start site (TSS) or 500 bp upstream of the TSS. First, TF-gene co-expression modules are defined in a data-driven manner with GENIE3 v1.8.0. Subsequently, those modules are refined via RcisTarget by keeping only those genes that contain the respective transcription factor binding motif (TFBS). Once the regulons are constructed, the method AUCell scores individual cells by assessing for each TF separately whether target genes are enriched in the top quantile of the cell signature.

4.9. Weighted gene co-expression network analysis

Weighted gene co-expression network analysis (WGCNA) was performed with functions in the WGCNA v1.69 R package following the previously published study by Tosches and colleagues [64]. According to the methods, the analyses were performed on pseudocells, calculated as averages of 100 cells randomly chosen within each cluster. The top 2000 highly variably expressed genes determined in Seurat were used for analysis. Briefly, the topological overlap matrix (TOM) was constructed with softPower and was set to 2. The hub genes for each module were identified as module eigengene. The GO enrichment analysis was performed by ClusterProfiler [65] R package using hub gene data sets and the BH method was employed for multiple test correction. GO terms

with a *P*-value lower than 0.05 were considered significantly enriched.

4.10. Gene function analysis

To get the lists of marker genes, we first extracted the genes' UMIs across cells within each cluster and then assigned cells to the BW or AW group. Based on the gene x cells matrix, we utilized edgeR to detect marker genes for each cluster between BW and AW (Table S9). We used lists of genes differentially expressed in each of the six clusters for GO and KEGG using Cytoscape 3.8.0 analyses with the ClueGO app [66]. Fisher exact test was used to measure gene enrichment in annotation terms. FDR corrected *P*-values were used to search for significantly enriched terms. GO terms and KEGG pathways with a *P*-value lower than 0.05 were considered significantly enriched (Table S10).

4.11. GWAS signal enrichment analysis

We previously reported details of the single-marker GWAS and fine-mapping analyses for the body type, reproduction, and production traits from 27,214 U.S. Holstein bulls, for health traits from 11,880–24,699 bulls, and for feed efficiency (i.e., RFI) from 3947 Holstein cows [52,67–69]. Because the complex traits being studied here are highly polygenic, we applied the sum-based marker-set test approach shown in eq. 1, as implemented in QGG package v1.0 [70], to determine whether GWAS signals were enriched in marker genes of distinct cell clusters and marker genes of AW vs. BW. We added 20-kb windows around gene regions to include the potential *cis*-regulatory variants. Previous studies showed that this approach had at least equal power when compared to other commonly used GWAS signal enrichment methods in humans [71,72], *Drosophila melanogaster* [73], and livestock [74–76], especially for the highly polygenic traits.

$$T_{sum} = \sum_{i=1}^{m_f} b_i^2 \quad (1)$$

In this expression, m_f is the number of genomic markers within a list of genes (marker genes of each cell cluster or marker genes of AW vs. BW in each cell cluster), and b is the marker effect from single-marker GWAS. We controlled marker-set sizes and linkage disequilibrium patterns among markers through applying the following genotype cyclical permutation strategy [70]. Briefly, we first ordered marker effects (i.e., b^2) using their chromosome positions (i.e., $b_1^2, b_2^2, \dots, b_{m-1}^2, b_m^2$). We then randomly selected one marker (i.e., b_k^2) from this vector as the first place, and shifted the remaining ones to new positions, while retaining their original orders (i.e., $b_k^2, b_{k+1}^2, \dots, b_{m-1}^2, b_m^2, b_1^2, \dots, b_{k-1}^2$) to maintain LD patterns among markers. We calculated a new summary statistic for a given list of genes using their original chromosome locations. To obtain an empirical *P*-value for the list of genes, we repeated this permutation procedure 10,000 times. We employed a one-tailed test of the proportion of random summary statistics greater than that observed.

4.12. Cross-species comparison

We downloaded a single-cell RNA-seq dataset of the human stomach from GSE134355. We first merged expression matrices of the two species (cattle and humans) based on the detected homologous genes' intersection. Next, we performed expression matrix preprocessing separately for the two species using the Seurat v3 R package, followed by integrating three datasets using functions in Seurat v3 [26]. The resolution was set to 0.4 to yield 13 cell clusters.

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ygeno.2021.04.039>.

Ethics approval and consent to participate

All samples were collected with the approval of the US Department of

Agriculture (USDA) Agriculture Research Service (ARS) Institutional Animal Care and Use Committee under Protocol 07–025. Consent to participate: not applicable.

Consent for publication

Not applicable.

Availability of data and material

All cell and gene expression matrix data are available as supplemental data S1. The GWAS summary statistics for all complex traits have been submitted to Figshare, i.e., body type, production, and reproduction traits under <https://figshare.com/s/ea726fa95a5bac158ac1>, and the remaining ones under <https://figshare.com/s/94540148512ddd7ed32>. All scripts and source codes can be found in the Supplemental Material, as well as in <https://github.com/YahGao/Rumen-scRNA-seq>.

Author statement

None.

Funding

This work was supported in part by USDA National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (AFRI) grant numbers 2013-67015-20951, 2016-67015-24886, 2019-67015-29321, and 2021-67015-33409 and the US-Israel Binational Agricultural Research and Development (BARD) Fund grant number US-4997-17. This work was also supported in part by USDA ARS appropriated projects 8042-31000-001-00-D, 8042-31000-002-00-D, and 8042-31310-078-00-D.

Authors' contributions

GEL and CJL conceived and designed the experiments. EEC, RLB, and JBC collected samples and/or generated NGS data. YG, LF, CJL, CPVT, LM, and GEL performed in silico prediction and computational analyses. YG, CJL, and GEL wrote the paper. All authors read and approved the final manuscript.

Declaration of Competing Interest

All authors declare no potential conflict of interest.

Acknowledgements

We thank Reuben Anderson, Mary Bowman, Donald Carbaugh, Christina Clover, Sarah McQueeney, Mary Niland, Marsha Campbell, Dennis Hucht, and Research Animal Services staff at the Beltsville Dairy Unit for technical assistance. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture (USDA). The USDA is an equal opportunity provider and employer.

References

- [1] L. Lin, F. Xie, D. Sun, J. Liu, W. Zhu, S. Mao, Ruminal microbiome-host crosstalk stimulates the development of the ruminal epithelium in a lamb model, *Microbiome* 7 (1) (2019) 83.
- [2] S. Lin, L. Fang, X. Kang, S. Liu, M. Liu, E.E. Connor, R.L. VI Baldwin, G. Liu, C.J. Li, Establishment and transcriptomic analyses of a cattle rumen epithelial primary cells (REPC) culture by bulk and single-cell RNA sequencing to elucidate interactions of butyrate and rumen development, *Heliyon* 6 (6) (2020), e04112.
- [3] R.L. VI Baldwin, Use of isolated ruminal epithelial cells in the study of rumen metabolism, *J. Nutr.* 28 (2) (1998) 293S–296S.
- [4] P. Gálfi, S. Neogrady, T. Sakata, 3 - Effects of volatile fatty acids on the epithelial cell proliferation of the digestive tract and its hormonal mediation, in: T. Tsuda, Y. Sasaki, R. Kawashima (Eds.), *Physiological Aspects of Digestion and Metabolism in Ruminants*, Academic Press, San Diego, 1991, pp. 49–59.
- [5] C. Stevens, Fatty acid transport through the rumen epithelium, *Physiol. Digest. Metab. Ruminant* (1970) 101–112.
- [6] R.L. VI Baldwin, E.E. Connor, Rumen function and development, *Vet. Clin. N. Am. Food Anim. Pract.* 33 (3) (2017) 427–439.
- [7] E.N. Bergman, Energy contributions of volatile fatty acids from the gastrointestinal tract in various species, *Physiol. Rev.* 70 (2) (1990) 567–590.
- [8] A.T. Phillipson, Physiology of digestion and metabolism in the ruminant. Proceedings of the Third International Symposium, Cambridge, August 1969, in: *Physiology of digestion and metabolism in the ruminant Proceedings of the Third International Symposium*, Cambridge, August 1969, Oriel Press Ltd., 32 Ridley Place, Newcastle upon Tyne, NE1 8LH, 1970.
- [9] R.L. VI Baldwin, S. Wu, W. Li, C. Li, B.J. Bequette, R.W. Li, Quantification of transcriptome responses of the rumen epithelium to butyrate infusion using RNA-seq technology, *Gene Regul. Syst. Biol.* 6 (2012) 67–80.
- [10] A. Naeem, J.K. Drackley, J.S. Lanier, R.E. Everts, S.L. Rodriguez-Zas, J.J. Loo, Ruminal epithelium transcriptome dynamics in response to plane of nutrition and age in young Holstein calves, *Funct. Integr. Genomics* 14 (1) (2014) 261–273.
- [11] R. Xiang, J. McNally, S. Rowe, A. Jonker, C.S. Pinares-Patino, V.H. Oddy, P. E. Vercoe, J.C. McEwan, B.P. Dalrymple, Gene network analysis identifies rumen epithelial cell proliferation, differentiation and metabolic pathways perturbed by diet and correlated with methane production, *Sci. Rep.* 6 (2016) 39022.
- [12] K. Zhao, Y.H. Chen, G.B. Penner, M. Oba, L.L. Guan, Transcriptome analysis of ruminal epithelia revealed potential regulatory mechanisms involved in host adaptation to gradual high fermentable dietary transition in beef cattle, *BMC Genomics* 18 (2017) 976.
- [13] W. Li, S. Gelsinger, A. Edwards, C. Riehle, D. Koch, Changes in meta-transcriptome of rumen epimural microbial community and liver transcriptome in young calves with feed induced acidosis, *Sci. Rep.* 9 (1) (2019) 18967.
- [14] I. Kanter, T. Kalisky, Single cell transcriptomics: methods and applications, *Front. Oncol.* 5 (2015) 53.
- [15] R.L. VI Baldwin, Use of isolated ruminal epithelial cells in the study of rumen metabolism, *J. Nutr.* 128 (2) (1998) 293S–296S.
- [16] L. Fang, S. Liu, M. Liu, X. Kang, S. Lin, B. Li, E.E. Connor, R.L. VI Baldwin, A. Tenesa, L. Ma, et al., Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations, *BMC Biol.* 17 (1) (2019) 68.
- [17] M.J. Peters, R. Joeanes, L.C. Pilling, C. Schurmann, K.N. Conneely, J. Powell, E. Reinmaa, G.L. Sutphin, A. Zhernakova, K. Schramm, et al., The transcriptional landscape of age in human peripheral blood, *Nat. Commun.* 6 (2015) 8570.
- [18] H. Dueck, M. Khaladkar, T.K. Kim, J.M. Spaethling, C. Francis, S. Suresh, S. A. Fisher, P. Seale, S.G. Beck, T. Bartfai, et al., Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation, *Genome Biol.* 16 (2015) 122.
- [19] F. Zhou, X. Li, W. Wang, P. Zhu, J. Zhou, W. He, M. Ding, F. Xiong, X. Zheng, Z. Li, et al., Tracing haematopoietic stem cell formation at single-cell resolution, *Nature* 537 (7604) (2016) 487–492.
- [20] L. Li, J. Dong, L. Yan, J. Yong, X. Liu, Y. Hu, X. Fan, X. Wu, H. Guo, X. Wang, et al., Single-cell RNA-seq analysis maps development of human Germline cells and gonadal niche interactions, *Cell Stem Cell* 20 (6) (2017) 858–873 (e854).
- [21] A.K. Shalek, R. Satija, J. Shuga, J.J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J.T. Gaublot, N. Yosef, et al., Single-cell RNA-seq reveals dynamic paracrine control of cellular variation, *Nature* 510 (7505) (2014) 363–369.
- [22] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, et al., Construction of a human cell landscape at single-cell level, *Nature* 581 (7808) (2020) 303–309.
- [23] A.L. Haber, M. Biton, N. Rogel, R.H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M.R. Howitt, Y. Katz, et al., A single-cell survey of the small intestinal epithelium, *Nature* 551 (7680) (2017) 333–339.
- [24] S. Gao, L. Yan, R. Wang, J. Li, J. Yong, X. Zhou, Y. Wei, X. Wu, X. Wang, X. Fan, et al., Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing, *Nat. Cell Biol.* 20 (6) (2018) 721–734.
- [25] J. Chen, B.T. Lau, N. Andor, S.M. Grimes, C. Handy, C. Wood-Bouwens, H.P. Ji, Single-cell transcriptome analysis identifies distinct cell types and niche signaling in a primary gastric organoid model, *Sci. Rep.* 9 (1) (2019) 4536.
- [26] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W.M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data, *Cell* 177 (7) (2019) 1888–1902 (e1821).
- [27] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, *J. Stat. Mech.: Theory Exp.* (2008) 10.
- [28] A. Duo, M.D. Robinson, C. Soneson, A systematic performance evaluation of clustering methods for single-cell RNA-seq data, *F1000Res* 7 (2018) 1141.
- [29] S. Freytag, L. Tian, I. Lonnstedt, M. Ng, M. Bahlo, Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data, *F1000Res* 7 (2018) 1297.
- [30] D. Aran, A.P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R.P. Naikawadi, P. J. Wolters, A.R. Abate, et al., Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage, *Nat. Immunol.* 20 (2) (2019) 163–172.
- [31] L.P. Chung, S. Keshav, S. Gordon, Cloning the human lysozyme cDNA: inverted Alu repeat in the mRNA and in situ hybridization for macrophages and Paneth cells, *Proc. Natl. Acad. Sci. U. S. A.* 85 (17) (1988) 6227–6231.

- [32] J.H. Martens, H.G. Stunnenberg, BLUEPRINT: mapping human blood cell epigenomes, *Haematologica* 98 (10) (2013) 1487–1489.
- [33] E.P. Consortium, An integrated encyclopedia of DNA elements in the human genome, *Nature* 489 (7414) (2012) 57–74.
- [34] S. Aibar, C.B. Gonzalez-Blas, T. Moerman, V.A. Huynh-Thu, H. Imrichova, G. Hulselmanns, F. Rambow, J.C. Marine, P. Geurts, J. Aerts, et al., SCENIC: single-cell regulatory network inference and clustering, *Nat. Methods* 14 (11) (2017) 1083–1086.
- [35] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H.A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories, *Nat. Methods* 14 (10) (2017) 979–982.
- [36] G.X. Zheng, J.M. Terry, P. Belgrader, P. Ryvkin, Z.W. Bent, R. Wilson, S.B. Ziraldo, T.D. Wheeler, G.P. McDermott, J. Zhu, et al., Massively parallel digital transcriptional profiling of single cells, *Nat. Commun.* 8 (2017) 14049.
- [37] H.G. Stunnenberg, International Human Epigenome C, Hirst M: The International Human Epigenome Consortium, A Blueprint for scientific collaboration and discovery, *Cell* 167 (5) (2016) 1145–1149.
- [38] M.S. Jackson, M. Rocchi, G. Thompson, T. Hearn, M. Crosier, J. Guy, D. Kirk, L. Mulligan, A. Ricco, S. Piccinini, et al., Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications, and unstable sequences with homologies to telomeric and other centromeric locations, *Hum. Mol. Genet.* 8 (1999) 205–215.
- [39] P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network analysis, *BMC Bioinformatics* 9 (2008) 559.
- [40] G.P. Smith, Evolution of repeated DNA sequences by unequal crossover, *Science* 191 (1976) 528–535.
- [41] A.F. Smit, Interspersed repeats and other mementos of transposable elements in mammalian genomes, *Curr. Opin. Genet. Dev.* 9 (6) (1999) 657–663.
- [42] X. Pan, Y. Wang, Z. Li, X. Chen, R. Heller, N. Wang, C. Zhao, Y. Cai, H. Xu, S. Li, et al., Tracing the origin of a new organ by inferring the genetic basis of rumen evolution, *bioRxiv* (2020), <https://doi.org/10.1101/2020.02.19.955872>.
- [43] C. Schell, N. Wanner, T.B. Huber, Glomerular development—shaping the multicellular filtration unit, *Semin. Cell Dev. Biol.* 36 (2014) 39–49.
- [44] A. Smit, A. Riggs, MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation, *Nucleic Acids Res.* 23 (1995) 98–102.
- [45] R.G. LeBaron, K.I. Bezverkov, M.P. Zimmer, R. Pavelec, J. Skonier, A.F. Purchio, Beta IG-H3, a novel secretory protein inducible by transforming growth factor-beta, is present in normal skin and promotes the adhesion and spreading of dermal fibroblasts in vitro, *J. Invest. Dermatol.* 104 (5) (1995) 844–849.
- [46] J.J. Bond, A.J. Donaldson, J.V.F. Coumans, K. Austin, D. Ebert, D. Wheeler, V. H. Oddy, Protein profiles of enzymatically isolated rumen epithelium in sheep fed a fibrous diet, *J. Anim. Sci. Biotechnol.* 10 (2019) 5.
- [47] O. Tatti, P. Vehviläinen, K. Lehti, J. Keski-Oja, MT1-MMP releases latent TGF-beta1 from endothelial cell extracellular matrix via proteolytic processing of LTBP-1, *Exp. Cell Res.* 314 (13) (2008) 2501–2514.
- [48] M. Smidt, I. Kirsch, L. Ratner, Deletion of Alu sequences in the fifth c-sis intron in individuals with meningiomas, *J. Clin. Investig.* 86 (4) (1990) 1151–1157.
- [49] J. Kalucka, A. Ettinger, K. Franke, S. Mamlook, R.P. Singh, K. Farhat, A. Muschter, S. Olbrich, G. Breier, D.M. Katschinski, et al., Loss of epithelial hypoxia-inducible factor prolyl hydroxylase 2 accelerates skin wound healing in mice, *Mol. Cell. Biol.* 33 (17) (2013) 3426–3438.
- [50] E.E. Connor, R.L. VI Baldwin, M.P. Walker, S.E. Ellis, C. Li, S. Kahl, H. Chung, R. W. Li, Transcriptional regulators transforming growth factor-beta1 and estrogen-related receptor-alpha identified as putative mediators of calf rumen epithelial tissue development and function during weaning, *J. Dairy Sci.* 97 (7) (2014) 4193–4207.
- [51] N.O. Ku, P. Strnad, H. Bantel, M.B. Omary, Keratins: biomarkers and modulators of apoptotic and necrotic cell death in the liver, *Hepatology* 64 (3) (2016) 966–976.
- [52] J. Jiang, J.B. Cole, E. Freebern, Y. Da, P.M. VanRaden, L. Ma, Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls, *Commun. Biol.* 2 (1) (2019) 212.
- [53] C. Perez-Stable, C.K. Shen, Competitive and cooperative functioning of the anterior and posterior promoter elements of an Alu family repeat, *Mol. Cell. Biol.* 6 (6) (1986) 2041–2052.
- [54] C. Vascotto, D. Fantini, M. Romanello, L. Cesaratto, M. Deganuto, A. Leonardi, J. P. Radicella, M.R. Kelley, C. D'Ambrosio, A. Scalon, et al., APE1/Ref-1 interacts with NPM1 within nucleoli and plays a role in the rRNA quality control process, *Mol. Cell. Biol.* 29 (7) (2009) 1834–1854.
- [55] L. Perez-Jurado, R. Peoples, P. Kaplan, B. Hamel, U. Francke, Molecular definition of the chromosome 7 deletion in Williams syndrome and parent-of-origin effects on growth, *Am. J. Hum. Genet.* 59 (1996) 781–791.
- [56] L.M. Perelygina, N.V. Tomilin, O.I. Podgoraia, Nekotorye kharakteristiki belkov iz kletok HeLa, spetsificheski svyazyvayushchikh ALU-posledovatel'nost' cheloveka, *Mol. Biol.* 21 (6) (1987) 1610–1619.
- [57] A. Edwards, H.A. Hammond, L. Jin, C.T. Caskey, R. Chakraborty, Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups, *Genomics* 12 (1992) 241–253.
- [58] L. Edelmann, E. Spiteri, K. Koren, V. Pulijal, M.G. Bialer, A. Shanske, R. Goldberg, B.E. Morrow, AT-rich palindromes mediate the constitutional t(11;22) translocation, *Am. J. Hum. Genet.* 68 (1) (2001) 1–13.
- [59] L. Edelmann, P. Stankiewicz, E. Spiteri, R.K. Pandita, L. Shaffer, J.R. Lupski, B. E. Morrow, Two functional copies of the DGCR6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus, *Genome Res.* 11 (2) (2001) 208–217.
- [60] L. Edelmann, R.K. Pandita, B.E. Morrow, Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome, *Am. J. Hum. Genet.* 64 (4) (1999) 1076–1086.
- [61] P.M. Bingham, T.B. Chou, I. Mims, Z. Zacher, On/off regulation of gene expression at the level of splicing, *Trends Genet.* 4 (1988) 134–138.
- [62] F. Bigoni, R. Stanyon, U. Koehler, A. Morescalchi, J. Wienberg, Mapping homology between human and black and white colobine monkey chromosomes by fluorescent in situ hybridization, *Am. J. Primatol.* 42 (1997) 289–298.
- [63] B.D. Rosen, D.M. Bickhart, R.D. Schnabel, S. Koren, C.G. Elsik, E. Tseng, T. N. Rowan, W.Y. Low, A. Zimin, C. Couldrey, et al., De novo assembly of the cattle reference genome with single-molecule sequencing, *Gigascience* (2020) 9(3).
- [64] M.A. Tosches, T.M. Yamawaki, R.K. Naumann, A.A. Jacobi, G. Tushev, G. Laurent, Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles, *Science* 360 (6391) (2018) 881.
- [65] G. Yu, L.G. Wang, Y. Han, Q.Y. He, ClusterProfiler: an R package for comparing biological themes among gene clusters, *OMICS* 16 (5) (2012) 284–287.
- [66] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W. H. Fridman, F. Pages, Z. Trajanoski, J. Galon, ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics* 25 (8) (2009) 1091–1093.
- [67] B. Li, L. Fang, D.J. Null, J.L. Hutchison, E.E. Connor, P.M. VanRaden, M. J. VandeHaar, R.J. Tempelman, K.A. Weigel, J.B. Cole, High-density genome-wide association study for residual feed intake in Holstein dairy cattle, *J. Dairy Sci.* 102 (12) (2019) 11067–11080.
- [68] L. Fang, W. Cai, S. Liu, O. Canela-Xandri, Y. Gao, J. Jiang, K. Rawlik, B. Li, S. G. Schroeder, B.D. Rosen, et al., Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle, *Genome Res.* 30 (5) (2020) 790–801.
- [69] E. Freebern, D.J.A. Santos, L. Fang, J. Jiang, K.L. Parker Gaddis, G.E. Liu, P. M. VanRaden, C. Maltecca, J.B. Cole, L. Ma, GWAS and fine-mapping of livability and six disease traits in Holstein cattle, *BMC Genomics* 21 (1) (2020) 41.
- [70] P.D. Rohde, I. Fourie Sørensen, P. Sørensen, qgg: an R package for large-scale quantitative genetic analyses, *Bioinformatics* 36 (8) (2019) 2614–2615.
- [71] S. Liu, Y. Yu, S. Zhang, J.B. Cole, A. Tenesa, T. Wang, T.G. McDanel, L. Ma, G. E. Liu, L. Fang, Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of complex traits in cattle and human, *BMC Biol.* 18 (1) (2020) 80.
- [72] P.D. Rohde, D. Demontis, B.C.D. Cuyabano, A.D. Børglum, P. Sørensen, Covariance association test (CVAT) identifies genetic markers associated with schizophrenia in functionally associated biological processes, *Genetics* 203 (4) (2016) 1901–1913.
- [73] I.F. Sørensen, S.M. Edwards, P.D. Rohde, P. Sørensen, Multiple trait covariance association test identifies gene ontology categories associated with chill coma recovery time in *Drosophila melanogaster*, *Sci. Rep.* 7 (1) (2017) 2413.
- [74] P. Sarup, J. Jensen, T. Ostensen, M. Henryron, P. Sørensen, Increased prediction accuracy using a genomic feature model including prior information on quantitative trait locus regions in purebred Danish Duroc pigs, *BMC Genet.* 17 (1) (2016) 11.
- [75] L. Fang, G. Sahana, G. Su, Y. Yu, S. Zhang, M.S. Lund, P. Sørensen, Integrating sequence-based GWAS and RNA-Seq provides novel insights into the genetic basis of mastitis and milk production in dairy cattle, *Sci. Rep.* 7 (1) (2017) 45560.
- [76] L. Fang, G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M.S. Lund, P. Sørensen, Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection, *Genet. Sel. Evol.* 49 (1) (2017) 44.

RESEARCH ARTICLE

Open Access



Genome-wide association study of *Mycobacterium avium* subspecies *Paratuberculosis* infection in Chinese Holstein

Yahui Gao¹, Jianping Jiang¹, Shaohua Yang¹, Jie Cao², Bo Han¹, Yachun Wang¹, Yi Zhang¹, Ying Yu¹, Shengli Zhang¹, Qin Zhang¹, Lingzhao Fang³, Bonnie Cantrell⁴ and Dongxiao Sun^{1*}

Abstract

Background: Paratuberculosis is a contagious, chronic and enteric disease in ruminants, which is caused by *Mycobacterium avium* subspecies *paratuberculosis* (MAP) infection, resulting in enormous economic losses worldwide. There is currently no effective cure for MAP infection or a vaccine, it is thus important to explore the genetic variants that contribute to host susceptibility to infection by MAP, which may provide a better understanding of the mechanisms of paratuberculosis and benefit animal genetic improvement. Herein we performed a genome-wide association study (GWAS) to identify genomic regions and candidate genes associated with susceptibility to MAP infection in dairy cattle.

Results: Using Illumina Bovine 50 K (54,609 SNPs) and GeneSeek HD (138,893 SNPs) chips, two analytical approaches were performed, GRAMMAR-GC and ROADTRIPS in 937 Chinese Holstein cows, among which individuals genotyped by the 50 K chip were imputed to HD SNPs with Beagle software. Consequently, 15 and 11 significant SNPs ($P < 5 \times 10^{-5}$) were identified with GRAMMAR-GC and ROADTRIPS, respectively. A total of 10 functional genes were in proximity to (i.e., within 1 Mb) these SNPs, including *IL4*, *IL5*, *IL13*, *IRF1*, *MyD88*, *PACSIN1*, *DEF6*, *TDP2*, *ZAP70* and *CSF2*. Functional enrichment analysis showed that these genes were involved in immune related pathways, such as interleukin, T cell receptor signaling pathways and inflammatory bowel disease (IBD), implying their potential associations with susceptibility to MAP infection. In addition, by examining the publicly available cattle QTLdb, a previous QTL for MAP was found to be overlapped with one of regions detected currently at 32.5 Mb on BTA23, where the *TDP2* gene was anchored.

Conclusions: In conclusion, we identified 26 SNPs located on 15 chromosomes in the Chinese Holstein population using two GWAS strategies with high density SNPs. Integrated analysis of GWAS, biological functions and the reported QTL information helps to detect positional candidate genes and the identification of regions associated with susceptibility to MAP traits in dairy cattle.

Keywords: GWAS, Paratuberculosis, SNP, Chinese Holstein

* Correspondence: sundx@cau.edu.cn

¹Key Laboratory of Animal Genetics and Breeding of Ministry of Agriculture, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China
Full list of author information is available at the end of the article



© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Paratuberculosis, also known as Johne's disease (JD), is a contagious, chronic and enteric disease in ruminants caused by *Mycobacterium avium* subspecies *paratuberculosis* (MAP) [1]. Symptoms of the disease include diarrhea and weight loss that eventually leads to death. This disease has a long period of incubation [2]. In cattle, JD cannot be diagnosed until symptoms are observed because animals can have MAP in their systems, but not have JD. It is always difficult to determine if an animal is at risk of contracting JD by screening for MAP. Once a cow shows symptoms of JD, there is no treatment so the only effective means to get rid of the disease is to cull. There is also no vaccine, so farmers cannot increase resistance to JD within their herds, which causes huge economic losses. The disease can easily be spread through the farm from contact with MAP infected feces or milk from infected cows. Changes in herd management could help to reduce JD, but understanding resistance in cattle will allow for better management of the disease. Genomic selection for disease-resistant animals may be a promising way to increase the ability of animals to resist MAP infection. Exploring the genetic variants that contribute to host susceptibility to infection by MAP is important both for animal genetic improvement programs and for a better understanding of the underlying mechanisms of disease.

Heritability estimates in Holstein and Jersey cows for infection with MAP range from 0.031 to 0.283 [3–13]. By employing a case-control design, several functional genes have been reported to be associated with susceptibility to MAP infection in cattle. This includes *CLEC7A* [14], *IL10RA* [15], *IL12RB1*, *IL12RB2*, *IL23R*, *IFNGR2* [16], *NOD2* [17–19], *PGLYRP1* [20], *SLC11A1* [21, 22], *SP110* [23], *TLR1* [24, 25], *TLR2* [24–26] and *TLR4* [24, 25]. Multiple QTLs are located on BTA7 [27] and BTA20 [28]. These studies indicated that genetic factors contribute to the susceptibility of MAP infection in cattle.

Nowadays, genome-wide association study (GWAS) is a popular strategy to identify candidate genes for specified traits. Earlier GWAS studies for MAP infection in Holstein cattle were based on serum ELISA, milk ELISA, fecal culture test or a comprehensive test for MAP infection [29–37]. Various SNPs associated with susceptibility to MAP infection are extensively distributed across all autosomes in different Holstein cattle populations. Several candidate genes for MAP infection were subsequently identified, such as *EDN2*, *PRDM1*, *LAMB4*, *DLD*, *LDLRAD3*, *CACNA1B*, *TIMD4*, *ITK*, *C*, *BTN1A1* and *TDP2* [29–37]. Zare et al. reported 9 SNPs on BTA3, BTA6, BTA17 and BTA23 in the US Jersey cattle population and suggested *SLC17A1*, *UBD*, *HIVEP1*, *CCDC17*, *ZNF684*, *UBE2L3*, *UBE2K*, *FAM109A* and *FAM5C* genes as candidates for susceptibility to MAP

infection [35]. In the present study, the objectives were to identify genetic markers and genomic regions that are associated with susceptibility to MAP infection by performing a GWAS in Chinese Holsteins, and to provide further molecular information for the MAP resistance breeding program.

Methods

Data description

A total of 8214 Chinese Holstein cows from 7 dairy farms belonging to the Beijing Sanyuan Dairy Farm Center were fed under the same management throughout this study. We collected a 500 µL blood sample from the caudal vein of each cow and performed a regular quarantine inspection of the farms during September 2014. Serum extracted from blood samples were stored at 4 °C until testing; within 5 days after collection. The commercially available, ELISA kit (IDEXX Laboratories, Inc., Westbrook, ME, USA) was used according to the manufacturer instructions to measure the antibody levels of each serum sample. The MAP status of an animal was expressed as a percentage of the sample to positive ratio (S/P) with the formula: $S/P \text{ ratio} = [(optical \text{ density (OD) of the sample} - OD \text{ of the negative control}) / (OD \text{ of positive sample} - OD \text{ of the negative control})]$, where ≤ 0.45 is negative; $0.45 < S/P < 0.55$ is suspect; $S/P \geq 0.55$ is positive. ELISA suspect results were excluded because of their uncertainty. Out of the 8214 detected cows, 185 positive individuals (case) and 760 negative individuals (control) from 6 herds were used for GWAS. ELISA results were employed as a binary trait (0 = negative, 1 = positive).

Genotyping

The individuals in this study were divided into 2 sub-groups for genotyping. Five hundred and thirty three cows belonging to the first sub-group were genotyped with the Illumina Bovine SNP50 BeadChip (54,609 SNPs, Illumina, San Diego, CA, USA) after extracting DNA from whole blood using routine procedures. DNA was isolated from whole blood with a commercially available kit, the DP318 Blood DNA Kit (Tiangen Biotech Co., China). The DNA of the remaining 412 cows in the second sub-group were extracted from hair by GeneSeek with QIAamp® DNA Mini Kit (QIAGEN Inc., Valencia, CA, USA) and then genotyped with the GeneSeek Genomic Profiler HD v2 (138,893 SNPs, GeneSeek, Lincoln, NE, USA). The genotype data were deposited in the Additional files (Additional files 1 and 2).

Imputation and quality control

To make full use of SNPs originating from the GeneSeek Genomic Profiler HD v2 (GeneSeek), individuals genotyped by the Illumina Bovine SNP50 BeadChip were

imputed to GeneSeek Genomic Profiler HD v2. Imputation was performed using BEAGLE 3.3.2 [38] default options. Allelic R^2 was estimated as an indicator of imputation accuracy based on the genotype probabilities.

To further evaluate imputation accuracy, we randomly selected 100 cows genotyped by the high-density chip, obtained the common SNPs between the two panels, and masked genotypes of SNPs left. We classified this small subset of cows as the study population, while the remaining cows genotyped by the high-density panel were classified as the reference population. After imputation, we compared the imputed and masked actual genotypes of the selected 100 cows to calculate the percentage of genotypes that are consistent between them.

Then we implemented PLINK [39] and removed SNPs with call rates < 95%, minor allele frequencies < 0.01, a deviation from Hardy-Weinberg equilibrium (HWE) P values < 10^{-6} and > 5% missing genotypes. A dataset containing 109,607 SNPs and 937 animals (182 cases and 755 controls) was used for further analysis. All SNP positions were determined according to the *Bos taurus* UMD 3.1 assembly [40].

Population stratification

Differences in allele frequencies between subpopulations of admixed populations can lead to false association in a GWAS [41]. In order to determine whether stratification exists in our study population, a principle component analysis (PCA) was performed by GCTA 1.24 [42] and results were visualized by R 3.3.1 [43].

GWAS

GRAMMAR-GC

We performed a GWAS using the Genome-wide Rapid Association using Mixed Model and Regression-Genomic Control (GRAMMAR-GC) approach [44, 45], a single-marker method implemented within the GenABEL package [46] for R [43]. This approach can account for the potential population structure and infer relationships using SNP data without pedigree information. There have been multiple GWAS studies in cattle that utilized this approach [31, 34–37]. GRAMMAR-GC is comprised of three steps that use the regression of phenotypes on the genotypes of individuals for one SNP at a time. First, to account for familial dependence among individuals, phenotypes were corrected by conducting a polygenic analysis using a genomic kinship matrix based on the SNP genotypes. Residuals from the polygenic analysis were then used as dependent quantitative traits for association analysis of each SNP with a linear regression model. Finally, genomic control (GC) was applied to correct the test statistic using the genomic inflation factor (λ), which is the regression

coefficient of the observed statistic on the expected statistic. We performed an association test for each SNP based on the following linear mixed model:

$$y = W\alpha + x\beta + u + \epsilon,$$

where y is the liability vector for case/control observations; W is a matrix of covariates (fixed effects that contain herd and parity); α is a vector of the corresponding coefficients including the intercept; x is a vector of genotypes of a marker at the locus tested; β is the effect size of the marker; u is a vector of random polygenic effects with a covariance structure as $\beta \sim N(0, V_g)$, V_g is the polygenic additive variance; ϵ is a vector of residual errors with $\epsilon \sim N(0, IV_e)$, I is the identity matrix, and V_e is the residual variance component.

In general, for GRAMMAR-GC, the value $T^2/\hat{\zeta}$ of each SNP with one-degree freedom is compared with χ^2_1 to determine whether the locus is significantly associated with the trait. Here $T_k^2 = \hat{\beta}_k^2 / \text{var}(\hat{\beta}_k)$, where $\hat{\beta}_k$ is the effect of the k^{th} SNP. The deflation factor ζ is estimated as $\zeta = \text{median}(T_1^2, T_2^2, \dots, T_k^2) / 0.456$.

ROADTRIPS

A second GWAS approach was implemented with ROADTRIPS 2.0 [47]. An important advantage of ROADTRIPS 2.0 is that it can analyze data with pedigree information and population admixture simultaneously. Based on the genome-wide SNP data, an empirical covariance matrix was constructed to adjust for potential population admixture and relatedness among individuals and maintain the advantage of utilizing known pedigree information when available. The ROADTRIPS 2.0 test statistic based on χ^2_1 distribution for each SNP takes the form:

$$\frac{(V^T Y)^2}{\hat{\sigma}^2 V^T \hat{\Psi} V} \sim \chi^2_1$$

Here $Y = (Y_1, Y_2, \dots, Y_n)^T$, is genotype vector at a test SNP for n individuals (coded using an allelic coding). V is a vector of length n coding for phenotype information (disease status) and known relationships. $\hat{\sigma}^2 \hat{\Psi}$ is an estimate of the null variance/covariance matrix of Y . $\hat{\sigma}^2$ is an estimate of $\text{Var}(Y)$ in an outbred population and $\hat{\Psi}$ is an estimated matrix used to simultaneously adjust for unknown relatedness/pedigree relationship errors and population stratification.

ROADTRIPS 2.0 provides three association tests named RM test, R_x test and RW test. According to the authors' recommendation, the RM test is the most powerful among the three tests when pedigree information is available. Compared with the R_x test and the RW test, the RM test can use the phenotypic information of

individuals with missing genotypes provided that they have a genotyped relative at the tested marker. Considering the features of the RM test and the data structure of this study being based on a corrected pedigree, we adopted the RM test for association analysis. *P* values of SNPs were derived from an asymptotic chi-square distribution with 1 degree of freedom. In addition, the fixed effects used here was the same as above.

Following the suggestion of the Wellcome Trust Case Control Consortium [48], two *P* value thresholds of 5×10^{-7} and 5×10^{-5} were considered as genome-wide “strong” and “moderate” association respectively.

Gene contents and functional annotation

Using BioMart in the Ensembl database (Ensembl Genes 92), genes within 1 Mb of the significant SNPs were retrieved based on the UMD 3.1 assembly. To provide insight into the functional enrichment of genes identified, we carried out GO (Gene Ontology) and Pathway analysis using KOBAS 3.0 [49]. KOBAS annotates a set of genes with putative pathways and disease relationships by mapping to genes with a known annotation. In addition, we compared the regions within 1 Mb of the significant SNPs with the reported cattle to QTLs for JD tolerance and *MAP* susceptibility in the Animal QTL database (<http://www.animalgenome.org/cgi-bin/QTLdb/index>) [50].

Results

Imputation accuracy

After imputation, we discarded SNPs with allelic $R^2 < 0.85$ and found an average allelic R^2 of 96.7% for imputed genotypes. Then we took a small subset including 100 cows genotyped by the high-density chip for calculating the imputation accuracy. Finally, the percentage of consistent genotypes was 97.03%, which suggested a high accuracy of imputation.

GWAS based on GRAMMAR-GC

With GCTA 1.24, a slight population substructure was revealed (Additional file 3: Figure S1). The inflation factor (λ), estimated to be 0.9399 (SE = 0.0002), indicates population substructure was a minor issue and that our results can be accepted for further analysis. The GC-corrected *P* values for the majority of SNPs corresponded well to the expected *P* values under the null hypothesis of no association. However, a few departures which mean the *P* values of these SNPs were higher than the expected *P* values under the null hypothesis indicated associations with the trait being studied (Additional file 4: Figure S2). As shown in Tables 1, 2 SNPs passed the strong association threshold and 13 SNPs passed the moderate threshold (Fig. 1).

Table 1 Results of GRAMMAR-GC genome-wide association analysis for susceptibility to *MAP* infection (df = 1)

SNP	BTA	Position	<i>P</i> value
BovineHD2200003382	22	11,359,993	1.13E-14
BovineHD0700017491	7	60,947,866	2.76E-09
BovineHD0700007000	7	25,403,106	2.54E-06
BovineHD2300007824	23	28,173,531	6.90E-06
BTB-00030699	1	61,806,466	8.85E-06
BovineHD0800002130	8	6,659,432	1.18E-05
BovineHD0200036516	2	125,910,848	1.49E-05
Hapmap44402-BTA-73818	13	27,398,154	1.64E-05
BovineHD4100015938	23	8,975,441	1.98E-05
ARS-BFGL-BAC-29490	23	20,166,517	2.75E-05
BovineHD2300009447	23	32,516,000	3.82E-05
ARS-BFGL-NGS-36626	18	54,052,117	4.50E-05
BovineHD0700006647	7	24,259,310	4.62E-05
BovineHD1800010086	18	33,293,455	4.62E-05
BovineHD2700011748	27	40,498,309	4.83E-05

GWAS based on ROADTRIPS

RM test was implemented for the association analysis. As shown in Additional file 5: Figure S3, the *P* values for the majority of SNPs exhibited a good correspondence to the expected values with a limited number of SNPs indicating their associations with the studied trait. In total, 4 and 7 SNPs passed the threshold of strong and moderate association, respectively (Fig. 2, Table 2).

Gene contents and functional annotation

There were 15 significant SNPs detected by GRAMMAR-GC in total. Utilizing BioMart in the Ensembl database (Ensembl Genes 92), we obtained the 232 IDs for genes located within or overlapped with the regions nearby these SNPs (< 1 Mb) (Additional file 6:

Table 2 Results of ROADTRIPS genome-wide association analysis for susceptibility to *MAP* infection (df = 1)

SNP	BTA	Position	<i>P</i> value
BTB-01281916	3	71,760,025	1.21E-08
Hapmap49590-BTA-38619	16	35,765,411	4.37E-08
ARS-BFGL-NGS-25380	11	2,260,123	1.98E-07
BTB-00452217	11	2,983,521	2.81E-07
BTA-79476-no-rs	7	60,933,059	9.71E-07
BovineHD0700006447	7	23,495,415	2.10E-06
BovineHD0200035709	2	123,174,023	7.90E-06
ARS-USDA-AGIL-chr6-117,920,790-000733	6	117,920,790	1.87E-05
BovineHD2400014963	24	52,857,741	3.33E-05
Hapmap39753-BTA-50189	20	28,186,842	4.80E-05
BovineHD1300015832	13	55,734,067	4.92E-05

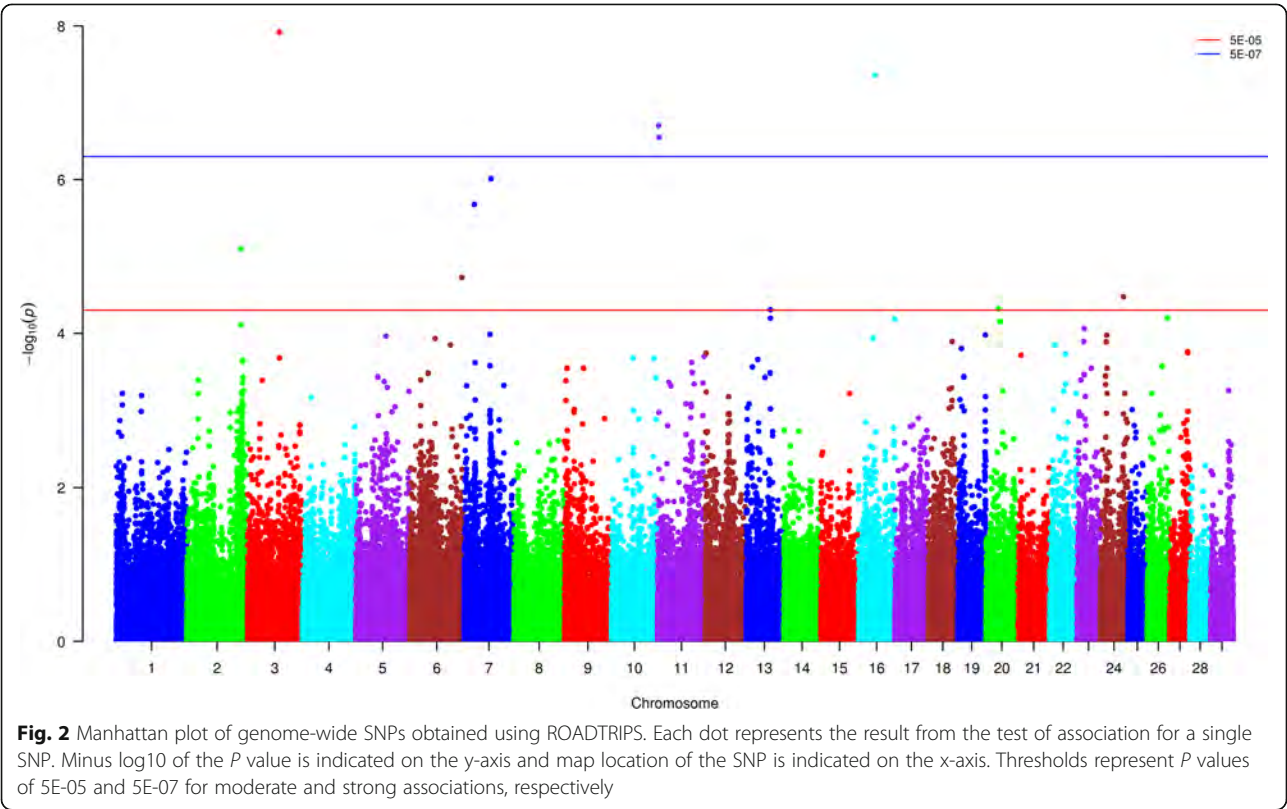
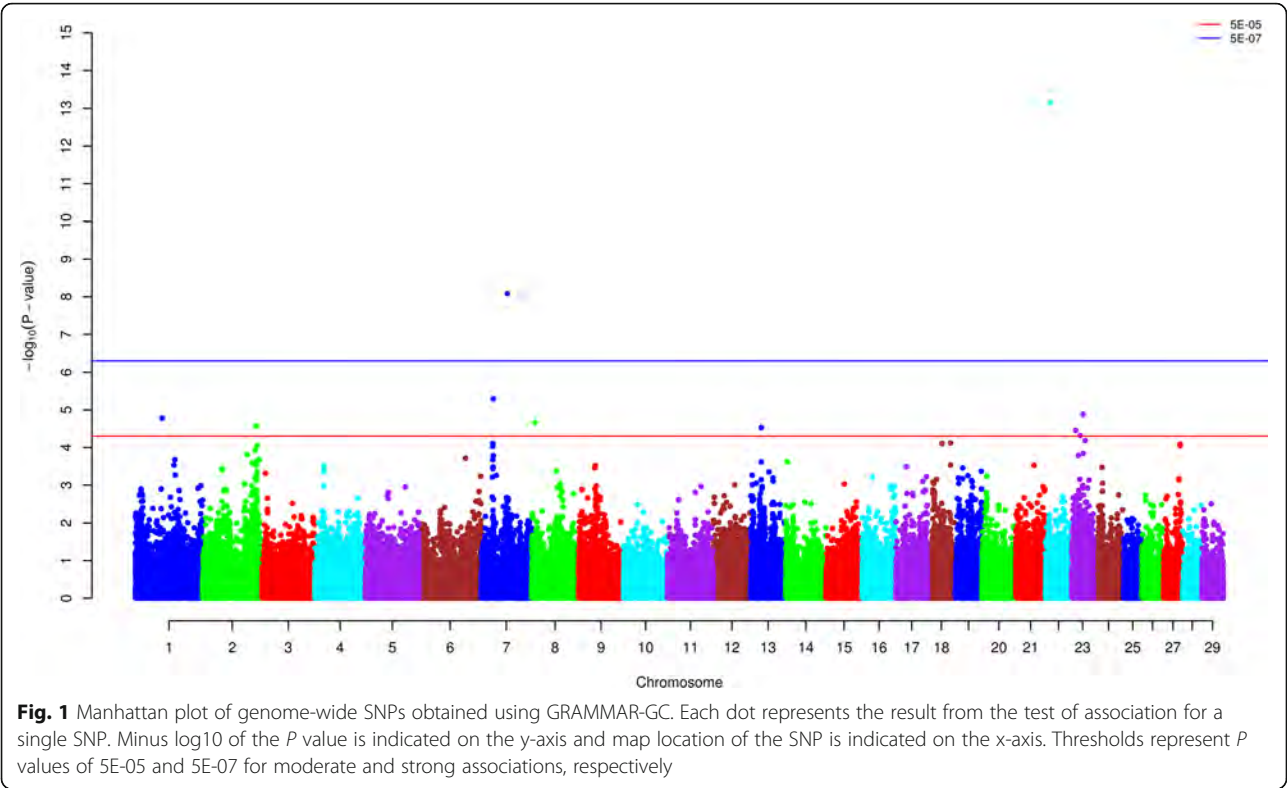


Table S1). Based on the 11 significant SNPs detected by ROADTRIPS, 123 functional gene IDs were identified (Additional file 7: Table S2). After the combination of results, a total of 343 genes were obtained, including 283 protein-coding genes, 21 miRNA genes, 6 pseudogenes, 12 snRNA, 15 snoRNA, 4 rRNA and 2 miscRNA (Additional file 8: Table S3).

GO and Pathway analysis were performed by KOBAS 3.0 to determine the biological functions of the 343 genes. Finally, 348 significant GO terms were detected, including those related to immune response ($P < 0.05$), such as immune response-regulating signaling pathway, regulation of leukocyte proliferation and immune response-activating signal transduction. Fifteen significant pathways were found, including those related to immune responses ($P < 0.05$) such as autoimmune thyroid disease and enrichment of the interleukin signaling pathway (Additional file 9: Table S4). In addition, T cell receptor signaling pathway [51] and inflammatory bowel disease (IBD) were detected but not significant.

Quantitative traits locus overlapped with SNPs

Until now, 6 JD tolerance and 161 *MAP* susceptibility QTLs have been reported in the cattle QTL database (<http://www.animalgenome.org/cgi-bin/QTLdb/BT/index>). After comparing these QTLs with the regions within 1 Mb of the 26 significant SNPs, 2 QTLs identified before [30, 37] located in BTA23 (~ 32.5 Mb) for *MAP* susceptibility were found. This implies the functional genes, such as *TDP2* (tyrosyl-DNA phosphodiesterase 2) around these SNPs are likely candidates for *MAP* susceptibility traits.

Discussion

There have been multiple GWASs conducted in different cattle population [29–37], and some candidate loci and genes have been identified. Although these studies found evidence of genomic regions associated with *MAP* infection, the consistency was not high. The genomic regions and genes regarded as candidates for the target traits were variable among previous studies. The difference between genomic regions identified by different studies is because of different trait definitions except for statistical methodologies [52]. There were four main definitions for infection cases in previous studies: ELISA positive [27, 31, 33], fecal culture positive [29], tissue culture positive [29] and comprehensive testing. Comprehensive testing includes tissue culture positive and fecal culture positive [29], ELISA positive or fecal positive [30], and ELISA positive or tissue culture positive [34].

We found 15 SNPs passing the threshold (5×10^{-5}) using GRAMMAR-GC with BTA23 owning the most SNPs. The most significant SNP, BovineHD2200003382

($P = 1.13\text{E-}14$) was identified on BTA22 (Table 1). Similarly, 11 SNPs passed the threshold (5×10^{-5}) using ROADTRIPS with the most significant SNP, BTB-01281916 ($P = 1.21\text{E-}08$) identified on BTA3 (Table 2). There was no common SNP sharing between these two methods because of some possible reasons. Firstly, different computational principles can cause different significant SNPs. It is common that the discrepancy caused by this reason, such as the report of Alpay et al. [36] and Sallam et al. [37]. Both ROADTRIPS and GRAMMAR-GC can correct sample structure. The ROADTRIPS program uses the quasi-likelihood methods (implemented in the MQLS and similar statistics), to obtain known kinship coefficients, which then together with the empirical covariance matrix estimated from genomic data to correct for known and unknown relatedness and population structure [47]. Instead of pedigree, GRAMMAR-GC program uses genomic kinship matrix estimated through genomic marker data to adjust for average allele sharing or relatedness among sample individuals and thus remove genetic stratification [44, 45].

Secondly, JD is affected by multiple genetic loci and SNPs identified using different methods were polygenic in present study. Those SNPs with large effect may be captured more easily by multiple GWAS methods. In addition, the size of study population may cause discrepancy. The more individuals, the higher the accuracy of GWAS result. So it seems normal that different methods identifying different SNPs in present study. Thus, it seems normal that different methods identify different SNPs.

While no SNPs were identified by two methods, but the SNPs between the two methods were located close to each other on the same chromosome. For example on BTA2 ~ 2.74 Mb was found between BovineHD0200036516 (GRAMMAR-GC) and BovineHD0200035709 (ROADTRIPS), on BTA7 ~ 14.81 Kb was found between BovineHD0700017491 (GRAMMAR-GC) and BTA-79476-no-rs (ROADTRIPS), on BTA7 ~ 1.91 Mb was found between BovineHD0700007000 (GRAMMAR-GC) and BovineHD0700006447 (ROADTRIPS) and ~ 0.76 Mb was found between BovineHD0700006647 (GRAMMAR-GC) and BovineHD0700006447 (ROADTRIPS). Combining the results of these two methods, 26 significant SNPs were obtained. The most SNPs were found on BTA7 followed by BTA23.

Among the 26 significant SNPs detected by two methods, BovineHD0700006447 (23.5 Mb) located on BTA7 in this study was close to SNPs detected by Pant et al. (20.6 Mb ~ 22.3 Mb) [27]. Genes nearby this SNP within less than 1 Mb were *IL4*, *IL5*, *IL13* and *IRF1*. The genes, *IL4* (interleukin 4), *IL5* and *IL13* are type 2 cytokines, that may be the predominant cytokines produced by CD4+ and other T cells in lymph nodes during the

subclinical infection of *MAP* [53]. As previously reported [54–56], in the clinical infection, the bovine *MAP* infection disease was characterized by a gradual shift in the immune responses from cell-mediated immune response to antibody mediated immune response while *IL4*, *IL5* and *IL13* can promote the T_H2 antibody-mediated immune response. Therefore, these three genes might play important roles in the pathogenesis of the disease [27]. *IRF1*, interferon regulatory factor 1, plays an important role in many immune responses including the Type 1 (T_H1) cell-mediated immune response. Cell mediated immunity is an important host defense mechanism against intracellular pathogens including *MAP* [57]. In addition, it can regulate the expression of many immune genes such as *IL6*, *IL12B*, and inducible nitric oxide synthase (*NOS2*) that function in the pathogenesis of human IBD [58–60].

The most significant SNP, BovineHD2200003382 detected by GRAMMAR-GC was located at 11.3 Mb on BTA22. Genes within 1 Mb of this location includes *MyD88* (myeloid differentiation primary response gene 88) which encodes a cytosolic adapter protein that plays a central role in the innate and adaptive immune response. *MyD88* functions as an essential signal transducer in the interleukin-1 and Toll-like receptor signaling pathways [61].

SNP BovineHD4100015938 on BTA23 (8.98 Mb) was close to ARS-BFGL-NGS-109956 (7.84 Mb) and ARS-BFGL-NGS-115177 (7.87 Mb) reported by Zare et al. [35]. The genes near to these two SNPs included *PACSIN1* and *DEF6* that are related to immune response. *PACSIN1* (protein kinase C and casein kinase substrate in neurons 1), belonging to a family of cytoplasmic phosphoproteins, participates in the regulation of endocytosis [62] and regulates the TLR7/9-mediated type I interferon response in plasmacytoid dendritic cells [63]. *DEF6* (*DEF6*, guanine nucleotide exchange factor) is a guanine nucleotide exchange factor (GEF) for RAC (MIM 602048) and CDC42 (MIM 116952) that are highly expressed in B and T cells [64]. SNP BovineHD2300009447 (32.5 Mb) located on BTA23 was very close to ARS-BFGL-NGS-1938 (32.6 Mb) reported by Zare et al. in Jersey cattle [35], and close to ss105264543 (33.6 Mb) reported by Minozzi et al. in Holstein cattle [34]. Gene within 1 Mb of this region was *TDP2* (tyrosyl-DNA phosphodiesterase 2). This gene encodes a member of a superfamily of divalent cation-dependent phosphodiesterases. The encoded protein associates with CD40, tumor necrosis factor (TNF) receptor-75 and TNF receptor associated factors (TRAFs) that inhibits nuclear factor-kappa-B activation. In addition, *TDP2* has sequence and structural similarities with APE1 endonuclease, which is involved in both DNA repair and the activation of transcription factors [65].

Interleukin, T cell receptor signaling pathway and inflammatory bowel disease (IBD) pathways are related to immune or inflammatory response. Two genes including *ZAP70* and *SRF*, except for *IL4*, *IL5*, *IL13* and *MyD88* stated above, were involved in the immune biological processes. *ZAP70* (zeta chain of T-cell receptor associated protein kinase 70) encodes an enzyme belonging to the protein tyrosine kinase family that plays a role in T-cell development and lymphocyte activation. This enzyme, phosphorylated on tyrosine residues upon T-cell antigen receptor (TCR) stimulation, functions in the initial step of TCR-mediated signal transduction in combination with the Src family kinases, Lck and Fyn and plays an essential role in the process of thymocyte development [66]. In addition, mutations in this gene cause selective T-cell defect, a severe combined immunodeficiency disease characterized by a selective absence of CD8-positive T-cells [67]. Leite et al. investigated the expression of *ZAP70* in cows naturally infected with *MAP* and revealed that the surface expression of *ZAP70* was decreased in CD4+ T cells of both subclinical and clinical animals indicating a change in T cell phenotype with disease state [68]. *CSF2* (colony stimulating factor 2), also known as *CSF* and *GM-CSF*, encodes a cytokine that controls the production, differentiation, and function of granulocytes and macrophages. This gene has been localized to a cluster of related genes at chromosome region 5q31 that are known to be associated with interstitial deletions in the 5q- syndrome and acute myelogenous leukemia. Other genes in the cluster include those encoding interleukins 4, 5, and 13 [69].

Furthermore, we found a region nearby the BovineHD2300009447 (BTA23, 32.5 Mb) overlapped with one QTL associated with *MAP* susceptibility. Combining this information with related genes found above, the 32 ~ 33 Mb region of BTA23 may be a case of a genomic region associated with *MAP* infection, which corresponds to the report of Zare et al. [35].

The present study focused on the potential function of 10 candidate genes. Future analysis is necessary to investigate the biological processes and molecular mechanism of these genes to anchor immune alterations and possible triggers that result in clinical paratuberculosis.

Conclusions

We performed a case-control GWAS for *MAP* infection in Chinese Holstein cattle using two statistical approaches, GRAMMAR-GC and ROADTRIPS. Twenty-six significant SNPs located on 15 chromosomes were detected based on data after imputation. Ten genes within less than 1 Mb of these SNPs were

involved in immune response pathways, implying their potential associations with susceptibility to *MAP*. These genes included *IL4*, *IL5*, *IL13*, *IRF1*, *MyD88*, *PACSIN1*, *DEF6*, *TDP2*, *ZAP70* and *CSF2*. By examining the QTLdb, the 32 ~ 33 Mb region of BTA23 may be a genomic region associated with *MAP* infection.

Additional files

- Additional file 1:** MAP file of SNP data. (MAP 3577 kb)
- Additional file 2:** PED file of SNP data. (PARTIAL 776 kb)
- Additional file 3:** Figure S1. PCA plot based on SNP data. (TIFF 175 kb)
- Additional file 4:** Figure S2. Q-Q plot based on SNP data using GRAMMAR-GC. (TIFF 175 kb)
- Additional file 5:** Figure S3. Q-Q plot based on SNP data using ROAD-TRIPS. (TIFF 179 kb)
- Additional file 6:** Table S1. The features of genes based on GRAMMAR-GC. (XLSX 20 kb)
- Additional file 7:** Table S2. The features of genes based on RODATRIPS. (XLSX 15 kb)
- Additional file 8:** Table S3. The features of genes based on GRAMMAR-GC and RODATRIPS. (XLSX 26 kb)
- Additional file 9:** Table S4. Functional enrichment of GO and Pathway analysis of 343 genes. (XLSX 32 kb)

Abbreviations

GC: Genomic control; GRAMMAR-GC: Genome-wide Rapid Association using Mixed Model and Regression-Genomic Control; GWAS: Genome-wide association study; JD: Johne's disease; MAP: *Mycobacterium avium* subspecies *Paratuberculosis*; PCA: Principle component analysis; QQ: Quantile-quantile; QTL: Quantitative trait locus; SNP: Single nucleotide polymorphisms

Acknowledgements

Not applicable.

Funding

This work was supported financially by the National Natural Science Foundation (31472065, 31872330), Beijing Dairy Industry Innovation Team (BAIC06-2017/2018), Beijing Science and Technology Program (D171100002417001), earmarked fund for Modern Agro-industry Technology Research System (CARS-36), and the Program for Changjiang Scholar and Innovation Research Team in University (IRT_15R62IRT_15R62). Funding bodies did not have a role in the design of the study and collection, analysis, and interpretation of data, neither in writing the manuscript.

Availability of data and materials

All supporting data can be found within the additional files.

Authors' contributions

YG performed bioinformatics and statistical analysis, and also was a major contributor to manuscript preparation. JJ, SY, JC and BH performed experiments and sample collection. YW, YZ, YY, SZ and QZ participated in result interpretation, wrote, revised and approved the manuscript. LF and BC commented the manuscript and were major contribution to manuscript revision. DS conceived and designed the experiments and wrote the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

All protocols for collection of the blood and hair samples of China Holstein cows were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) at China Agricultural University. Blood and hair samples were collected specifically for this study following standard procedures with the full agreement of the Beijing Dairy Cattle Center who owned the animals.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Key Laboratory of Animal Genetics and Breeding of Ministry of Agriculture, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing 100193, China. ²College of Veterinary Medicine, China Agricultural University, Beijing 100193, China. ³Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA. ⁴Department of Animal and Veterinary Sciences, University of Vermont, Burlington, VT 05405, USA.

Received: 1 February 2018 Accepted: 18 December 2018

Published online: 27 December 2018

References

- Clarke CJ. The pathology and pathogenesis of paratuberculosis in ruminants and other species. *J Comp Pathol*. 1997;116(3):217–61.
- Whitlock RH, Buerge C. Preclinical and clinical manifestations of paratuberculosis (including pathology). *Vet Clin North Am Food Anim Pract*. 1996;12(2):345–56.
- Koets AP, Aduana G, Janss LL, van Weering HJ, Kalis CH, Wentink GH, et al. Genetic variation of susceptibility to *Mycobacterium avium* ssp. *paratuberculosis* infection in dairy cattle. *J Dairy Sci*. 2000;83:2702–8.
- Mortensen H, Nielsen SS, Berg P. Genetic variation and heritability of the antibody response to *Mycobacterium avium* subspecies *paratuberculosis* in Danish Holstein cows. *J Dairy Sci*. 2004;87(7):2108–13.
- Gonda MG, Chang YM, Shook GE, Collins MT, Kirkpatrick BW. Genetic variation of *Mycobacterium avium* ssp. *paratuberculosis* infection in US Holsteins. *J Dairy Sci*. 2006;89(5):1804–12.
- Hinger M, Brandt H, Erhardt G. Heritability estimates for antibody response to *Mycobacterium avium* subspecies *paratuberculosis* in German Holstein cattle. *J Dairy Sci*. 2008;91(8):3237–44.
- Attalla SA, Seykora AJ, Cole JB, Heins BJ. Genetic parameters of milk ELISA scores for Johne's disease. *J Dairy Sci*. 2010;93(4):1729–35.
- Berry DP, Good M, Mullowney P, Cromie AR, More SJ. Genetic variation in serological response to *Mycobacterium avium* subspecies *paratuberculosis* and its association with performance in Irish Holstein-Friesian dairy cows. *Livest Sci*. 2010;131:102–7.
- van Hulzen KJ, Nielsen M, Koets AP, de Jong G, van Arendonk JA, Heuven HC. Effect of herd prevalence on heritability estimates of antibody response to *Mycobacterium avium* subspecies *paratuberculosis*. *J Dairy Sci*. 2011;94(2):992–7.
- Küpper J, Brandt H, Donat K, Erhardt G. Heritability estimates for *Mycobacterium avium* subspecies *paratuberculosis* status of German Holstein cows tested by fecal culture. *J Dairy Sci*. 2012;95(5):2734–9.
- Shook GE, Chaffer M, Wu XL, Ezra E. Genetic parameters for paratuberculosis infection and effect of infection on production traits in Israeli Holsteins. *Anim Genet*. 2012;43(Suppl 1):56–64.
- Zare Y, Shook GE, Collins MT, Kirkpatrick BW. Short communication: heritability estimates for susceptibility to *Mycobacterium avium* subspecies *paratuberculosis* infection defined by ELISA and fecal culture test results in Jersey cattle. *J Dairy Sci*. 2014;97(7):4562–7.
- Gao Y, Cao J, Zhang S, Zhang Q, Sun D. Short communication: heritability estimates for susceptibility to *Mycobacterium avium* ssp. *paratuberculosis* infection in Chinese Holstein cattle. *J Dairy Sci*. 2018;101(8):7274–9.
- Pant SD, Verschoor CP, Schenkel FS, You Q, Kelton DF, Karrow NA. Bovine CLEC7A genetic variants and their association with seropositivity in Johne's disease ELISA. *Gene*. 2014;537(2):302–7.
- Verschoor CP, Pant SD, You Q, Schenkel FS, Kelton DF, Karrow NA. Polymorphisms in the gene encoding bovine interleukin-10 receptor alpha are associated with *Mycobacterium avium* ssp. *paratuberculosis* infection status. *BMC Genet*. 2010;11:23.

16. Pant SD, Verschoor CP, Skelding AM, Schenkel FS, You Q, Biggar GA, et al. Bovine IFNGR2, IL12RB1, IL12RB2, and IL23R polymorphisms and MAP infection status. *Mamm Genome*. 2011;22(9–10):583–8.
17. Pinedo PJ, Buergele CD, Donovan GA, Melendez P, Morel L, Wu R, et al. Association between CARD15/NOD2 gene polymorphisms and paratuberculosis infection in cattle. *Vet Microbiol*. 2009;134(3–4):346–52.
18. Ruiz-Larrahaga O, Garrido JM, Iriondo M, Manzano C, Molina E, Koets AP, et al. Genetic association between bovine NOD2 polymorphisms and infection by *Mycobacterium avium* subsp. paratuberculosis in Holstein-Friesian cattle. *Anim Genet*. 2010;41(6):652–5.
19. Küpper JD, Brandt HR, Erhardt G. Genetic association between NOD2 polymorphism and infection status by *Mycobacterium avium* ssp. paratuberculosis in German Holstein cattle. *Anim Genet*. 2014;45(1):114–6.
20. Pant SD, Verschoor CP, Schenkel FS, You Q, Kelton DF, Karrow NA. Bovine PGLYRP1 polymorphisms and their association with resistance to *Mycobacterium avium* ssp. paratuberculosis. *Anim Genet*. 2011;42(4):354–60.
21. Pinedo PJ, Buergele CD, Donovan GA, Melendez P, Morel L, Wu R, et al. Candidate gene polymorphisms (BoIFNG, TLR4, SLC11A1) as risk factors for paratuberculosis infection in cattle. *Prev Vet Med*. 2009;91(2–4):189–96.
22. Ruiz-Larrahaga O, Garrido JM, Manzano C, Iriondo M, Molina E, Gil A, et al. Identification of single nucleotide polymorphisms in the bovine solute carrier family 11 member 1 (SLC11A1) gene and their association with infection by *Mycobacterium avium* subspecies paratuberculosis. *J Dairy Sci*. 2010;93(4):1713–21.
23. Ruiz-Larrahaga O, Garrido JM, Iriondo M, Manzano C, Molina E, Montes I, et al. SP110 as a novel susceptibility gene for *Mycobacterium avium* subspecies paratuberculosis infection in cattle. *J Dairy Sci*. 2010;93(12):5950–8.
24. Mucha R, Bhide MR, Chakurkar EB, Novak M, Mikula I Sr. Toll-like receptors TLR1, TLR2 and TLR4 gene mutations and natural resistance to *Mycobacterium avium* subsp. paratuberculosis infection in cattle. *Vet Immunol Immunopathol*. 2009;128(4):381–8.
25. Ruiz-Larrahaga O, Manzano C, Iriondo M, Garrido JM, Molina E, Vazquez P, et al. Genetic variation of toll-like receptor genes and infection by *Mycobacterium avium* ssp. paratuberculosis in Holstein-Friesian cattle. *J Dairy Sci*. 2011;94(7):3635–41.
26. Koets A, Santema W, Mertens H, Oostenrijk D, Keestra M, Overdijk M, et al. Susceptibility to paratuberculosis infection in cattle is associated with single nucleotide polymorphisms in toll-like receptor 2 which modulate immune responses against *Mycobacterium avium* subspecies paratuberculosis. *Prev Vet Med*. 2010;93(4):305–15.
27. Pant SD, Schenkel FS, Verschoor CP, You Q, Kelton DF, Moore SS, et al. A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in Holstein cattle. *Genomics*. 2010;95(3):176–82.
28. Gonda MG, Kirkpatrick BW, Shook GE, Collins MT. Identification of a QTL on BTA20 affecting susceptibility to *Mycobacterium avium* ssp. paratuberculosis infection in US Holsteins. *Anim Genet*. 2007;38(4):389–96.
29. Settles M, Zanella R, McKay SD, Schnabel RD, Taylor JF, Whitlock R, et al. A whole genome association analysis identifies loci associated with *Mycobacterium avium* subsp. paratuberculosis infection status in US holstein cattle. *Anim Genet*. 2009;40(5):655–62.
30. Kirkpatrick BW, Shi X, Shook GE, Collins MT. Whole-genome association analysis of susceptibility to paratuberculosis in Holstein cattle. *Anim Genet*. 2011;42(2):149–60.
31. Minozzi G, Buggiotti L, Stella A, Strozzi F, Luini M, Williams JL. Genetic loci involved in antibody response to *Mycobacterium avium* ssp. paratuberculosis in cattle. *PLoS One*. 2010;5(6):e11117.
32. Zanella R, Settles ML, McKay SD, Schnabel R, Taylor J, Whitlock RH, et al. Identification of loci associated with tolerance to Johne's disease in Holstein cattle. *Anim Genet*. 2011;42(1):28–38.
33. van Hulzen KJ, Schopen GC, van Arendonk JA, Nielen M, Koets AP, Schrooten C, et al. Genome-wide association study to identify chromosomal regions associated with antibody response to *Mycobacterium avium* subspecies paratuberculosis in milk of Dutch Holstein-Friesians. *J Dairy Sci*. 2012;95(5):2740–8.
34. Minozzi G, Williams JL, Stella A, Strozzi F, Luini M, Settles ML, et al. Meta-analysis of two genome-wide association studies of bovine paratuberculosis. *PLoS One*. 2012;7(3):e32578.
35. Zare Y, Shook GE, Collins MT, Kirkpatrick BW. Genome-wide association analysis and genomic prediction of *Mycobacterium avium* subspecies paratuberculosis infection in US Jersey cattle. *PLoS One*. 2014;9(2):e88380.
36. Alpay F, Zare Y, Kamalludin MH, Huang X, Shi X, Shook GE, et al. Genome-wide association study of susceptibility to infection by *Mycobacterium avium* subspecies paratuberculosis in Holstein cattle. *PLoS One*. 2014;9(12):e111704.
37. Sallam AM, Zare Y, Alpay F, Shook GE, Collins MT, Alsheikh S, et al. An across-breed genome wide association analysis of susceptibility to paratuberculosis in dairy cattle. *J Dairy Res*. 2017;84(1):61–7.
38. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*. 2009;84(2):210–23.
39. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75.
40. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, et al. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol*. 2009;10(4):R42.
41. Haldar T, Ghosh S. Effect of population stratification on false positive rates of population-based association analyses of quantitative traits. *Ann Hum Genet*. 2012;76(3):237–45.
42. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82.
43. R Core Team. R: A language and environment for statistical computing. R Foundation for statistical Computing Vienna, Austria. 2015. <https://www.R-project.org/>. Accessed 21 Dec 2018.
44. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*. 2007;177(1):577–85.
45. Amin N, van Duijn CM, Aulchenko YS. A genomic background based method for association analysis in related individuals. *PLoS One*. 2007;2(12):e1274.
46. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. 2007;23(10):1294–6.
47. Thornton T, McPeck MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet*. 2010;86(2):172–84.
48. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661–78.
49. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res*. 2011;39(Web Server issue):W316–22.
50. Hu ZL, Fritz ER, Reecy JM. AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic Acids Res*. 2007;35(Database issue):D604–9.
51. Chen L. Co-inhibitory molecules of the B7-CD28 family in the control of T-cell immunity. *Nat Rev Immunol*. 2004;4(5):336–47.
52. Küpper J, Brandt H, Donat K, Erhardt G. Phenotype definition is a main point in genome-wide association studies for bovine *Mycobacterium avium* ssp. paratuberculosis infection status. *Animal*. 2014;8(10):1586–93.
53. Coussens PM. *Mycobacterium paratuberculosis* and the bovine immune system. *Anim Health Res Rev*. 2001;2(2):141–61.
54. Coussens PM, Pudrith CB, Skovgaard K, Ren X, Suchyta SP, Stabel JR, et al. Johne's disease in cattle is associated with enhanced expression of genes encoding IL-5, GATA-3, tissue inhibitors of matrix metalloproteinases 1 and 2, and factors promoting apoptosis in peripheral blood mononuclear cells. *Vet Immunol Immunopathol*. 2005;105(3–4):221–34.
55. Sohal JS, Singh SV, Tyagi P, Subhodh S, Singh PK, Singh AV, et al. Immunology of mycobacterial infections: with special reference to *Mycobacterium avium* subspecies paratuberculosis. *Immunobiology*. 2008;213(7):585–98.
56. Stabel JR. Transitions in immune responses to *Mycobacterium paratuberculosis*. *Vet Microbiol*. 2000;77(3–4):465–73.
57. Koets A, Rutten V, Hoek A, van Mil F, Müller K, Bakker D, et al. Progressive bovine paratuberculosis is associated with local loss of CD4(+) T cells, increased frequency of gamma delta T cells, and related changes in T-cell function. *Infect Immun*. 2002;70(7):3856–64.
58. Lohoff M, Ferrick D, Mittrucker HW, Duncan GS, Bischof S, Rollinghoff M, et al. Interferon regulatory factor-1 is required for a T helper 1 immune response in vivo. *Immunity*. 1997;6(6):681–9.
59. McElligott DL, Phillips JA, Stillman CA, Koch RJ, Mosier DE, Hobbs MV. CD4+ T cells from IRF-1-deficient mice exhibit altered patterns of cytokine expression and cell subset homeostasis. *J Immunol*. 1997;159(9):4180–6.

60. Taki S, Sato T, Ogasawara K, Fukuda T, Sato M, Hida S, et al. Multistage regulation of Th1-type immune responses by the transcription factor IRF-1. *Immunity*. 1997;6(6):673–9.
61. He J, You X, Zeng Y, Yu M, Zuo L, Wu Y. Mycoplasma genitalium-derived lipid-associated membrane proteins activate NF-kappaB through toll-like receptors 1, 2, and 6 and CD14 in a MyD88-dependent pathway. *Clin Vaccine Immunol*. 2009;16(12):1750–7.
62. Pérez-Otaño I, Luján R, Tavalin SJ, Plomann M, Modregger J, Liu XB, et al. Endocytosis and synaptic removal of NR3A-containing NMDA receptors by PACSIN1/syndapin1. *Nat Neurosci*. 2006;9(5):611–21.
63. Esashi E, Bao M, Wang YH, Cao W, Liu YJ. PACSIN1 regulates the TLR7/9-mediated type I interferon response in plasmacytoid dendritic cells. *Eur J Immunol*. 2012;42(3):573–9.
64. Gupta S, Lee A, Hu C, Fanzo J, Goldberg I, Cattoretti G, et al. Molecular cloning of IBP, a SWAP-70 homologous GEF, which is highly expressed in the immune system. *Hum Immunol*. 2003;64(4):389–401.
65. Rodrigues-Lima F, Josephs M, Katan M, Cassinat B. Sequence analysis identifies TTRAP, a protein that associates with CD40 and TNF receptor-associated factors, as a member of a superfamily of divalent cation-dependent phosphodiesterases. *Biochem Biophys Res Commun*. 2001;285(5):1274–9.
66. Gu Y, Chae HD, Siefing JE, Jasti AC, Hildeman DA, Williams DA. RhoH GTPase recruits and activates Zap70 required for T cell receptor signaling and thymocyte development. *Nat Immunol*. 2006;7(11):1182–90.
67. Shirkani A, Shahrooei M, Azizi G, Rokni-Zadeh H, Abolhassani H, Farrokhi S, et al. Novel mutation of ZAP-70-related combined immunodeficiency: first case from the National Iranian Registry and review of the literature. *Immunol Investig*. 2017;46(1):70–9.
68. Leite FL, Eslabão LB, Pesch B, Bannantine JP, Reinhardt TA, Stabel JR. ZAP-70, CTLA-4 and proximal T cell receptor signaling in cows infected with *Mycobacterium avium* subsp. *paratuberculosis*. *Vet Immunol Immunopathol*. 2015;167(1–2):15–21.
69. Murphy JM, Soboleva TA, Mirza S, Ford SC, Olsen JE, Chen J, et al. Clarification of the role of N-glycans on the common beta-subunit of the human IL-3, IL-5 and GM-CSF receptors and the murine IL-3 beta-receptor in ligand-binding and receptor activation. *Cytokine*. 2008;42(2):234–42.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions





Short communication: Heritability estimates for susceptibility to *Mycobacterium avium* ssp. *paratuberculosis* infection in Chinese Holstein cattle

Y. Gao,* J. Cao,† S. Zhang,* Q. Zhang,* and D. Sun*¹

*Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, National Engineering Laboratory of Animal Breeding, College of Animal Science and Technology, and

†College of Veterinary Medicine, China Agricultural University, Beijing 100193, China

ABSTRACT

Paratuberculosis in ruminants, which is caused by *Mycobacterium avium* ssp. *paratuberculosis* (MAP), is a contagious, chronic enteric disease associated with economic losses, animal welfare, and health implications in dairy cattle production. In this study, we estimated the variance components and heritability of susceptibility to MAP infection in Chinese Holstein cattle. We collected 4,937 serum samples from cows in 7 dairy herds in the Beijing region of China and used the ELISA test to detect antibodies to MAP. Three statistical models were implemented to estimate heritabilities: (1) a linear model (ELISA sample-to-positive ratios as a continuous trait); (2) a binary threshold model (positive/negative from ELISA results); and (3) an ordered threshold model (ELISA results as an ordered categorical model with categories 1 to 5 corresponding to negative, uncertain, mildly positive, intermediate positive, and strongly positive). The heritability estimates ranged from 0.0389 to 0.1069, indicating that genetic factors affect MAP infection susceptibility in Chinese Holstein cattle.

Key words: paratuberculosis, Johne's disease, heritability, Chinese Holstein

Short Communication

Paratuberculosis, or Johne's disease (JD) as it is commonly known, is caused by *Mycobacterium avium* ssp. *paratuberculosis* (MAP) and is a contagious, chronic, enteric disease of ruminants (Clarke, 1997). Animals suffering from JD show several classic clinical signs including diarrhea and weight loss and they eventually die. Ten percent of infected cows die of this disease every year, and the remaining 90% of diseased

cows are slaughtered (Chen, 2016). *Mycobacterium avium* ssp. *paratuberculosis* can hide in many substrates such as colostrum, milk, and feces, from where it can be transmitted to susceptible individuals (Nielsen et al., 2008), with the major route being the fecal-oral route (Lombard, 2011). Since 2000, JD has been reported around the world with an average herd prevalence (the proportion of herds with at least one JD case) between 30 and 50% (Boelaert et al., 2000; Muskens et al., 2000; Tiwari et al., 2006; NAHMS, 2007; Haghkhah et al., 2008; Defra, 2009; Good et al., 2009; Nielsen, 2009a,b,c). According to recent reports, herd prevalence has reached 91.1% in the United States (Lombard et al., 2013) and 65.4% in Egypt (Amin et al., 2015). The associated costs of JD, such as decreased reproductive and productive efficiency and the need for diagnostic testing, have resulted in JD affecting the global dairy industry severely (Ott et al., 1999). Unfortunately, because no effective cure or vaccines for JD exist, the only way to decrease its prevalence is to implement control programs such as those described by Ferrouillet et al. (2009). The main steps of control programs include identifying positive cows using ELISA or other detection methods, isolating them to eliminate transmission routes, and finally eliminating positive cows that show clinical symptoms.

Selection for disease-resistant animals could produce offspring with increased average ability to resist MAP infection; thus, genetic selection for animals with JD resistance would be an effective means to control JD. Several early studies performed on Holstein and Jersey cattle that estimated the heritability of infection with MAP reported estimates ranging from 0.031 to 0.283 (Koets et al., 2000; Mortensen et al., 2004; Gonda et al., 2006; Hinger et al., 2008; Attalla et al., 2010; Berry et al., 2010; van Hulzen et al., 2011; Küpper et al., 2012; Shook et al., 2012; Zare et al., 2014). Despite variable results, these previous studies show that MAP-infection resistance in cattle has a genetic background.

Chinese Holstein cattle are the result of cross breeding between Chinese Yellow cattle and European Holsteins

Received May 31, 2017.

Accepted February 7, 2018.

¹ Corresponding author: sundx@cau.edu.cn

over the past 100 yr. Foreign Holstein bulls, semen, and embryos, mainly from the United States, and a few from Canada and Europe, have been continuously imported, and these have been used directly for AI or for crossing with Chinese Holstein cows through planned mating to generate breeding bulls (Sun et al., 2009). According to recent investigations on MAP in large-scale dairy farms in some Chinese provinces, herd-level prevalence of MAP infection has reached 100% (Sun et al., 2015; Cui et al., 2016) and within-herd prevalence has also gradually increased. To the best of our knowledge, no systematic study to analyze the genetics of susceptibility to MAP infection in Chinese Holstein cattle has been undertaken. To genetically improve resistance to MAP infection, implementation of a comprehensive genetic evaluation is a crucial step. Therefore, the objective of this study was to estimate the heritability of susceptibility to MAP infection based on antibody titers from Chinese Holstein cattle, and to use the data in the context of dairy cattle selection.

The protocols for collecting blood samples from the experimental animals were reviewed and approved by the Institutional Animal Care and Use Committee at China Agricultural University (Permit Number DK996). All the experiments were performed in accordance with the relevant approved guidelines and regulations. In total, 8,214 Chinese Holstein cows from 7 herds at the Beijing Sanyuan Dairy Farm Center were sampled, 4,937 of which were ≤ 24 mo of age and used for heritability estimation. All cows were fed under the same feeding and management system, and regular quarantine inspections of the herds were conducted so that these 7 herds represented the situation of JD infection status of dairy herds in the Beijing region. Although all 7 herds belong to the Beijing Sanyuan Dairy Farm Centre, there was no movement of animals between the herds because of the strict management system based on regular quarantine inspections twice annually. In addition, there were no records of other diseases such as tuberculosis or subclinical mastitis in the cows included in this study. The cows belonged to 436 sire families with an average of 11.3 daughters per sire. Pedigree relationships for the 4,937 cows were traced back 5 generations. Blood samples (500 μ L) were collected from the caudal vein of each cow during the regular quarantine inspection of the farms in September 2014. All cows within a herd were sampled on the same day. Serum extracted from blood samples was stored at 4°C until testing, which took place within 5 d of collection. With the ELISA method, the antibody levels in the serum samples were determined using the *Mycobacterium paratuberculosis* Antibody Test Kit (Idexx Laboratories Inc., Westbrook, ME) following the manufacturer's instructions. The MAP status of an animal was expressed

as the sample-to-positive (S/P) ratio multiplied by 100: S/P ratio = $100 \times [(\text{optical density (OD) value of the sample} - \text{OD of the negative control}) / (\text{OD of a positive sample} - \text{OD of the negative control})]$, where $S/P \leq 0.45$ is negative; $0.45 < S/P < 0.55$ is uncertain; $0.55 \leq S/P < 1$ is mildly positive; $1 \geq S/P < 2$ is intermediate positive; and $S/P > 2$ is strongly positive. Parity levels were from 0 (nulliparous) to 4 (\geq fourth parity). The ages of the cows sampled ranged from 25 to 162 mo. Ages were grouped in 17 levels by 6-mo intervals (all cows older than 120 mo were in level 21), and 77.17% of individuals were between 25 and 60 mo old. Three traits were defined for heritability estimation according to the ELISA results: (1) ELISA S/P were taken as a continuous trait; (2) ELISA results were taken as a binary trait (0 = negative, 1 = positive), where mildly positive, intermediate positive, and strongly positive ELISA results were considered ELISA positive (77 uncertain ELISA results were excluded); and (3) ELISA results were taken as an ordered categorical trait with categories 1 to 5 corresponding to negative, uncertain, mildly positive, intermediate positive, and strongly positive, respectively.

We used an animal genetic model and 3 statistical models: a linear model, a binary threshold model, and an ordered threshold model to estimate the variance components. Because of non-normality, log-transformed ELISA S/P ratios were used in the linear model [$\text{ELISA} = \log_{10}(\text{ELISA} + 0.01)$; 0.01 was added to avoid $\log(0)$]. Before estimating variance components, a preliminary analysis of the fixed effects including herd and parity was conducted using the different models. The fixed effects for herd and parity showed significant ($P < 0.01$) effects in all 3 models.

The following linear model was used for the analysis of log-transformed ELISA S/P ratios:

$$y_{ijkn} = p_i + h_j + a_k + e_{ijkn},$$

where y_{ijkn} is the transformed ELISA S/P ratio, p_i is the fixed effect of parity, h_j is the fixed effect of herd, a_k is the additive genetic effect, and e_{ijkn} is the residual. Random effects were assumed to be normally distributed:

$$a \sim N(0, \mathbf{A}\sigma_a^2),$$

$$e \sim N(0, \mathbf{I}\sigma_e^2),$$

where \mathbf{A} is the numerator relationship matrix, σ_a^2 represents the individual additive variance, \mathbf{I} is an identity matrix, and σ_e^2 represents the residual variance. Variance components were estimated by the DMUAI mod-

Table 1. Distribution of cow *Mycobacterium avium* ssp. *paratuberculosis* (MAP) infection status by herd and parity

Variable	Level	Negative		Uncertain		Positive ¹		Total
		No. of cows	%	No. of cows	%	No. of cows	%	
Total		4,132	83.69	77	1.56	728	14.75	4,937
Herd	1	413	77.78	10	1.88	108	20.34	531
	2	1,071	80.95	16	1.21	236	17.84	1,323
	3	485	83.33	1	0.17	96	16.49	582
	4	585	88.91	20	3.04	53	8.05	658
	5	753	83.76	20	2.22	126	14.02	899
	6	536	89.63	7	1.17	55	9.20	598
	7	289	83.53	3	0.87	54	15.61	346
Parity	0	238	88.81	2	0.75	28	10.45	268
	1	1,657	86.53	31	1.62	227	11.8	1,915
	2	1,054	79.67	28	2.12	241	18.22	1,323
	3	552	82.39	3	0.45	115	17.16	670
	≥4	631	82.92	13	1.71	117	15.37	761

¹Mildly positive, intermediate positive, and strongly positive ELISA results were considered ELISA positive.

ule using the REML method in the DMU software (Madsen and Jensen, 2013).

In the binary threshold model, only ELISA-negative and ELISA-positive results were used. The estimation of the variance components was carried out by the RJMC module using the Bayesian Markov chain Monte Carlo (MCMC) method in the DMU software (Madsen and Jensen, 2013). The binary threshold model was

$$l_{ijkn} = p_i + h_j + a_k + e_{ijkn},$$

where l_{ijkn} is the liability to JD based on ELISA S/P ratios for parity i , herd j , and animal k , and the other parameters are defined above.

In the ordered threshold model, the estimation of the variance components was carried out by the RJMC module using the Bayesian MCMC method in the DMU software (Madsen and Jensen, 2013). The fitted model was

$$\lambda_{ijkn} = p_i + h_j + a_k + e_{ijkn},$$

where λ_{ijkn} is the liability to JD based on ELISA S/P ratios for parity i , herd j , and animal k , and the other parameters are defined as in the linear model.

Narrow-sense heritability was calculated as

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2},$$

where σ_a^2 is the individual additive variance and σ_e^2 is the residual variance. Standard errors were reported.

After screening the 4,937 Chinese Holstein cows using ELISA methodology, 728 cows, distributed across the 7 herds, were found to be positive for MAP. The frequency of positive cows across these herds varied from

8.05 to 20.34%, with an average prevalence of 14.51% (Table 1). Second- and third-parity cows showed the highest positivity rates at 18.22 and 17.16%, respectively (Table 1). The frequency of ELISA-positive cows by age, as shown in Figure 1, varied from 3.23% (97–102 mo) to 20.83% (43–48 mo), with the highest frequency recorded between 25 and 96 mo of age. The average frequencies of the 5 categories were 83.69% (negative), 1.56% (uncertain), 4.33% (mildly positive), 8.47% (intermediate positive), and 1.94% (strongly positive).

The estimated heritability obtained from the 3 models in this study ranged from 0.0389 (ordered threshold model) to 0.1069 (binary threshold model; Table 2), a finding in line with the range reported in previous studies (0.031 to 0.283). The discrepancies in the estimated heritabilities between these studies may be explained by several factors: (1) the different population sizes, from 3,020 (Koets et al., 2000) to 684,364 (van Hulzen et al., 2011); (2) the different statistical methods (such as linear model vs. threshold model, and sire model vs. animal model); and (3) the different diagnostic tests used to determine infection status. Additionally, infection prevalence in different herds can contribute to variations in the estimated heritability (van Hulzen et al., 2011), and it is known that variance components are population-specific and will not necessarily be the same across different breeds or different populations. The heritabilities estimated in both the linear model (0.0488 ± 0.0200) and the ordered threshold model (0.0389 ± 0.0098) were lower than that in the binary threshold model (0.1069 ± 0.0394). According to the study by Zare et al. (2014), the ELISA S/P ratio as a continuous trait captures a whole spectrum of variations in susceptibility to MAP infection among study animals, whereas classifying study results into binary categories compresses information, which may result in a lower heritability estimate. However, we cannot

Table 2. Heritability with standard errors and variance component estimates¹ for different models and data sets

No. of cows	Model	$h^2 \pm \text{SE}$	V_A	V_P
4,937	Linear ²	0.0488 ± 0.0200	0.0156	0.3199
4,860	Binary threshold ³	0.1069 ± 0.0394	0.1197	1.1197
4,937	Ordered threshold ⁴	0.0389 ± 0.0098	0.0358	0.9214

¹ V_A = additive genetic variance; V_P = phenotypic variance.

²Logarithmically transformed ELISA sample-to-positive (S/P) ratios.

³Optical density (OD) ≤ 0.45 were considered ELISA negative; $0.45 < \text{OD} < 0.55$ were considered ELISA uncertain and therefore removed from further analysis; $\text{OD} \geq 0.55$ were considered ELISA positive.

⁴The 5 categories (negative, uncertain, mildly positive, intermediate positive, and strongly positive) were defined according to the ELISA S/P ratios.

be sure that a binary model would result in a loss of information that affects mostly the genetic component of the total variance. Binarization is a simplification of the original problem, but heritability may not necessarily be smaller. In the study by Hinger et al. (2008), the estimated heritability from the linear model was larger than that from the binary model, whereas the opposite was found by Berry et al. (2010). According to Zare et al. (2014), the estimated heritability from linear model was lower than that from the ordered threshold model. Different from their results, the estimated heritability in our linear model (0.0498 ± 0.0200) was higher than that from the ordered threshold model (0.0389 ± 0.0098). However, Attalla et al. (2010) reported that there was no evidence to recommend a linear model over a threshold model for genetic evaluation. In the current study, the non-normality of the ELISA S/P ratio, even after transformation, may have introduced bias into heritability estimates.

In addition, we estimated heritability in 3 age groups (25–36 mo, 37–48 mo, and >48 mo) based on 3 models with almost 1,500 cows in each group (Table 3). The heritabilities estimated by linear and ordered threshold models were consistent with the estimates in previous studies (Koets et al., 2000; Mortensen et al., 2004; Gonda et al., 2006; Hinger et al., 2008; Attalla et al., 2010; Berry et al., 2010; van Hulzen et al., 2011; Küpper et al., 2012; Shook et al., 2012; Zare et al., 2014). In addition, we observed that the additive genetic variance did not reach convergence in the binary threshold model so the estimated heritability was higher than that in other 2 models (0.9965 vs. 0.1712 or 0.1096) in the age group of 25 to 36 mo, indicating that the estimate of 0.9965 was not convincing. The reason for this is likely that the number of cows in this age group was relatively small for heritability estimation.

Several diagnostic tests were used in the early studies on MAP. In the first study to estimate the heritability

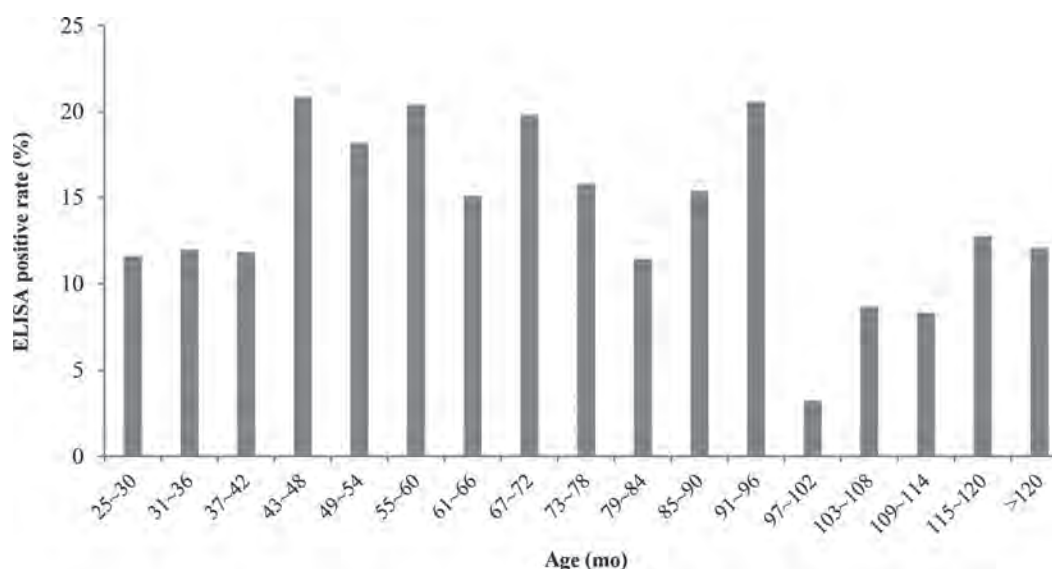
**Figure 1.** Estimated frequency of *Mycobacterium avium* ssp. *paratuberculosis* (MAP) infection-positive cows by age (mo).

Table 3. Heritability by age with standard errors and variance component estimates¹ for different models and data sets

Model	Age (mo)	No. of cows	$h^2 \pm SE$	V_A	V_P
Linear ²	25–36	1,745	0.1712 ± 0.0615	0.0505	0.2948
	37–48	1,310	0.0433 ± 0.0454	0.014	0.3241
	>48	1,882	0.0076 ± 0.0249	0.0026	0.3391
Binary threshold ³	25–36	1,724	0.9965 ± 0.0030	285.1671	286.1671
	37–48	1,276	0.1445 ± 0.0817	0.1689	1.1689
	>48	1,860	0.1381 ± 0.0688	0.1602	1.1602
Ordered threshold ⁴	25–36	1,745	0.1096 ± 0.0284	0.0372	0.3399
	37–48	1,310	0.1041 ± 0.0283	0.0399	0.3827
	>48	1,882	0.0961 ± 0.0255	0.0449	0.4672

¹ V_A = additive genetic variance; V_P = phenotypic variance.

²Logarithmically transformed ELISA sample-to-positive (S/P) ratios.

³Optical density (OD) ≤ 0.45 were considered ELISA negative; $0.45 < OD < 0.55$ were considered ELISA uncertain and therefore removed from further analysis; $OD \geq 0.55$ were considered ELISA positive.

⁴The 5 categories (negative, uncertain, mildly positive, intermediate positive, and strongly positive) were defined according to the ELISA S/P ratios.

of susceptibility to MAP, Koets et al. (2000) detected 3,020 MAP-positive Dutch dairy cattle based on post-mortem data. Although postmortem analysis is one of the most accurate methods, it is not applicable for large-scale and routine disease diagnosis. In subsequent research, 3 detection methods were used to determine the MAP status of dairy cows. The methods included milk ELISA (Mortensen et al., 2004; Attalla et al., 2010; van Hulzen et al., 2011), fecal culture (Gonda et al., 2006; Küpper et al., 2012; Zare et al., 2014), and serum ELISA (Gonda et al., 2006; Hinger et al., 2008; Berry et al., 2010; Shook et al., 2012; Zare et al., 2014). When methods differ in their sensitivities, this will cause variation in the MAP status of the cows. Eamens et al. (2000) suggested that there is a low correlation between fecal culture and serum ELISA in judging MAP status, whereas Gonda et al. (2006) concluded that the analysis results for serum and milk ELISA and fecal culture showed similar heritabilities. In addition, Hinger et al. (2008) and Küpper et al. (2012) used serum ELISA and fecal culture to determine the MAP status of German Holstein cattle and both methods showed the same tendency of MAP prevalence over age groups. Zare et al. (2014) estimated the heritability for susceptibility to MAP infection in US Jersey cattle using serum ELISA and fecal culture simultaneously. Similar heritabilities for both diagnostic tests were obtained, but the authors emphasized that the heritability estimated from the combination of serum ELISA and fecal culture (0.197) might be closer to the true heritability value.

Whitlock and Buergetl (1996) classified MAP infections into 4 categories according to the appearance of clinical signs and disease severity (i.e., “silent” infection, subclinical disease, clinical disease, and advanced clinical disease). They found that after newborns or

heifers were infected with MAP, it took 2 yr for them to show clinical signs of the disease. In the present study, most ELISA-positive cows were aged 25 to 96 mo, a finding consistent with the known infection patterns of JD. Additionally, the age distribution of MAP-positive animals was consistent with the results of Hinger et al. (2008) and Küpper et al. (2012).

In conclusion, this is the first study to estimate the heritability of susceptibility to MAP infection in Chinese Holstein cattle. The estimated heritability ranged from 0.0389 to 0.1069, which is in agreement with the range reported in previous studies. Although the heritability value is low, it should be possible to select individuals in breeding programs for resistance to MAP infections as a means of controlling the prevalence of JD.

ACKNOWLEDGMENTS

This work was supported financially by the National Natural Science Foundation (31472065, Beijing), Beijing Dairy Industry Innovation Team (BAIC06-2018, Beijing), National Science and Technology Program of China (2013AA102504, Beijing), and the Program for Changjiang Scholar and Innovation Research Team in University (IRT1191, Beijing). We thank Sandra Cheesman, PhD, from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

REFERENCES

- Amin, A. S., C. Y. Hsu, S. F. Darwish, P. Ghosh, E. M. AbdEl-Fatah, T. S. Behour, and A. M. Talaat. 2015. Ecology and genomic features of infection with *Mycobacterium avium* subspecies *paratuberculosis* in Egypt. *Microbiology* 161:807–818. <https://doi.org/10.1099/mic.0.000051>.

- Attalla, S. A., A. J. Seykora, J. B. Cole, and B. J. Heins. 2010. Genetic parameters of milk ELISA scores for Johne's disease. *J. Dairy Sci.* 93:1729–1735. <https://doi.org/10.3168/jds.2009-2625>.
- Berry, D. P., M. Good, P. Mullooney, A. R. Cromie, and S. J. More. 2010. Genetic variation in serological response to *Mycobacterium avium* subspecies *paratuberculosis* and its association with performance in Irish Holstein-Friesian dairy cows. *Livest. Sci.* 131:102–107. <https://doi.org/10.1016/j.livsci.2010.03.007>.
- Boelaert, F., K. Walravens, P. Biont, J. P. Vermeersch, D. Berkvens, and J. Godfroid. 2000. Prevalence of paratuberculosis (Johne's disease) in the Belgian cattle population. *Vet. Microbiol.* 77:269–281.
- Chen, F. 2016. Epidemiological survey of Johne's disease in Beijing dairy herds and initial implementation of control program. MS Thesis. China Agricultural Univ., Beijing.
- Clarke, C. J. 1997. The pathology and pathogenesis of paratuberculosis in ruminants and other species. *J. Comp. Pathol.* 116:217–261.
- Cui, J. L., Y. C. Zhao, T. Wu, and Z. Y. Shen. 2016. Epidemiological investigation of bovine paratuberculosis in scaled dairy farms in some provinces of China in 2015. *Anim. Husb. Feed Sci.* 37:97–99.
- Defra. 2009. SB4022: An Integrated Strategy to Determine the Herd Level Prevalence of Johne's Disease in the UK Dairy Herd. Department for Environment, Food and Rural Affairs (Defra). Accessed Apr. 15, 2017. <http://webarchive.nationalarchives.gov.uk/20130402151656/http://archive.defra.gov.uk/foodfarm/farmanimal/diseases/atoz/johnes/index.htm>.
- Eamens, G. J., R. J. Whittington, I. B. Marsh, M. J. Turner, V. Saunders, P. D. Kemsley, and D. Rayward. 2000. Comparative sensitivity of various faecal culture methods and ELISA in dairy cattle herds with endemic Johne's disease. *Vet. Microbiol.* 77:357–367. [https://doi.org/10.1016/S0378-1135\(00\)00321-7](https://doi.org/10.1016/S0378-1135(00)00321-7).
- Ferrouillet, C., S. J. Wells, W. L. Hartmann, S. M. Godden, and J. Carrier. 2009. Decrease of Johne's disease prevalence and incidence in six Minnesota, USA, dairy cattle herds on a long-term management program. *Prev. Vet. Med.* 88:128–137. <https://doi.org/10.1016/j.prevetmed.2008.08.001>.
- Gonda, M. G., Y. M. Chang, G. E. Shook, M. T. Collins, and B. W. Kirkpatrick. 2006. Genetic variation of *Mycobacterium avium* ssp. *paratuberculosis* infection in US Holsteins. *J. Dairy Sci.* 89:1804–1812. [https://doi.org/10.3168/jds.S0022-0302\(06\)72249-4](https://doi.org/10.3168/jds.S0022-0302(06)72249-4).
- Good, M., T. Clegg, H. Sheridan, D. Yearsely, T. O'Brien, J. Egan, and P. Mullooney. 2009. Prevalence and distribution of paratuberculosis (Johne's disease) in cattle herds in Ireland. *Ir. Vet. J.* 62:597–606. <https://doi.org/10.1186/2046-0481-62-9-597>.
- Haghkhah, M., M. Ansari-Lari, A. M. Novin-Baheran, and A. Bahramy. 2008. Herd-level prevalence of *Mycobacterium avium* subspecies *paratuberculosis* by bulk-tank milk PCR in Fars province (southern Iran) dairy herds. *Prev. Vet. Med.* 86:8–13. <https://doi.org/10.1016/j.prevetmed.2008.03.010>.
- Hinger, M., H. Brandt, and G. Erhardt. 2008. Heritability estimates for antibody response to *Mycobacterium avium* subspecies *paratuberculosis* in German Holstein cattle. *J. Dairy Sci.* 91:3237–3244. <https://doi.org/10.3168/jds.2008-1021>.
- Koets, A. P., G. Adugna, L. L. Janss, H. J. van Weering, C. H. Kalis, G. H. Wentink, V. P. Rutten, and Y. H. Schukken. 2000. Genetic variation of susceptibility to *Mycobacterium avium* ssp. *paratuberculosis* infection in dairy cattle. *J. Dairy Sci.* 83:2702–2708. [https://doi.org/10.3168/jds.S0022-0302\(00\)75164-2](https://doi.org/10.3168/jds.S0022-0302(00)75164-2).
- Küpper, J., H. Brandt, K. Donat, and G. Erhardt. 2012. Heritability estimates for *Mycobacterium avium* subspecies *paratuberculosis* status of German Holstein cows tested by fecal culture. *J. Dairy Sci.* 95:2734–2739. <https://doi.org/10.3168/jds.2011-4994>.
- Lombard, J. E. 2011. Epidemiology and economics of *paratuberculosis*. *Vet. Clin. North Am. Food Anim. Pract.* 27:525–535. <https://doi.org/10.1016/j.cvfa.2011.07.012>.
- Lombard, J. E., I. A. Gardner, S. R. Jafarzadeh, C. P. Fossler, B. Harris, R. T. Capsel, B. A. Wagner, and W. O. Johnson. 2013. Herd-level prevalence of *Mycobacterium avium* ssp. *paratuberculosis* infection in United States dairy herds in 2007. *Prev. Vet. Med.* 108:234–238. <https://doi.org/10.1016/j.prevetmed.2012.08.006>.
- Madsen, P., and J. Jensen. 2013. A user's guide to DMU. A package for analyzing multivariate mixed models. Version 6. release 5.2. Center for Quantitative Genetics and Genomics, Dept. of Molecular Biology and Genetics, University of Aarhus, Research Centre Foulum, Tjele, Denmark.
- Mortensen, H., S. S. Nielsen, and P. Berg. 2004. Genetic variation and heritability of the antibody response to *Mycobacterium avium* subspecies *paratuberculosis* in Danish Holsteins cows. *J. Dairy Sci.* 87:2108–2113. [https://doi.org/10.3168/jds.S0022-0302\(04\)70029-6](https://doi.org/10.3168/jds.S0022-0302(04)70029-6).
- Muskens, J., H. W. Barkema, E. W. Russchen, C. van Maanen, Y. H. Schukken, and D. Bakker. 2000. Prevalence and regional distribution of paratuberculosis in dairy herds in the Netherlands. *Vet. Microbiol.* 77:253–261.
- National Animal Health Monitoring System (NAHMS). 2007. Part I: Reference of Dairy Health and Management in the United States. USDA, Animal and Plant Health Inspection Service, Veterinary Service, Center for Epidemiology and Animal Health, Fort Collins, CO.
- Nielsen, S. S. 2009a. Parameters used to assess the efforts to control paratuberculosis in Denmark. Pages 14–20 in Monitoring success of paratuberculosis programs: Proc. 2nd Paratuberculosis Forum. Minneapolis, MN. IDF, Brussels, Belgium.
- Nielsen, S. S. 2009b. Paratuberculosis in dairy cattle—Epidemiological studies used for design of a control programme in Denmark. Dr. Med. Vet. Thesis. Department of Large Animal Sciences. University of Copenhagen, Denmark.
- Nielsen, S. S. 2009c. Programmes on Paratuberculosis in Europe. Pages 101–108 in Proc. 10th Int. Colloq. Paratuberculosis, Minneapolis, MN. Univ. of Minnesota, Minneapolis.
- Nielsen, S. S., H. Bjerre, and N. Toft. 2008. Colostrum and milk as risk factors for infection with *Mycobacterium avium* subspecies *paratuberculosis* in dairy cattle. *J. Dairy Sci.* 91:4610–4615. <https://doi.org/10.3168/jds.2008-1272>.
- Ott, S. L., S. J. Wells, and B. A. Wagner. 1999. Herd-level economic losses associated with Johne's disease on US dairy operations. *Prev. Vet. Med.* 40:179–192.
- Shook, G. E., M. Chaffer, X. L. Wu, and E. Ezra. 2012. Genetic parameters for paratuberculosis infection and effect of infection on production traits in Israeli Holsteins. *Anim. Genet.* 43(Suppl. 1):56–64. <https://doi.org/10.1111/j.1365-2052.2012.02349.x>.
- Sun, D. X., J. Jia, Y. Ma, Y. Zhang, Y. C. Wang, Y. Yu, and Y. Zhang. 2009. Effects of DGAT1 and GHR on milk yield and milk composition in the Chinese dairy population. *Anim. Genet.* 40:997–1000.
- Sun, Y., S. C. Ma, H. Dong, and X. Y. Wang. 2015. Serological investigation and analysis of bovine paratuberculosis in some provinces of China. *Progress in Veterinary Medicine* 36:118–120.
- Tiwari, A., J. A. VanLeeuwen, S. L. McKenna, G. P. Keefe, and H. W. Barkema. 2006. Johne's disease in Canada. Part I: Clinical symptoms, pathophysiology, diagnosis, and prevalence in dairy herds. *Can. Vet. J.* 47:874–882.
- van Hulzen, K. J., M. Nielen, A. P. Koets, G. de Jong, J. A. van Arendonk, and H. C. Heuven. 2011. Effect of herd prevalence on heritability estimates of antibody response to *Mycobacterium avium* subspecies *paratuberculosis*. *J. Dairy Sci.* 94:992–997. <https://doi.org/10.3168/jds.2010-3472>.
- Whitlock, R. H., and C. Buergelt. 1996. Preclinical and clinical manifestations of paratuberculosis (including pathology). *Vet. Clin. North Am. Food Anim. Pract.* 12:345–356.
- Zare, Y., G. E. Shook, M. T. Collins, and B. W. Kirkpatrick. 2014. Short communication: Heritability estimates for susceptibility to *Mycobacterium avium* subspecies *paratuberculosis* infection defined by ELISA and fecal culture test results in Jersey cattle. *J. Dairy Sci.* 97:4562–4567. <https://doi.org/10.3168/jds.2013-7426>.

Benchmarking 24 combinations of genotype pre-phasing and imputation software for SNP arrays in pigs

Haonan Zeng¹, Kaixuan Guo¹, Zhanming Zhong¹, Jinyan Teng¹, Zhiting Xu¹, Chen Wei¹, Shaolei Shi¹, Zhe Zhang¹, Yahui Gao^{1,*}

¹State Key Laboratory of Swine and Poultry Breeding Industry, National Engineering Research Center for Breeding Swine Industry, Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

Haonan Zeng, E-mail: hnzeric@hotmail.com

*Correspondence Yahui Gao, E-mail: yahui.gao@scau.edu.cn

Abstract

Genotype imputation is essential for increasing marker density and maximizing the utility of existing SNP array data in animal breeding. Although a wide range of software is available for genotype imputation, a comprehensive benchmark in pigs is still lacking. In this study, we benchmarked 24 combinations of genotype imputation software for SNP arrays in pigs, comprising six independent pre-phasing software (fastPHASE, MaCH, BIMBAM, Eagle, SHAPEIT, Beagle) and four distinct imputation software (pbwt, Minimac, IMPUTE, Beagle), using 1,602 whole-genome sequencing (WGS) pigs from a multibreed pig genomics reference panel (PGRP) in PigGTEx. Our results indicated that the combination of Beagle for pre-phasing and Minimac for imputation achieves the highest imputation accuracy with a concordance of 0.983, especially for low-frequency SNPs (MAF<0.05). Finally, we proposed three recommended strategies: i) the combination of Beagle and Minimac is optimal for achieving the highest accuracy; ii) the combination of Beagle and Beagle is recognized for its convenience and relatively high accuracy despite it being memory-intensive; iii) the combination of Eagle and pbwt is feasible for its minimal computational cost with relatively high accuracy. This study provides valuable insights for implementing genotype imputation for pig SNP arrays toward sequence data and offers a basis for applications in livestock and poultry breeding.

Keywords: pig, genotype imputation, benchmark

Highlight

Evaluated 24 combinations of imputation software for pigs.

Beagle-Minimac combination provided the best imputation accuracy.

Beagle-Beagle combination stands out for convenience.

Eagle-pbwt combination showed excellent performance in resource efficiency.

1. Introduction

High-quality genomic data is crucial for genetic statistical analyses, including genome-wide association studies (GWAS) (Abdellaoui *et al.* 2023), human genomic risk prediction (Han *et al.* 2023), and genomic prediction in animals (Teng *et al.* 2022; Wang *et al.* 2024) and plants (Desta *et al.* 2014; Liu *et al.* 2024). Accurate and high-coverage genomic data such as whole genome sequencing (WGS) can boost the power of GWAS and contribute to precise candidate gene mapping (Cai *et al.* 2022; Fang *et al.* 2019; Sun *et al.* 2023). However, due to the high cost of WGS, single nucleotide polymorphism (SNP) arrays containing a limited number of SNPs are still widely used in animal breeding (Zhang *et al.* 2023), which reduces the efficiency of identifying the trait-associated genomic variants and genes in GWAS and the accuracy of genomic prediction.

Genotype imputation, an approach that increases marker density by accurately inferring ungenotyped

variants based on linkage disequilibrium (LD) using a high-density reference panel, is the key to maximizing the utility of existing SNP array data in animal breeding (Li *et al.* 2022). Currently, most genotype imputation models are based on the Li and Stephens Hidden Markov Model (HMM) (Li *et al.* 2003) described in 2003. This model assumes that each study haplotype is an imperfect mosaic of haplotypes in the reference panel, allowing for calculating the probabilities of each haplotype in the reference panel for each study variant. Although some new methods based on deep learning have recently been developed (Chen *et al.* 2019; Song *et al.* 2022; Kojima *et al.* 2020), the Li and Stephens HMM remains the widely acknowledged model for genotype imputation due to its advantage in accuracy, stability, and computational efficiency (De Marino *et al.* 2022; Naito *et al.* 2024). The process of genotype imputation consists of two main steps: pre-phasing, and imputation. Pre-phasing is the first step, where the phase or arrangement of alleles on chromosomes is inferred from the observed genotype data. This step involves estimating the haplotypes for each individual, and identifying which alleles are inherited together. Pre-phasing simplifies the subsequent imputation step by reducing computational complexity and enhancing accuracy, as the haplotypes are already inferred, allowing for more precise prediction of missing genetic variants in the imputation step. Imputation is the second step in the process, where missing or untyped genetic variants in a study sample are predicted using a reference panel of known haplotypes. This step involves comparing the phased haplotypes from the study sample to the reference panel to infer the most likely genotypes for the missing data. Imputation enhances the density and accuracy of genetic data without additional costly sequencing.

Since the genotype imputation method was first proposed, researchers worldwide have focused on the factors affecting the imputation accuracy (Ye *et al.* 2019; Zhang *et al.* 2022) and proposed many algorithms and platforms (Das *et al.* 2016; McCarthy *et al.* 2016; Zhang *et al.* 2024). At present, there is an abundance of software available for this purpose. Differences exist among the various software (De Marino *et al.* 2022; Ye *et al.* 2019), making the selection of the software combination strategy for imputation crucial for researchers conducting post-analysis. Recent benchmarks of pre-phasing and imputation software have been conducted on human (De Marino *et al.* 2022), cattle (Teng *et al.* 2022), and tilapia (Ye *et al.* 2024). However, a comprehensive benchmark of software in pigs is still lacking, resulting in confusion in pig genomic research. A comprehensive evaluation of imputation software in pigs is crucial for improving genotype accuracy, particularly in low-frequency variants, and enhancing genomic prediction for breeding programs. It helps optimize computational efficiency, reduce costs by maximizing the use of SNP arrays, and ensure more reliable trait mapping, supporting precision breeding strategies.

In this study, we employed a comprehensive benchmark of six haplotype pre-phasing software (Beagle, SHAPEIT, Eagle, MaCH, fastPHASE, and BIMBAM) and four imputation software (Beagle, pbwt, IMPUTE, and Minimac) in pairwise combinations (24 combinations in total) for SNP array genotypes in pigs. We assessed the performance of genotype imputation by measuring imputation accuracy and computational memory size and runtime. Our goal is to determine the optimal combination of pre-phasing and imputation software for pig SNP arrays in different scenarios.

2. Materials and methods

2.1. Data collection and pre-processing

We collected 1,602 pigs from the PigGTEx project (a multibreed pig genomics reference panel, PGRP) (Teng *et al.* 2024) with WGS data covering 42,523,218 SNPs. To better assess imputation accuracy, we selected 80 Yorkshire pigs in PGRP for genotype masking (taken as a golden standard) by maximizing the expected genetic relationship (REL) proposed by Druet and Hayes (Druet *et al.* 2014), retaining only overlapped variants with the GeneSeek Genomic Profiler (GGP) Porcine SNP50 BeadChip (46,494 SNPs) and classified these 80 Yorkshire pigs as a part of the target panel, while the remaining 1,522 pigs were kept as the reference panel.

To reduce computational complexity, we extracted the longest Chromosome 1 for the subsequent analysis.

The reference panel contains 4,368,645 SNPs on chromosome 1, while the target panel contains 4,753 SNPs. Coordinates for all SNPs are based on the Sus Scrofa 11.1 reference genome. Additionally, we conducted principal component analysis (PCA) to depict population structure using GCTA (Yang *et al.* 2011) with the parameter of ‘--pca 5’.

2.2. Genotype pre-phasing and imputation

We used six independent software for genotype pre-phasing and four software for imputation (**Table 1**), resulting in a total of 24 software combinations. To meet the requirements of different software for input files, we converted the reference panel files into the recommended formats corresponding to each software and used the default parameters for the analysis.

Table 1 Summary of pre-phasing and imputation software

Software	Function	Version	Released Year	Publication	Format of reference panel (suffix)
fastPHASE	Pre-phasing	v1.4.8	2006	(Scheet <i>et al.</i> 2006)	.recode.phase.inp
MaCH	Pre-phasing	v1.0.18.c	2007	(Li <i>et al.</i> 2009)	.haps/.snps
BIMBAM	Pre-phasing	v1.0	2010	(Servin <i>et al.</i> 2007)	recode.geno.txt/.recode.phe no.txt/.recode.pos.txt
pbwt	Imputation	v3.1	2014	(Durbin 2014)	.pbwt/.samples/.sites
Eagle	Pre-phasing	v2.4.1	2018	(Loh <i>et al.</i> 2016)	.bcf
Minimac	Imputation	v1.0.2	2019	(McCarthy <i>et al.</i> 2016)	.m3vcf.gz
IMPUTE	Imputation	v1.1.5	2020	(Rubinacci <i>et al.</i> 2020)	.imp5
SHAPEIT	Pre-phasing	v4.2.0	2021	(Delaneau <i>et al.</i> 2019)	.vcf.gz
Beagle	Pre-phasing & Imputation	v5.4	2022	(Browning <i>et al.</i> 2021, 2018)	.bref3

2.3. Imputation accuracy assessment

To evaluate the imputation accuracy, we extracted imputed genotypes of 4,364,610 masked SNPs from the 80 golden standard Yorkshires for accuracy calculation. We used three statistics of squared correlation (r^2), imputation quality score (IQS), and concordance rate (CR) to assess the imputation accuracy. Detailed calculation methods for each statistic can be found in **Appendix A**. Briefly, r^2 is the squared correlation between the expected additive genotype dosages and the known true additive genotype dosages according to the effect allele (i.e., 0, 1, 2). The CR is the concordance rate between the expected additive genotype dosages and the known true additive genotype dosages. The IQS is the adjusted CR controlling for allele frequencies by accounting for chance agreement (Lin *et al.* 2010). Three statistics were calculated for each SNP, and then the average value across SNPs was determined to represent the overall imputation accuracy for a single combination.

To investigate the imputation accuracy of different software combinations, we set the target panel size and the reference panel size to 400 and 1,000, respectively. Genotype imputation was performed using 24 different software combinations. The minor allele frequency (MAF) was calculated based on the target panel. MAF was divided into 13 groups (i.e., 0, (0,0.01], (0.01,0.02], (0.02,0.03], (0.03,0.04], (0.04,0.05], (0.05,0.07], (0.07,0.10], (0.10,0.15], (0.15,0.20], (0.20,0.30], (0.30,0.40], (0.40,0.50]), while the group of MAF equal 0 is not used for calculating the r^2 .

2.4. Computational runtime and memory record

The computational environment is supported by the National Supercomputer Center in Guangzhou, China, which is Red Hat Enterprise Linux Server release 7.3 (Maipo), the core processor of Intel Xeon E5-2692 v2

12*2, and a memory size of 128Gb.

We ran each of the software using a single thread and recorded the runtime, the maximum memory size, and the average processor usage. Accounting for the difference that exists in processor usage among software though setting single thread, runtime was adjusted by the average number of processor threads using the formula: $t \times cpu_{average}$.

3. Results

3.1. Overview of study design

The pipeline of this study was shown in **Appendix B**. In brief, we selected 80 Yorkshire pigs as golden standard pigs from 1,602 WGS pigs (PGRP) in the PigGTEx project (Teng *et al.* 2024), to form a subset of the target panel, showing a robust and representative population structure of Yorkshire in both the target and reference panel (**Appendix C**). The remaining 1,522 individuals were used as a reference panel containing 42,523,218 SNPs. Simultaneously, to ensure robust population statistics (e.g., MAF), we genotyped 320 additional Yorkshire pigs using the GGP Porcine SNP50 BeadChip (46,494 SNPs) and combined this data with that of the previously selected 80 WGS Yorkshire pigs, creating a total panel of 400 pigs. We masked the loci in the 400 pigs panel that were not present on the GGP Porcine SNP50 BeadChip to form a target panel. We then performed genotype pre-phasing and imputation using 24 software combinations. After imputation, we extracted the masked SNPs of 80 Yorkshire pigs in the target panel for accuracy assessment.

3.2. Imputation accuracy

To evaluate the accuracy of 24 software combinations, we calculated the overall imputation accuracy using three statistics (**Figure 1a-c**). The result revealed that the imputation accuracy of combinations including BIMBAM, fastPHASE, and MaCH for pre-phasing (mean CR=0.940, mean r^2 =0.499, mean IQS=0.405) was much lower than those including Beagle, Eagle, and SHAPEIT (mean CR=0.981, mean r^2 =0.814, mean IQS=0.751). The combinations involving BIMBAM for pre-phasing showed the lowest accuracy. The combinations containing Beagle, Eagle, and SHAPEIT for pre-phasing demonstrated a high level of accuracy, and Beagle exhibited the highest accuracy. In most cases, combinations including Minimac for imputation had the highest accuracy (mean CR=0.963, mean r^2 =0.674, mean IQS=0.606).

To explore whether imputation accuracy was affected by the size of the reference panel, we imputed the target panel using four imputation software based on nine reference panels of different sizes after pre-phasing performed using Beagle (**Figure 1d**). The results showed that a larger reference panel size leads to higher imputation accuracy.

To fully understand the pattern of imputation accuracy of 24 combinations of SNPs at different MAFs, we calculated the imputation accuracy according to different MAF groups (**Figure 1e-g**). The results showed that the statistics of r^2 and IQS were positively correlated with MAF (**Figure 1e,f**) while CR was the opposite (**Figure 1g**), which was in line with those of other research (Teng *et al.* 2022; Ding *et al.* 2023). In the result of r^2 and IQS, combinations containing Beagle, Eagle, and SHPAEIT for pre-phasing showed a higher level of accuracy compared with the combinations involving the other three pre-phasing software. Beagle for pre-phasing had the highest accuracy in low-frequency SNPs (MAF<0.05), and the Minimac for imputation displayed the highest accuracy in most combinations. It was noted that the accuracy exhibited a rapid upward trend when MAF<0.05, and a slight increase appeared when MAF≥0.05.

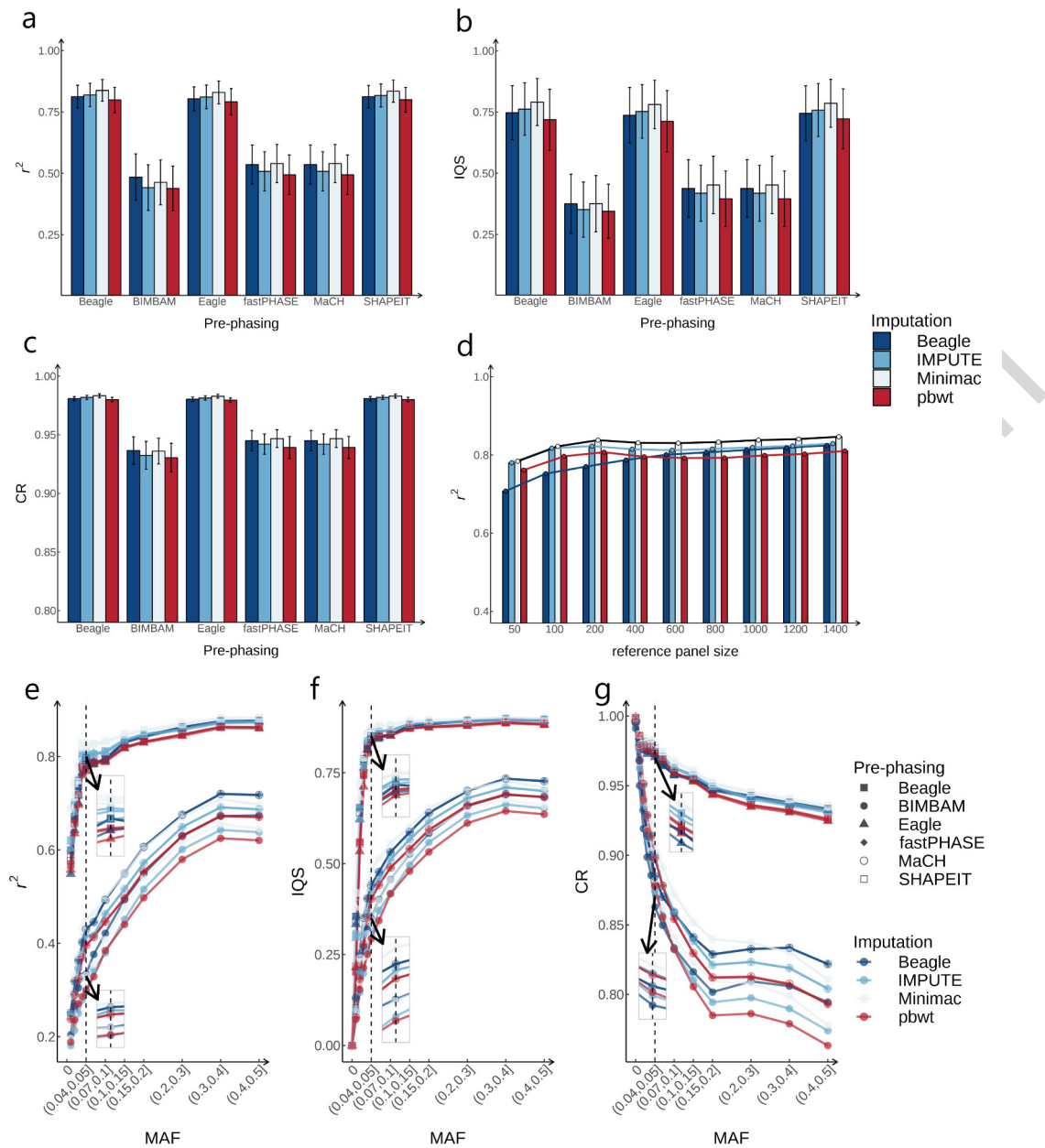


Figure 1 Imputation accuracy of 24 software combinations. a-c) The mean imputation accuracy of r^2 , IQS, and CR, respectively. The error bar in each column represents variance. d) The pattern of the impact of reference sample size on imputation accuracy. The line in four colors represents the corresponding trend of imputation software. e-g) Imputation accuracy of r^2 , IQS, CR, respectively. Dashed lines represent the value of MAF equal to 0.05, and sub figures in grids represent the detail of imputation accuracy in the value of MAF equal to 0.05. MAF was divided into 13 groups but showed 8 groups on the x-axis.

3.3. Computational cost

To explore the computational performance, we recorded the runtime and the maximum memory size for different sample sizes of the target and reference panels (**Figure 2**). The memory requirements and runtime of all the software increase with the sample size of either the target panel or reference panel. For the step of pre-phasing, among the three low-accuracy software, BIMBAM had the most memory requirement, fastPHASE had the longest runtime, and fastPHASE exhibited the greatest resource-efficiency. Among the other three high-accuracy software, SHAPEIT had the greatest computational efficiency and resource

efficiency, while Beagle showed the most memory requirement. In terms of the step of imputation, pbwt was the most resource-efficient with the least runtime and memory requirement. In contrast, Beagle had the highest computational demands, extremely in memory requirement.

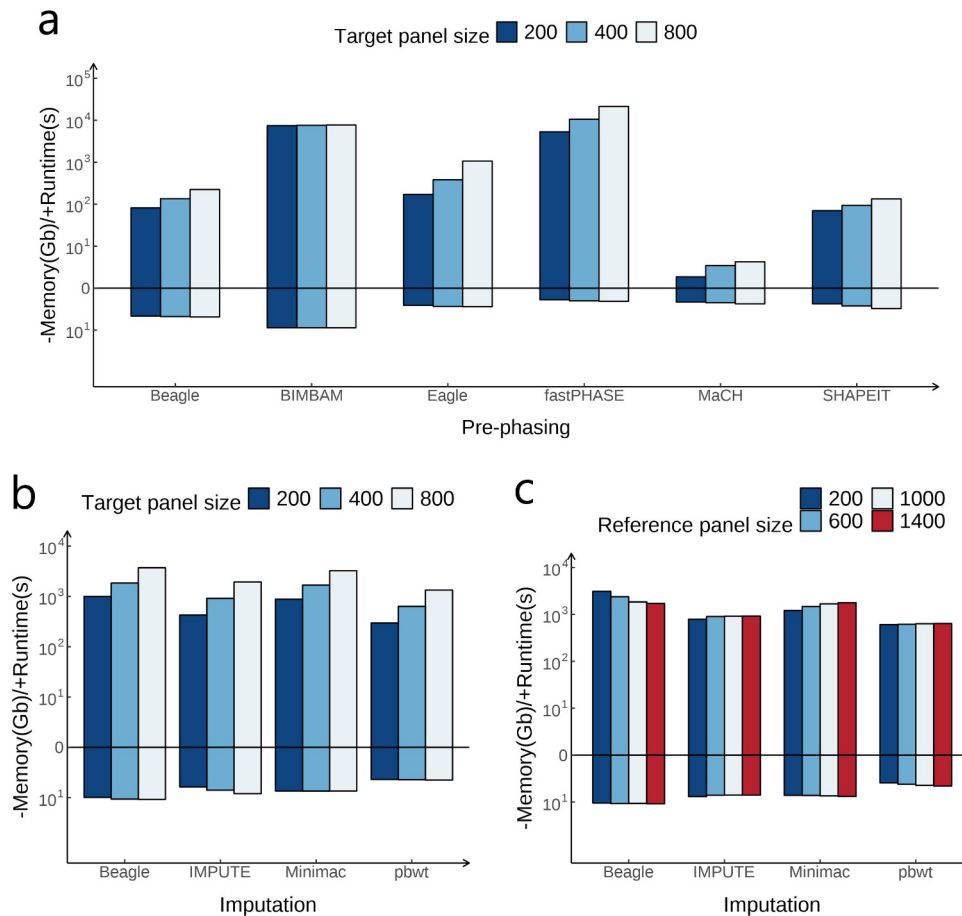


Figure 2 Computational performance of all software. The upper y-axis is the runtime (s), and the lower of y-axis is the maximum memory size (Gb). a) The computational performance of pre-phasing software versus target panel size. b) The computational performance of Imputation software versus target panel size. c) The computational performance of Imputation software versus reference panel size.

4. Discussion

We presented the most comprehensive benchmark for SNP array genotype imputation in pigs, employing a total of 24 combinations of six pre-phasing and four imputation software. In this study, all combinations achieved a high level of imputation accuracy, while commonly low accuracy for low-frequency SNPs ($MAF < 0.01$), which was in line with the pattern of genotype imputation (Browning *et al.* 2016; Rubinacci *et al.* 2021; Ye *et al.* 2019). Additionally, significant differences among the combinations made them suitable for diverse scenarios.

For pre-phasing, Beagle, Eagle, and SHAPEIT demonstrated superior imputation performance (accuracy and computational performance) compared to BIMBAM, fastPHASE, and MaCH. This may be attributed to the fact that BIMBAM, fastPHASE, and MaCH were three software developed at an early time and have not been updated in recent years. The latest update dates for BIMBAM, fastPHASE, and MaCH are 2010, 2008, and 2011, respectively. As a result, their algorithms may not be optimally adapted to the current genotype data, which involves larger sample sizes, higher call rates, and the use of the PLINK format. Among the pre-phasing software, Beagle exhibited the highest accuracy, particularly for low-frequency SNPs. However,

it also required the most memory size.

In terms of imputation, four software could reach high imputation accuracy, among which Minimac showed the highest. Although pbwt showed the lowest accuracy, it also exhibited the best computational efficiency with the least runtime and the best resource efficiency with the least memory size, that was suitable for the limited computational environment. Beagle was widely considered the most convenient and fast software for genotype imputation, effectively utilizing multiple threads and memory, but it also needed the most memory size, so it was suitable for achieving the best computing efficiency in a sufficiently powerful computing environment.

5. Conclusion

In this study, we comprehensively benchmarked 24 combinations of genotype pre-phasing and imputation software for SNP arrays in pigs, comprising six independent pre-phasing software and four distinct imputation software. Our results indicated that fastPHASE, BIMBAM, and MaCH were not suitable for our data due to their low accuracy. Beagle and Minimac were the highest accuracy software in pre-phasing and imputation, respectively. In terms of computational efficiency, pbwt for imputation proved the most resource-efficient with the least runtime and memory size. Based on this study, we concluded that 1) Using Beagle for pre-phasing and Minimac for imputation is the optimal software combination for the highest imputation accuracy. 2) Using Beagle for pre-phasing and Beagle for imputation is the optimal software combination for prioritizing convenience in a robust computational environment. 3) Using Eagle for pre-phasing and pbwt for imputation is the optimal software combination for pursuing the minimum computational requirement.

Future studies should explore innovative software that integrates cutting-edge algorithms for enhanced imputation accuracy. Moreover, analyzing a broader array of datasets, particularly from different pig breeds or genetic lines, will help assess the generalizability of current methods and identify potential improvements tailored to specific populations or breeding goals.

Ethical approval

Not applicable since no biological samples were collected and no animal handling was performed for this study. All data was collected from existing databases provided by South China Agricultural University (Guangzhou 510642, China).

Declaration of competing interest

The authors declare that they have no conflict of interest.

Acknowledgments

This study was supported by the National Key R&D Program of China (2022YFF1000900), the earmarked fund for China Agriculture Research System (CARS-35), Guangxi Science and Technology Program Project (GuikeJB23023003), the Local Innovative and Research Teams Project of Guangdong Province (2019BT02N630), Dedicated Funds for the Construction of Key Disciplines in Targeted Universities (2023B10564001), and the Young Scientists Fund of the National Natural Science Foundation of China (32402714). We thank National Supercomputer Center in Guangzhou China for its support in providing computing resources.

References

Abdellaoui A, Yengo L, Verweij K J H, Visscher P M. 2023. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics*, **110**, 179–194.

- Han J, van Hylckama Vlieg A, Rosendaal F R. 2023. Genomic science of risk prediction for venous thromboembolic disease: convenient clarification or compounding complexity. *Journal of Thrombosis and Haemostasis*, **21**, 3292–3303.
- Teng J, Ye S, Gao N, Chen Z, Diao S, Li X, Yuan X, Zhang H, Li J, Zhang X, Zhang Z. 2022. Incorporating genomic annotation into single-step genomic prediction with imputed whole-genome sequence data. *Journal of Integrative Agriculture*, **21**, 1126–1136.
- Wang Z, Li W, Tang Z. 2024. Enhancing the genomic prediction accuracy of swine agricultural economic traits using an expanded one-hot encoding in CNN models¹. *Journal of Integrative Agriculture*,.
- Desta Z A, Ortiz R. 2014. Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, **19**, 592–601.
- Liu P, Ma L, Jian S, He Y, Yuan G, Ge F, Chen Z, Zou C, Pan G, Lübberstedt T, Shen Y. 2024. Population genomic analysis reveals key genetic variations and the driving force for embryonic callus induction capability in maize. *Journal of Integrative Agriculture*, **23**, 2178–2195.
- Cai Z, Christensen O F, Lund M S, Ostensen T, Sahana G. 2022. Large-scale association study on daily weight gain in pigs reveals overlap of genetic factors for growth in humans. *BMC Genomics*, **23**, 133.
- Fang L, Jiang J, Li B, Zhou Y, Freebern E, Vanraden P M, Cole J B, Liu G E, Ma L. 2019. Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. *Communications Biology*, **2**, 100.
- Sun Y, Li Y, Zhao C, Teng J, Wang Y, Wang T, Shi X, Liu Z, Li H, Wang J, Wang W, Ning C, Wang C, Zhang Q. 2023. Genome-wide association study for numbers of vertebrae in Dezhou donkey population reveals new candidate genes. *Journal of Integrative Agriculture*, **22**, 3159–3169.
- Zhang Z, Xing S, Qiu A, Zhang N, Wang W, Qian C, Zhang J, Wang C, Zhang Q, Ding X. 2023. The development of a porcine 50K SNP panel using genotyping by target sequencing and its application¹. *Journal of Integrative Agriculture*,.
- Li Y, Bai X, Liu X, Wang W, Li Z, Wang N, Xiao F, Gao H, Guo H, Li H, Wang S. 2022. Integration of genome-wide association study and selection signatures reveals genetic determinants for skeletal muscle production traits in an F2 chicken population. *Journal of Integrative Agriculture*, **21**, 2065–2075.
- Li N, Stephens M. 2003. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, **165**, 2213–2233.
- Chen J, Shi X. 2019. Sparse Convolutional Denoising Autoencoders for Genotype Imputation. *Genes*, **10**, 652.
- Song M, Greenbaum J, Luttrell J I, Zhou W, Wu C, Luo Z, Qiu C, Zhao L J, Su K-J, Tian Q, Shen H, Hong H, Gong P, Shi X, Deng H-W, Zhang C. 2022. An autoencoder-based deep learning method for genotype imputation. *Frontiers in Artificial Intelligence*, **5**,.
- Kojima K, Tadaka S, Katsuoka F, Tamiya G, Yamamoto M, Kinoshita K. 2020. A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. *PLOS Computational Biology*, **16**, e1008207.
- De Marino A, Mahmoud A A, Bose M, Bircan K O, Terpolovsky A, Bamunusinghe V, Bohn S, Khan U, Novkovic B, Yazdi P G. 2022. A comparative analysis of current phasing and imputation software. *PLOS One*, **17**, e0260177.
- Naito T, Okada Y. 2024. Genotype imputation methods for whole and complex genomic regions utilizing deep learning technology. *Journal of Human Genetics*, 1–6.
- Ye S, Yuan X, Huang S, Zhang H, Chen Z, Li J, Zhang X, Zhang Z. 2019. Comparison of genotype imputation strategies using a combined reference panel for chicken population. *Animal*, **13**, 1119–1126.
- Zhang K, Peng X, Zhang S, Zhan H, Lu J, Xie S, Zhao S, Li X, Ma Y. 2022. A comprehensive evaluation of factors affecting the accuracy of pig genotype imputation using a single or multi-breed reference

- population. *Journal of Integrative Agriculture*, **21**, 486–495.
- Das S, Forer L, Schönherr S, Sidore C, Locke A E, Kwong A, Vrieze S I, Chew E Y, Levy S, McGue M, Schlessinger D, Stambolian D, Loh P-R, Iacono W G, Swaroop A, Scott L J, Cucca F, Kronenberg F, Boehnke M, Abecasis G R, *et al.* 2016. Next-generation genotype imputation service and methods. *Nature Genetics*, **48**, 1284–1287.
- McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood A R, Teumer A, Kang H M, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott L J, Zhang H, Mahajan A, Veldink J, *et al.* 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, **48**, 1279–1283.
- Zhang K, Liang J, Fu Y, Chu J, Fu L, Wang Y, Li W, Zhou Y, Li J, Yin X, Wang H, Liu X, Mou C, Wang C, Wang H, Dong X, Yan D, Yu M, Zhao S, Li X, *et al.* 2024. AGIDB: a versatile database for genotype imputation and variant decoding across species. *Nucleic Acids Research*, **52**, D835–D849.
- Teng J, Zhao C H, Wang D, Chen Z, Tang H, Li J B, Mei C, Yang Z P, Ning C, Zhang Q. 2022. Assessment of the performance of different imputation methods for low-coverage sequencing in Holstein cattle. *Journal of Dairy Science*, **105**, 3355–3366.
- Ye S, Zhou X, Lai Z, Ikhwanuddin M, Ma H. 2024. Systematic comparison of genotype imputation strategies in aquaculture: A case study in Nile tilapia (*Oreochromis niloticus*) populations. *Aquaculture*, **592**, 741175.
- Teng J, Gao Y, Yin H, Bai Z, Liu S, Zeng H, Bai L, Cai Z, Zhao B, Li X, Xu Z, Lin Q, Pan Z, Yang W, Yu X, Guan D, Hou Y, Keel B N, Rohrer G A, Lindholm-Perry A K, *et al.* 2024. A compendium of genetic regulatory effects across pig tissues. *Nature Genetics*, **56**, 112–123.
- Druet T, Macleod I M, Hayes B J. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity*, **112**, 39–47.
- Yang J A, Lee S H, Goddard M E, Visscher P M. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. *American Journal of Human Genetics*, **88**, 76–82.
- Scheet P, Stephens M. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *The American Journal of Human Genetics*, **78**, 629–644.
- Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype Imputation. *Annual Review of Genomics and Human Genetics*, **10**, 387–406.
- Servin B, Stephens M. 2007. Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLOS Genetics*, **3**, e114.
- Durbin R. 2014. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, **30**, 1266–1272.
- Loh P-R, Danecek P, Palamara P F, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis G R, Durbin R, L Price A. 2016. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, **48**, 1443–1448.
- Rubinacci S, Delaneau O, Marchini J. 2020. Genotype imputation using the Positional Burrows Wheeler Transform. *PLOS Genetics*, **16**, e1009049.
- Delaneau O, Zagury J-F, Robinson M R, Marchini J L, Dermitzakis E T. 2019. Accurate, scalable and integrative haplotype estimation. *Nature Communications*, **10**, 5436.
- Browning B L, Tian X, Zhou Y, Browning S R. 2021. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, **108**, 1880–1890.
- Browning B L, Zhou Y, Browning S R. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics*, **103**, 338–348.
- Lin P, Hartz S M, Zhang Z H, Saccone S F, Wang J, Tischfield J A, Edenberg H J, Kramer J R, Goate A M,

- Bierut L J, Rice J P, COGA C C C, GENEVA. 2010. A New Statistic to Evaluate Imputation Reliability. *PLOS One*, **5**, e9697.
- Ding R, Savegnago R, Liu J, Long N, Tan C, Cai G, Zhuang Z, Wu J, Yang M, Qiu Y, Ruan D, Quan J, Zheng E, Yang H, Li Z, Tan S, Bedhane M, Schnabel R, Steibel J, Gondro C, *et al.* 2023. The SWine IMputation (SWIM) haplotype reference panel enables nucleotide resolution genetic mapping in pigs. *Communications Biology*, **6**, 577.
- Browning B L, Browning S R. 2016. Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics*, **98**, 116–126.
- Rubinacci S, Ribeiro D M, Hofmeister R J, Delaneau O. 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, **53**, 120–126.

Advanced Publication

《中国畜牧兽医》编委会

主任委员：秦玉昌

委 员：（按英文字母/汉字笔画排序）

Julang Li, Canada

印遇龙 院士	任继周 院士	麦康森 院士	沈建忠 院士	张改平 院士		
陈化兰 院士	侯水生 院士	姚 斌 院士	夏咸柱 院士	譙仕彦 院士		
马月辉	王凤来	王文杰	王立贤	王加启	王楚端	文 杰
田见晖	朱 奎	朱化彬	朱鸿飞	刘国世	刘建新	李 奎
李发弟	李俊雅	李胜利	杨汉春	吕于明	张文广	张军民
张宏福	张家骅	陈继兰	范红结	罗会颖	罗海玲	周正奎
单安山	赵茹茜	高艳霞	咎林森	秦贵信	贾玉山	郭 勇
郭 鑫	唐湘方	曹兵海	曹果清	韩雪松	储明星	童光志
熊本海						

主 编：李 琍

副 主 编：卢庆萍 臧长江

编辑部主任：戴 晔

本期值班编辑：荣光辉

中 国 畜 牧 兽 医

（月刊，1974年创刊）

2024年8月5日 第51卷 第8期

China Animal Husbandry & Veterinary Medicine

(Monthly, Started in 1974)

Vol.51 No.8 August 5, 2024

主 管 中华人民共和国农业农村部

主 办 中国农业科学院北京畜牧兽医研究所

编辑出版 《中国畜牧兽医》编辑部

地 址 北京市海淀区圆明园西路2号

邮 编 100193

电 话 010-62816020 62811226 62810371

网 址 <http://www.chvm.net>

电子邮件 zgxsmy@caas.cn

印刷装订 北京科信印刷有限公司

国内发行 北京报刊发行局

订 购 处 全国各地邮电局

国外发行 中国国际图书贸易集团有限公司
（北京399信箱）

国内定价 50元/期 600元/年

Superintended by Ministry of Agriculture and Rural Affairs, P. R. China

Sponsored by Institute of Animal Sciences, Chinese Academy of
Agricultural Sciences (CAAS)

Published by Editorial Department of *China Animal Husbandry &
Veterinary Medicine*

Address No. 2 Yuanmingyuan West Road, Haidian, Beijing, China

Postcode 100193

Telephone 010-62816020 62811226 62810371

Website <http://www.chvm.net>

E-mail zgxsmy@caas.cn

Issued by Beijing Kexin Printing Co.,Ltd

Subscription Domestic Post Office

Issued Abroad by International Book Trading Corporation
P.O.Box 399, Beijing, P.R.China

ISSN 1671-7236 邮发代号：2-215 国外代号：BM3517
CN 11-4843/S 广告发布登记证号：京海工商广登字20170085号



中国科学引文数据库（CSCD）来源期刊
全国中文核心期刊 中国科技核心期刊
中国农林核心期刊 RCCSE中国核心学术期刊（A）

ISSN 1671-7236
CN 11- 4843/S

中国畜牧兽医

第51卷

第8期

2024年8月

中国畜牧兽医

CHINA ANIMAL HUSBANDRY & VETERINARY MEDICINE

第51卷
Vol. 51

第8期
No.8

2024年8月



主管： 中华人民共和国农业农村部
主办： 中国农业科学院北京畜牧兽医研究所

目次

· 生 物 技 术 ·

牛 MED28 基因编码区克隆、生物信息学分析及组织表达谱研究
..... 李晋男,王亚慧,邓天宇,梁 忙,杜丽丽,李柯安宁,薛青青,骞 里,高 雪,张路培,
朱 波,陈 燕,王泽昭,李俊雅,高会江(3225)

基于 CRISPR/Cas12a-RT-RAA 的猪繁殖与呼吸综合征病毒快速检测方法的建立
..... 于晶雪,韦珊珊,覃绍敏,吴健敏,杨丽华,陈凤莲,许力士,秦树英,
华 俊,韦 珏,方 芳,刘金凤(3237)

· 生 理 生 化 ·

番茄红素对顺铂所致大鼠脏器损伤的保护效果评价 孙 悦,田 雪,杨春雪,徐恩爽,郑家三(3247)

甜茶树苷 H 对高糖致小鼠肾足细胞损伤的作用研究
..... 陈 毓,孙伟翔,李昊达,张 婷,杨海峰,陈晓兰,李锋涛(3256)

· 营 养 与 饲 料 ·

诃子提取物对脂多糖刺激肉鸡肠道和肝脏损伤的影响 张晓涵,孙岚源,宋 转,侯永清,吴 涛(3267)

香芹酚对水貂生长性能、血清生化指标及肠道菌群的影响
..... 石华丽,陈嗣麒,云祥鸿,杨懿纯,张爱武(3279)

热应激对山羊粪便菌群结构及代谢产物的影响.....
..... 王常童,杨连弟,殷 丽,王 乐,张 进,左福元,黄文明(3288)

饲粮蛋白质水平对哺育期塔里木鸽种鸽生产性能、消化代谢和血浆生化指标的影响
..... 王 静,宋丽荣,刘 蓁,付 睿,王 孜,臧长江,李凤鸣(3301)

胍基乙酸对育肥猪生长性能、血清生化指标、抗氧化能力和免疫功能的影响.....
..... 李贞明,马现永,容 庭,崔艺燕,宋 敏,刘志昌,邓 盾,田志梅,余 苗(3311)

饲粮添加甲酸钙对老龄蛋鸡生产性能和鸡蛋品质的影响.....
..... 邱 凯,常心雨,高 珊,张海华,张海军,武书庚(3320)

不同营养水平饲粮对安黑杂交牛生长性能、养分消化和抗氧化性能的影响
..... 杜兰霞,杨瑞鑫,刘菲菲,郭艳丽,魏立明(3329)

地衣芽孢杆菌和丁酸梭菌对育肥羔羊生长、养分消化及血清生化指标的影响
..... 赵心念,李伯森,苏东遥,杨学颖,张新竹,纪海祥,张会文,高玉红(3337)

马铃薯干替代玉米对生长育肥猪生长性能、肉质性状和血常规指标的影响	颜 港,邢天琦,王钰滨,江山,张 帅,许 迪,张 坤,王梦影,梁 晶,兰干球,李 奎,黄三文(3345)
母马补喂支链氨基酸对哺乳马驹生长性能和粪便菌群的影响	刘笑天,贾雨欣,高 凤,袁鑫鑫,王永发,薛宇恒,孟 军,曾亚琦,李晓斌(3355)
短链脂肪酸对动物肠道健康和肠脑信号传递的影响研究进展	刘 森,任文义,徐晓锋,张力莉(3365)
膳食纤维在调节猪肠道健康与缓解肠道疾病中的作用及其机制	章 娜,李世强,杨凯丽,张 沙,班 博,方热军(3375)
湖羊对天蚕素抗菌肽的响应	冉 扬,邱 洁,申小云(3385)
牦牛肉品质研究进展	娄新建,马万浩,郝力壮,刘书杰,马世科,拜彬强(3394)
补喂酵母培养物对泌乳母马产奶量、乳成分及血液生化指标的影响	张积荣,张国庆,甫拉提江·艾力皮别克(3410)
· 遗 传 繁 育 ·	
藏猪、川乡黑猪妊娠期子宫体繁殖功能相关 lncRNA 鉴定和功能预测	王秋实,李江凌,赵素君,刘 锐(3417)
基于机器学习的地方鸡产蛋曲线拟合探索	郭 军,曲 亮,邵 丹,窦套存,王 强,李永峰,王星果,胡玉萍,童海兵(3428)
基于全基因组重测序检测中国地方猪的体型选择信号	马 烨,刘玉强,吴煜伟,李广祯,冯雪燕,刁淑琪,高亚辉(3438)
基于 GBS 测序技术分析云上黑山羊主配公羊亲缘关系及近交系数	李子健,江炎庭,兰 蓉,朱 兰,叶朗惠,龚 翔,欧阳依娜(3447)
鸡精液甘油简易冷冻保存技术改进与优化效果评价	刘伯承,刘 微,柳 颖,何晓娜,陈一峰,张光友,张明军,燕海峰(3461)
· 预 防 兽 医 ·	
CD163 基因敲除 iPAMs 的构建及其感染 PRRSV 的特征分析	董泽霞,林 鑫,周期律,王 楠,黄 雷,刘志国,冯 政,牟玉莲(3471)
牛诺如病毒感染相关腹泻犊牛粪便菌群的特征分析	赵清梅,崔省委,郭仕辉,余永涛,梁泰宇,李欢语(3484)
PEDV、TGEV 与 PDCoV S 蛋白表位基因三联疫苗的构建及其鉴定	刘 青,顾天越,包利霞,朱凡杰,王鑫源,刘婷婷,朱晓琛,鄢明华,董志民,王利丽,张东超,金天明(3495)
羊传染性脓疱病毒口唇和内脏感染株趋化因子结合蛋白基因差异分析及原核表达	李 蓉,富国文,曾 琴,吴 姣,袁嘉芮,刘亚波,四朗玉珍,樊月圆(3506)
o <i>ppCDF</i> 基因缺失对鼠伤寒沙门菌致病性的影响	尹景芬,张家莉,陈世雄,杨 婉,武绍碧,杨 琦(3519)

猪繁殖与呼吸综合征病毒诱导的炎症反应及抗炎药物研究进展·····	
····· 乔常宏,陈 晶,罗 琴,刘宝玲,何振文,刘丁语,陈翔宇,王晓虎,王 刚,白挨泉,蔡汝健(3528)	
1 株多重耐药大肠杆菌噬菌体 BP32 的分离鉴定及生物学特性分析 ·····	
····· 吴静楠,李占鸿,李富祥,张振兴,张以芳,宋建领(3540)	
· 基 础 兽 医 ·	
基于谱效关联分析和网络药理学研究连蒲双清散药效物质及作用靶点·····	
····· 齐小雨,刘星彤,孙一丹,候冉冉,李 秋,刘志海,刘聪敏(3552)	
甘肃省及周边地区畜禽屠宰及市售环节沙门菌的分离鉴定及生物学特性分析·····	
····· 马永辉,曹 青,范子秋,赵学慧,芝 吉,马金锐,何曾文,张浩浩,邓 静, 崇 倩,张坤中,宋维丽,苟惠天,薛惠文(3566)	
牛冠状病毒感染新生犊牛肺上皮细胞模型的建立·····	
····· 陈秋会,蒋珊珊,高萌萌,夏立勇,张国华,任亚超,周玉龙(3577)	
大肠杆菌肠菌素对空肠弯曲菌的促生长作用研究·····	
····· 时晨欣,崔一芳,申学阳,焦晓丽,郭芳芳,丁保安,徐福洲(3585)	
EGCG 调控 PI3K/Akt 通路对 CIS 诱导大鼠急性肾损伤的作用机制研究·····	
····· 杨春雪,孙 悦,田 雪,徐恩爽,郑家三(3595)	
河曲马源大肠杆菌耐药相关基因检测及外排泵抑制剂对其生物被膜的影响·····	
····· 赵 星,梁 军,李 阳,杨丹娇,陈朝喜(3603)	
1 株犬源多重耐药粪肠球菌的分离鉴定及其致病性分析 ··· 徐 飞,温贵兰,龚新勇,文 明,陈薄帆(3615)	
牛支原体感染牛巨噬细胞对线粒体途径介导凋亡的影响·····	
····· 唐 恬,王 振,余梦环,徐 坤,何瑞丽,秦田哲,陈创夫,马忠臣,王 勇(3625)	
过表达 miRNA-424-5p 靶向 AKT3 对奶牛子宫内膜上皮细胞凋亡的影响 ·····	
····· 姚伟佳,罗春海,刘佳金,王 薇,李丹阳,刘炳琦,付世新(3635)	
桑叶多糖对猪肺泡巨噬细胞的免疫调节作用 ····· 韩辰淼,王靖萱,杨海峰,陈晓兰,卜仕金(3643)	
桂郁金醇提物对脂多糖致小鼠肺损伤的保护作用研究·····	
····· 胡明霞,邓艳萍,何永芸,姚 玥,莫晓丹,黄 安,杨秀芬(3652)	
健脾祛湿方治疗非酒精性脂肪肝大鼠的作用研究 ····· 苏圆圆,于澄元,张密霞,张艳军,庄朋伟(3662)	
金黄色葡萄球菌和表皮葡萄球菌对奶牛中性粒细胞炎性因子分泌的影响·····	
····· 何兴丽,周佩瑶,张小雪,牟泉宙,李 杨,王昭元,刘 鹏,王 梓,宋杨阳,李晓琳,沈冰蕾(3676)	
· 环 境 安 全 ·	
冬季叠层笼养肉鸡舍内氨气、二氧化碳和空气颗粒物浓度分布特征 ·····	
····· 唐璐婵,王 俊,石志芳,李绚阳,席 磊(3687)	

• Biotechnology •

Cloning, bioinformatics analysis and tissue expression profile of coding region of bovine *MED28* gene
..... LI Jinnan, WANG Yahui, DENG Tianyu, LIANG Mang, DU Lili, LI Keanning, XUE Qingqing, QIAN Li,
GAO Xue, ZHANG Lupei, ZHU Bo, CHEN Yan, WANG Zezhao, LI Junya, GAO Huijiang(3225)

Establishment of a rapid detection method for Porcine reproductive and respiratory syndrome virus based on CRISPR/
Cas12a-RT-RAA YU Jingxue, WEI Shanshan, QIN Shaomin, WU Jianmin, YANG Lihua, CHEN Fenglian,
XU Lishi, QIN Shuying, HUA Jun, WEI Jue, FANG Fang, LIU Jinfeng(3237)

• Physiological and Biochemical •

Evaluation of the protective effect of lycopene on organ injury induced by cisplatin in rats
..... SUN Yue, TIAN Xue, YANG Chunxue, XU Enshuang, ZHENG Jiasan(3247)

Effect of cyclocarioside H on the damage of renal podocytes in mice stimulated by high glucose
..... CHEN Yu, SUN Weixiang, LI Haoda, ZHANG Ting, YANG Haifeng, CHEN Xiaolan, LI Fengtao(3256)

• Nutrition and Feed •

Effect of chebuli extract on intestinal and liver damage induced by lipopolysaccharide in broilers
..... ZHANG Xiaohan, SUN Lanyuan, SONG Zhuan, HOU Yongqing, WU Tao(3267)

Effect of carvacrol on growth performance, serum biochemical indexes and intestinal microbiota of minks
..... SHI Huali, CHEN Siqi, YUN Xianghong, YANG Yichun, ZHANG Aiwu(3279)

Effects of heat stress on the structure and metabolites of fecal flora of goats
..... WANG Changtong, YANG Liandi, YIN Li, WANG Le, ZHANG Jin, ZUO Fuyuan, HUANG Wenming(3288)

Effects of dietary protein levels on reproductive performance, digestion and metabolism and plasma biochemical indicators
of Tarim pigeons during breeding period
..... WANG Jing, SONG Lirong, LIU Zhen, FU Rui, WANG Zi, ZANG Changjiang, LI Fengming(3301)

Effects of guanidinoacetic acid on growth performance, serum biochemical indices, antioxidant capacity and immune
function of finishing pigs LI Zhenming, MA Xianyong, RONG Ting, CUI Yiyan, SONG Min,
LIU Zhichang, DENG Dun, TIAN Zhimei, YU Miao(3311)

Effects of dietary calcium formate supplementation on production performance and egg quality of elderly laying hens
..... QIU Kai, CHANG Xinyu, GAO Shan, ZHANG Haihua, ZHANG Haijun, WU Shugeng(3320)

Effects of diets with different nutrient levels on growth performance, nutrient digestion and antioxidant property of Anhei
crossbred cattle DU Lanxia, YANG Ruixin, LIU Feifei, GUO Yanli, WEI Liming(3329)

Effects of dietary *Bacillus licheniformis* and *Clostridium butyricum* on growth, digestion, and serum biochemical indexes
in fattening lambs ZHAO Xinnian, LI Bosen, SU Dongyao, YANG Xueying, ZHANG Xin Zhu,
JI Haixiang, ZHANG Huiwen, GAO Yuhong(3337)

Effects of substituting dried potatoes for corn on growth performance, meat quality traits and routine blood indexes of
growing-finishing pigs YAN Gang, XING Tianqi, WANG Yubin, JIANG Shan, ZHANG Shuai, XU Di,
ZHANG Kun, WANG Mengying, LIANG Jing, LAN Ganqiu, LI Kui, HUANG Sanwen(3345)

Effects of supplemental feeding of branched-chain amino acids to mares on growth performance and fecal flora of lactating foals LIU Xiaotian,JIA Yuxin,GAO Feng,YUAN Xinxin,WANG Yongfa,XUE Yuheng, MENG Jun,ZENG Yaqi,LI Xiaobin(3355)

Research progress on the effects of short-chain fatty acids on gut health and gut-brain signaling in animals LIU Miao,REN Wenyi,XU Xiaofeng,ZHANG Lili(3365)

The role and mechanism of dietary fiber in regulating porcine intestinal health and alleviating intestinal disorders ZHANG Na,LI Shiqiang,YANG Kaili,ZHANG Sha,BAN Bo,FANG Rejun(3375)

Response of fattening Hu sheep to cecropin antimicrobial peptides RAN Yang,QIU Jie,SHEN Xiaoyun(3385)

Progress of yak meat quality research LOU Xinjian,MA Wanhao,HAO Lizhuang,LIU Shujie,MA Shike,BAI Binqiang(3394)

The effect of supplementing yeast culture on milk production,milk composition and blood biochemical indicators in lactating mares ZHANG Jirong,ZHANG Guoqing,Bolatjan • Alipbek(3410)

• Genetics and Breeding •

Identification and function prediction of lncRNA related to reproductive function of uterine body in Tibetan and Chuanxiang Black pigs during pregnancy WANG Qiushi,LI Jiangling,ZHAO Sujun,LIU Rui(3417)

Exploration of egg production curve fitting of local chickens based on machine learning GUO Jun,QU Liang, SHAO Dan,DOU Taocun,WANG Qiang,LI Yongfeng,WANG Xingguo,HU Yuping,TONG Haibing(3428)

Detection of body shape selection signals in Chinese indigenous pigs based on whole genome resequencing MA Ye,LIU Yuqiang,WU Yuwei,LI Guangzhen,FENG Xueyan,DIAO Shuqi,GAO Yahui(3438)

Genetic relationship and inbreeding coefficients of Yunshang Black goats main mating rams using genotyping-by-sequencing LI Zijian,JIANG Yanting,LAN Rong,ZHU Lan,YE Langhui,GONG Xiang,OUYANG Yina(3447)

Evaluation of improvement and optimization for cryopreservation technology of chicken semen with glycerol LIU Bocheng,LIU Wei,LIU Ying,HE Xiaona,CHEN Yifeng, ZHANG Guangyou,ZHANG Mingjun,YAN Haifeng(3461)

• Preventive Veterinary Medicine •

Construction and PRRSV infection characteristic analysis of CD163 gene knockout iPAMs DONG Zexia,LIN Xin,ZHOU Qilyu,WANG Nan,HUANG Lei,LIU Zhiguo,FENG Zheng,MU Yulian(3471)

Characteristics of fecal microbiota in calves with diarrhea associated with Bovine norovirus infection ZHAO Qingmei,CUI Shengwei,GUO Shihui,YU Yongtao,LIANG Taiyu,LI Huanyu(3484)

Construction and identification of a triple vaccine of PEDV,TGEV and PDCoV S protein epitope gene LIU Qing,GU Tianyue,BAO Lixia,ZHU Fanjie,WANG Xinyuan,LIU Tingting,ZHU Xiaochen, YAN Minghua,DONG Zhimin,WANG Lili,ZHANG Dongchao,JIN Tianming(3495)

Analysis of the differences of chemokine-binding protein gene between Ovine contagious pustular dermatitis virus oral and visceral infection strains and their prokaryotic expression LI Rong,FU Guowen,ZENG Qin,WU Jiao,YUAN Jiarui,LIU Yabo,SILANG Yuzhen,FAN Yueyuan(3506)

Effects of *oppCDF* gene deletion on the pathogenicity of *Salmonella* Typhimurium YE Jingfen,ZHANG Jiali,CHEN Shixiong,YANG Wan,WU Shaobi,YANG Qi(3519)

Advances in inflammation induced by Porcine reproductive and respiratory syndrome virus and anti-inflammatory drugs QIAO Changhong,CHEN Jing,LUO Qin,LIU Baoling,HE Zhenwen,LIU Dingyu, CHEN Xiangyu,WANG Xiaohu,WANG Gang,BAI Aiquan,CAI Rujian(3528)

Isolation,identification and biological characterization analysis of phage BP32 from multidrug-resistant *Escherichia coli*
..... WU Jingnan,LI Zhanhong,LI Fuxiang,ZHANG Zhenxing,ZHANG Yifang,SONG Jianling(3540)

• Basic Veterinary Medicine •

Study on the pharmacodynamic substances and targets of Lianpu Shuangqing powder based on spectrum-effect relationship
analysis and network pharmacology
..... QI Xiaoyu,LIU Xingtong,SUN Yidan,HOU Ranran,LI Qiu,LIU Zhihai,LIU Congmin(3552)

Isolation,identification and biological characterization analysis of *Salmonella* from livestock and poultry slaughtering and
marketing in Gansu province and neighboring areas
..... MA Yonghui,CAO Qing,FAN Ziqiu,ZHAO Xuehui,ZHI Ji,MA Jinrui,HE Zengwen,ZHANG Haohao,
DENG Jing,CHONG Qian,ZHANG Kunzhong,SONG Weili,GOU Huitian,XUE Huiwen(3566)

Establishment of a neonatal calf lung epithelial cell model infected with Bovine coronavirus
... CHEN Qiuhui,JIANG Shanshan,GAO Mengmeng,XIA Liyong,ZHANG Guohua,REN Yachao,ZHOU Yulong(3577)

Effects of *Escherichia coli* enterobactin on growth promotion of *Campylobacter jejuni*
..... SHI Chenxin,CUI Yifang,SHEN Xueyang,JIAO Xiaoli,GUO Fangfang,DING Baoan,XU Fuzhou(3585)

Mechanism of EGCG regulating PI3K/Akt pathway on CIS-induced acute kidney injury in rats
..... YANG Chunxue,SUN Yue,TIAN Xue,XU Enshuang,ZHENG Jiasan(3595)

Detection of drug resistance genes and effect of efflux pump inhibitors on biofilm of *Escherichia coli* isolated from Hequ
horses ZHAO Xing,LIANG Jun,LI Yang,YANG Danjiao,CHEN Chaoxi(3603)

Isolation and identification of a multidrug-resistant strain of *Enterococcus faecalis* from dog and analysis of its pathogenicity
..... XU Fei,WEN Guilan,GONG Xinyong,WEN Ming,CHEN Bofan(3615)

Effect of *Mycoplasma bovis* infection on mitochondrial pathway mediated apoptosis in bovine macrophages
..... TANG Tian,WANG Zhen,YU Menghuan,XU Kun,HE Ruili,QIN Tianzhe,
CHEN Chuangfu,MA Zhongchen,WANG Yong(3625)

Effect of overexpression of miRNA-424-5p targeting AKT3 on apoptosis of endometrial epithelial cells in dairy cows
..... YAO Weijia,LUO Chunhai,LIU Jiajin,WANG Wei,LI Danyang,LIU Bingqi,FU Shixin(3635)

Immunomodulatory effects of mulberry leaf polysaccharide on porcine alveolar macrophages
..... HAN Chenmiao,WANG Jingxuan,YANG Haifeng,CHEN Xiaolan,BU Shijin(3643)

Protective effect of alcohol extract from *Curcuma kwangsiensis* on lung injury induced by LPS in mice
..... HU Mingxia,DENG Yanping,HE Yongyun,YAO Yue,MO Xiaodan,HUANG An,YANG Xiufen(3652)

Study on the efficacy of Jianpi Qushi decoction in the treatment of nonalcoholic fatty liver in rats
..... SU Yuanyuan,YU Chengyuan,ZHANG Mixia,ZHANG Yanjun,ZHUANG Pengwei(3662)

Effect of *Staphylococcus aureus* and *Staphylococcus epidermidis* on the secretion of inflammatory factors by neutrophils
in dairy cows HE Xingli,ZHOU Peiyao,ZHANG Xiaoxue,MOU Quanzhou,LI Yang,WANG Zhaoyuan,
LIU Peng,WANG Zi,SONG Yangyang,LI Xiaolin,SHEN Binglei(3676)

• Environmental Safety •

Distribution characteristics of ammonia,carbon dioxide and airborne particle concentrations in a broiler house with vertical
tiered cages in winter TANG Luchan,WANG Jun,SHI Zhifang,LI Xuanyang,XI Lei(3687)

基于全基因组重测序检测中国地方猪的 体型选择信号

马 烨,刘玉强,吴煜伟,李广祯,冯雪燕,刁淑琪,高亚辉[✉]

(国家生猪种业工程技术研究中心,华南农业大学动物科学学院,广东省农业动物基因组学及
分子育种重点实验室,广州 510642)

摘要:【目的】利用全基因组选择信号检测方法探索中国地方猪基因组中与体型相关的候选基因,分析中国地方猪在进化和驯化过程中的受选择情况。【方法】基于 310 头地方猪的全基因组重测序数据,利用跨群体扩展单倍型纯合(cross population extended haplotype homozygosity,XP-EHH)和固定分化指数 F_{ST} 统计量(fixation index)两种方法进行全基因组选择信号检测。取两种方法各前 0.1% 位点的交集作为候选位点,分别向上、下游延伸 200 kb 作为潜在受选择区域。通过富集分析进一步探索选择信号的生物学功能。【结果】使用 XP-EHH 和 F_{ST} 两种方法共检测到 633 个潜在受选择位点,获得 633 个潜在区域,共注释到 133 个基因。数量性状基因座(quantitative trait locus,QTL)富集分析发现,不同体型中国地方猪之间的选择信号与肉质和胴体性状相关;染色质状态富集结果发现,不同体型中国地方猪群体之间的差异组织为内脏、消化系统和大脑。GO 功能和 KEGG 通路富集分析发现 16 个基因在 6 个功能条目($P < 0.05$, count > 3)上显著富集,主要集中在生长代谢相关通路,其中 ACSM5、ACSM4、FXR1 和 FOXA2 作为候选基因主要富集于脂肪酸合成与代谢、肌肉生长发育等信号通路。【结论】本研究采用两种选择信号检测方法对不同体型的中国地方猪进行了分析,筛选到一系列候选位点和基因,重点挖掘了参与猪生长发育的候选基因,如 ACSM5、ACSM4、FXR1 和 FOXA2。

关键词: 中国地方猪;选择信号;全基因组重测序;跨群体扩展单倍型纯合(XP-EHH);固定分化指数(F_{ST})

中图分类号: S813.3

文献标识码: A

Doi: 10.16431/j.cnki.1671-7236.2024.08.022

开放科学(资源服务)标识码(OSID):



Detection of Body Shape Selection Signals in Chinese Indigenous Pigs Based on Whole Genome Resequencing

MA Ye, LIU Yuqiang, WU Yuwei, LI Guangzhen, FENG Xueyan, DIAO Shuqi, GAO Yahui[✉]

(National Engineering Research Center for Breeding Swine Industry, Guangdong

Provincial Key Laboratory of Agro-Animal Genomics and Molecular Breeding,

College of Animal Science, South China Agricultural University, Guangzhou 510642, China)

Abstract: 【Objective】 The purpose of this study was to investigate candidate genes associated with body size in the genomes of Chinese indigenous pigs using whole-genome selection signal detection methods, and analyze the selection patterns during their evolution and domestication of Chinese indigenous pigs. 【Method】 Based on whole-genome resequencing data of 310 indigenous pigs, two methodologies were employed: Cross population extended haplotype homozygosity (XP-EHH) and fixation index statistic (F_{ST}), a comprehensive genome-wide selection signal scan was conducted. The intersection of the top 0.1% sites identified by both methods was designated as candidate loci, and their upstream and downstream 200 kb regions were designated as potential

selection regions. Through enrichment analysis, the biological functions of selected signals were further explored. 【Result】 A total of 633 potential selection loci were detected by XP-EHH and F_{ST} methods, 633 potential regions were obtained, and 133 genes were annotated. The quantitative trait locus (QTL) enrichment analysis showed that the selection signals in different body types of Chinese indigenous pigs were related to meat and carcass traits. The results of chromatin state enrichment analysis showed that the tissues with differences in different body types of Chinese indigenous pigs were concentrated in viscera, digestive system and cerebellum. GO function and KEGG pathway enrichment analysis showed that 16 genes were significantly enriched in 6 functional items ($P < 0.05$, count > 3), mainly concentrated in growth and metabolism-related pathways. Among them, *ACSM5*, *ACSM4*, *FXR1* and *FOXA2* were candidate genes, which were mainly enriched in fatty acid synthesis and metabolism, muscle growth and development and other signaling pathways. 【Conclusion】 In this study, two detection methods were used to analyze Chinese indigenous pigs with different body sizes, and a series of genetic differentiation genes were screened. The candidate genes involved in the regulation of growth and development in pigs, such as *ACSM5*, *ACSM4*, *FXR1* and *FOXA2*, were excavated.

Key words: Chinese indigenous pigs; selective signals; whole-genome resequencing; XP-EHH; F_{ST}

猪作为最早被人类驯化的动物之一^[1],在生态、农业、医学研究中扮演着重要角色。约 1 万年前,野猪(*Sus scrofa*)主要分布在中国和近东^[2-3]。从约 9 000 年前首次被驯化以来,猪经历了复杂的进化和人工选择过程,形成了不同的品种品系,在繁殖性能、发情时间及妊娠期方面也有很大变化,且在外观、生长发育、肉质风味等多个方面均展现出了明显的多样性^[4-5]。中国有着丰富的猪品种遗传资源,全球有超过 1/3 的猪品种源于中国^[5]。其中,微型猪品种众多且品质优良,如香猪、藏猪、五指山猪和台湾的兰屿小耳猪,其成年猪体重在 40~45 kg,均低于国外的小型猪品种,如米尼苏达·荷曼小型猪、汉福特小型猪(成年猪体重在 70~90 kg)和高金根小型猪(成年猪体重约 53 kg)^[5]。猪的体型大小和生长速度是商业猪种选育的重点,从遗传上解析小型猪的形成机制并定位相关的突变位点,对猪种遗传差异的探索及品种的选育都具有重要的指导意义。

利用测序技术对畜禽基因组进行选择信号检测可以更好地揭示影响畜禽重要经济性状的候选功能基因。研究表明,猪体型大小受到多基因调控。Reimer 等^[6]利用小型猪和大型猪的重测序数据对猪体型变异进行分析,揭示了猪体型大小变化受多基因而非单基因或寡基因调控,并表明代谢功能的改变和胰岛素抵抗有助于猪的小型化;Kim 等^[7]通过研究尤卡坦小型猪体型的进化模型发现,尤卡坦小型猪在机体和细胞水平上均受到了选择;384 Simianer 等^[8]对哥廷根迷你猪的研究发现,该小型

猪品种的形成是一种垂体侏儒症的表现,而这种症状是由垂体腺生长激素分泌减少,进而导致胰岛素的分泌减少引起的。目前,大量与体型发育相关的基因和调控因子已被验证,包括 *HMGA2*^[9]、*BMP2*^[10]、*FGFR3*^[11]和 *IGF- I R*^[12]。

随着大量组学数据的产出,为多维深度解析选择信号的潜在分子生物学功能提供了便利。PigGTEx (Pig Genotype-Tissue Expression) 项目^[13]和 FAANG (Functional Annotation of Animal Genomes Project) 项目^[14]提供了大量的功能基因组注释信息及表达数量性状基因座 (expression quantitative trait locus, eQTL) 等分子表型信息,为探索选择信号在基因组及转录组对复杂表型的调控作用提供了更多层次的证据,将选择信号与复杂的性状紧密联系起来。同时,在猪大规模复杂性状遗传解析方面, PigBiobank (<http://pigbiobank.farmgtex.org/>) 数据库^[15]中包含了猪复杂性状的相关数据,为全球研究者提供了大数据支撑。

作为选育重点,猪体型大小对提高猪肉的生产效率和经济价值、满足市场和消费者的需求有重要意义。为了探索猪体型形成的遗传基础,国内外已有许多针对常见商品猪品种的不同体型性状的全基因组关联研究 (genome-wide association study, GWAS)^[16-18],但系统比较不同体型中国地方猪品种遗传差异的报道较少。鉴于此,本研究使用 PigGTEx 项目中的中国本土猪种的全基因组重测序数据及公共数据库的相关数据进行地方猪全基因

组选择信号检测,并结合大规模注释信息进行体型相关候选基因鉴定,以期解析中国猪体型发育的分子机理,为猪种的选育提供理论基础。

1 材料与方法

1.1 试验群体

本研究提取了猪全基因组参考面板 (Pig Genomics Reference Panel, PGRP)^[13] (v 1. 0) 中

23 个地方猪种共 310 头成年公猪的全基因组测序数据用于分析。同时,本研究统计了《中国畜禽遗传资源志·猪志》^[5] 所收录的研究用品种的体重、体长、胸围和体高 4 种生长性状平均数据,并使用标准化的表型数据计算了各品种间的欧氏距离。基于品种间欧氏距离将试验品种划分为大体型猪和小体型猪两组(图 1),其中大体型组包含 13 个品种共 192 个个体,小体型组包含 10 个品种共 118 个个体。

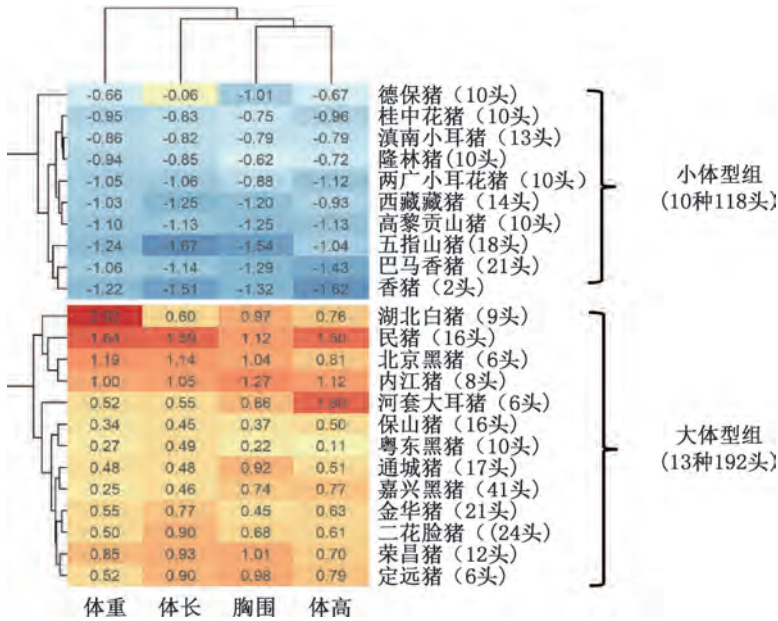


图 1 中国地方猪体重和体尺性状数据的聚类图

Fig. 1 Cluster diagrams for body weight and body size trait data of Chinese indigenous pigs

1.2 方法

1.2.1 数据下载与处理 本研究使用的测序数据来自于公共数据库与 PGRP 项目,使用 PLINK(v 1.90)软件^[19]对单核苷酸多态性(single nucleotide polymorphism, SNP)位点进行质控,质控条件为:①剔除次等位基因频率(minor allele frequency, MAF)<0.01 的位点,使用命令:“--maf 0.01”;②保留常染色体 SNP 位点,使用命令:“--autosome”;③保留检出率>0.95 的 SNP,使用命令:“--geno 0.05”。

1.2.2 主成分分析 本研究使用 GCTA(v 1.93.2)软件^[20]对试验群体进行主成分分析(principal component analysis, PCA),用于探究试验群体的分层情况和家系划分,并用 ggplot2 R 包(v 3.3.6)^[21]对结果进行可视化。

1.2.3 选择信号检测 本研究采用跨群体扩展单倍型纯合(cross population extended haplotype 385 points, F_{ST} 检测基于遗传变异的分布模式,通过比较同一基因座上不同群体的遗传差异计算 F_{ST} 值。 F_{ST}

计量(fixation index)^[23]两种选择信号检测方法进行全基因组选择信号检测,将同时被两种方法检测到的显著受选择位点定义为潜在显著受选择位点,并将其上、下游 200 kb 的区间定义为潜在受选择区间。

使用 Selscan(v 1.3.0)软件^[24]进行 XP-EHH 检验分析。XP-EHH 检测属于双尾检测,本研究将 0.1%分位点和 99.9%分位点设置为阈值线,当位点的 XP-EHH 分数低于 0.1%分位点时,说明该位点在群体 A 受到选择;当位点的 XP-EHH 分数高于 99.9%分位点时,说明该位点在群体 B 受到选择。其中,群体 A 代表大体型猪;群体 B 代表小体型猪。

使用 VCFtools(v 0.1.13)软件^[25]对比较组(大体型猪-小体型猪)进行基于单位点的 F_{ST} 统计量计算,将前 0.1%分位点的 SNP 作为显著受选择位点。 F_{ST} 检测基于遗传变异的分布模式,通过比较同一基因座上不同群体的遗传差异计算 F_{ST} 值。 F_{ST}

值通过计算群体间的遗传方差和总体遗传方差之间的比值获得,较高的 F_{ST} 值表示群体间的遗传差异较大,而较低的 F_{ST} 值则表示群体之间的遗传相似性较高。

1.2.4 富集分析 本研究将显著受选择位点分别与数量性状基因座 (quantitative trait locus, QTL) 区域和染色质状态进行富集分析。

通过 Animal QTLdb (Release 45) 数据库^[26] 将性状分为五大性状分级 (trait class), 又将性状分级具体细分为 20 种性状类型 (trait type)。根据 Animal QTLdb (Release 45) 数据库^[26] 中 QTL 信息, 本研究保留了显著 ($P<0.05$)、性状类型报道次数 >100 、长度 <1 Mb 的 QTL 与显著受选择位点进行富集分析。采用 region R 包^[27] 进行富集分析, 其中显著性阈值定义为 $P<0.05$ 。

将显著受选择位点与 FANNG 项目^[14] 产生的 14 类组织的 15 种染色质状态进行富集分析。15 种染色质状态主要划分为 6 类: ① 启动子 (TssA、TssAHet、TssBiv); ② 转录起始位置 (transcription start site, TSS) 近端转录区 (TxFlnk、TxFlnkWk、TxFlnkHet); ③ 增强子 (EnhA、EnhAMe、EnhAWk、EnhAHet、EnhPois); ④ ATAC 岛区域 (ATAC_Is); ⑤ 抑制区 (Repr、ReprWk); ⑥ 静态区 (Quiescent)。采用 LOLA 包^[28] 进行富集分析, 其

中显著性阈值定义为 $P<0.05$ 。

1.2.5 功能注释 根据 Ensembl 数据库^[29] 中猪基因组注释信息, 本研究基于基因物理位置与潜在受选择区域进行重叠检索, 将重叠基因定义为潜在候选基因。使用 DAVID (v 6.8) 数据库^[30] 对候选基因进行 GO 功能及 KEGG 通路富集分析, 保留 $P<0.05$ 和富集基因数 (count) >3 的条目。

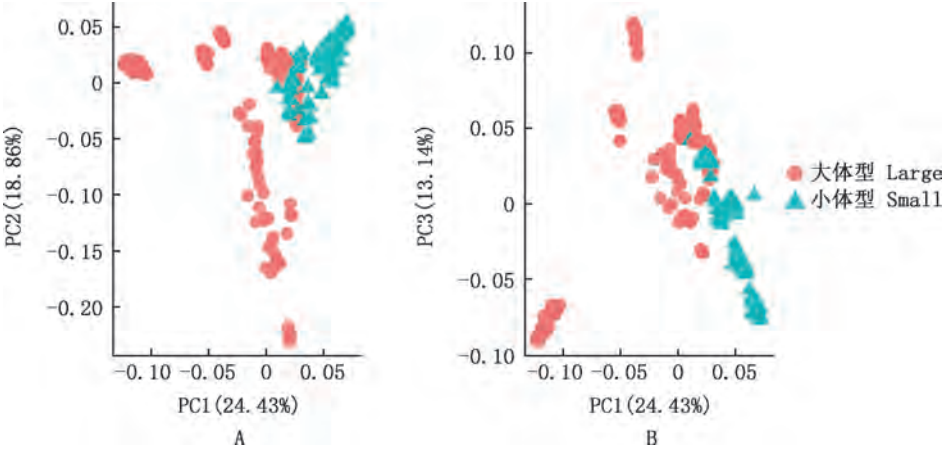
2 结 果

2.1 数据质控结果

在数据质控过程中, 下载数据包括 310 头中国地方猪与 42 523 218 个 SNPs 位点, 以 $MAF>0.01$ 、SNP 缺失率 <0.1 、样本的 SNP 缺失率 <0.1 和仅保留常染色体为筛选标准, 共有 10 366 323 个 SNPs 位点被删除, 剩余 310 个个体和 32 156 895 个 SNPs 位点用于后续分析。

2.2 PCA 结果

对 310 头中国地方猪进行 PCA 推断个体间的聚类模式, 结果见图 2。由图 2 可知, PC1、PC2 和 PC3 分别解释了 24.43%、18.86% 和 13.14% 的遗传变异。大体型地方猪的分布相对较广, 其分散程度相对小体型地方猪更大; 小体型地方猪主要分布于华南地区, 相对聚集。此外, PC1-PC3 均可大致区分不同分组的



A, PC1-PC2; B, PC1-PC3

图 2 中国地方猪群体主成分分析图

Fig. 2 PCA plot of Chinese indigenous pig population

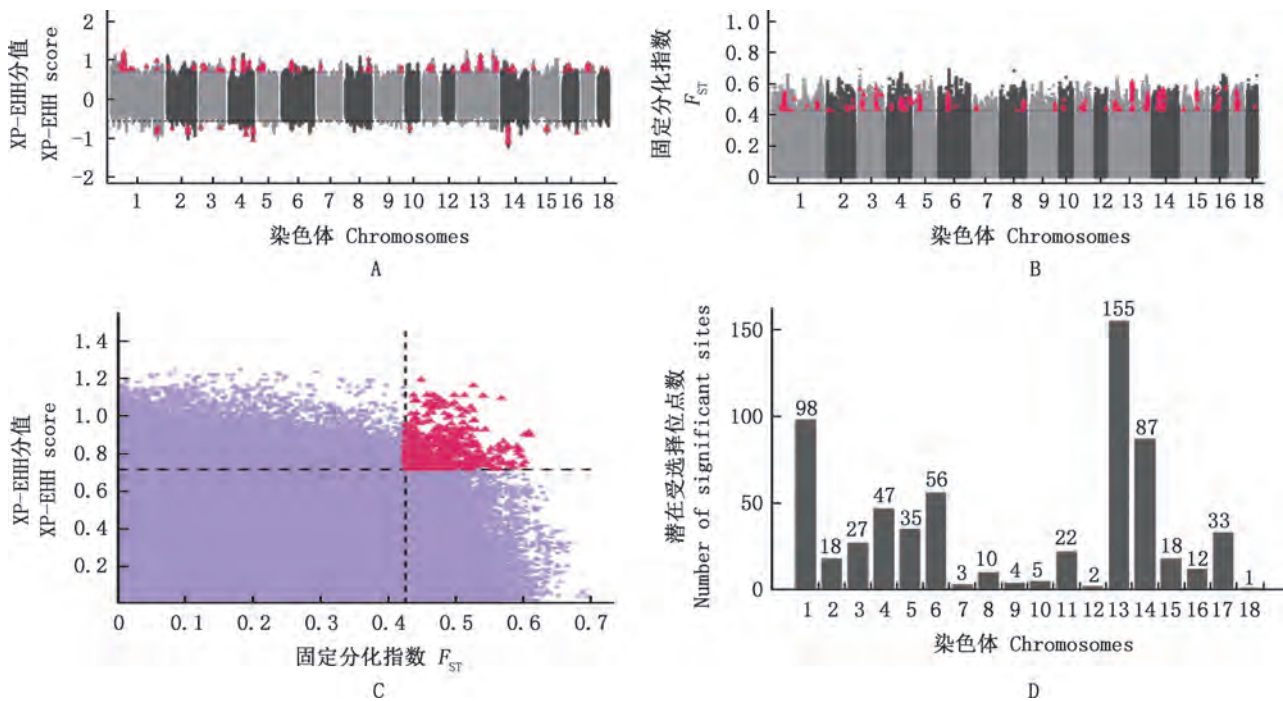
2.3 选择信号检测

本研究对大体型和小体型中国地方猪进行了 XP-EHH 和 F_{ST} 两种选择信号检测分析, 分别得到 32 602 和 32 576 个位点 (图 3A、3B), 且显著受选择位点在 1~18 号染色体上均有分布。全基因组范围

内, XP-EHH 方法检测到的最显著受选择位点位于 1 号染色体上, 其 XP-EHH 分数绝对值为 1.1957; 而 F_{ST} 方法检测到的最显著受选择位点位于 13 号染色体上, 其 F_{ST} 值为 0.6106。取两种方法检测结果的交集, 共获得 633 个潜在受选择位点 (图 3C),

在 18 条染色体上均有分布,其中 13 号染色体上分布了最多潜在受选择位点,占总显著受选择位点数

的 24.5%(图 3D)。



A,全基因组 XP-EHH 分析曼哈顿图;B,全基因组 F_{ST} 分析曼哈顿图;C,全基因组选择信号检测图(常染色体),虚线表示 99.9%置信区间,显著受选择位点用粉红色三角点表示;D,受选择位点分布图(常染色体)

A,Manhattan plot for genome-wide XP-EHH analysis;B,Manhattan plot for genome-wide F_{ST} analysis;C,Plot for selection signal detection in the whole genome (autosomes),the dashed line represents the 99.9% confidence interval,and significantly selected sites are indicated by pink triangles;D,Distribution map of selected sites (autosomes)

图 3 中国地方猪的全基因组选择信号检测图

Fig.3 Genome-wide selection signal detection map of Chinese indigenous pigs

2.4 QTL 区域富集分析

经筛选后本研究保留了 17 841 个 QTLs 用于和潜在受选择位点进行富集分析,从性状类别上看,大体型和小体型中国地方猪的潜在受选择位点多富集于肉质和胴体性状 QTL,主要包含 pH、脂肪酸含量、脂肪沉积、肉色纹理等性状的富集(图 4)。

2.5 染色质状态富集分析

将潜在受选择位点与 14 类组织的 15 种染色质状态进行富集分析,结果显示,潜在受选择位点在中国地方猪肝脏、下丘脑、皮质和小脑组织的 TssAHet 状态,以及脾脏、回肠、十二指肠和脂肪组织的 TxFlnk 状态显著富集(图 5)。TssAHet 和 TxFlnk 状态活性均较强,分别与启动子区域和基因

转录起始位点相关。

2.6 GO 功能与 KEGG 通路富集分析

将潜在受选择位点向上、下游各延伸 200 kb 的区间作为潜在受选择区间,用于基因注释,共注释到 133 个基因。对注释基因进行 GO 功能和 KEGG 通路富集分析,保留 $P<0.05$ 和富集基因数 >3 的条目,共筛选得到 6 条显著 GO 条目和 1 条 KEGG 通路(表 1),注释基因主要集中于不同的代谢过程。其中,3 个基因(ASCM4、ASCM5、ENSSSCG00000007858)同时富集于酰基-CoA 代谢过程、脂肪酸生物合成过程、脂肪酰-CoA 合成酶活性、脂肪酸连接酶活性的条目和丁酸代谢通路。对基因功能进行检索,其中 FOXA2 与 FXR1 基因和猪体型大小有关。

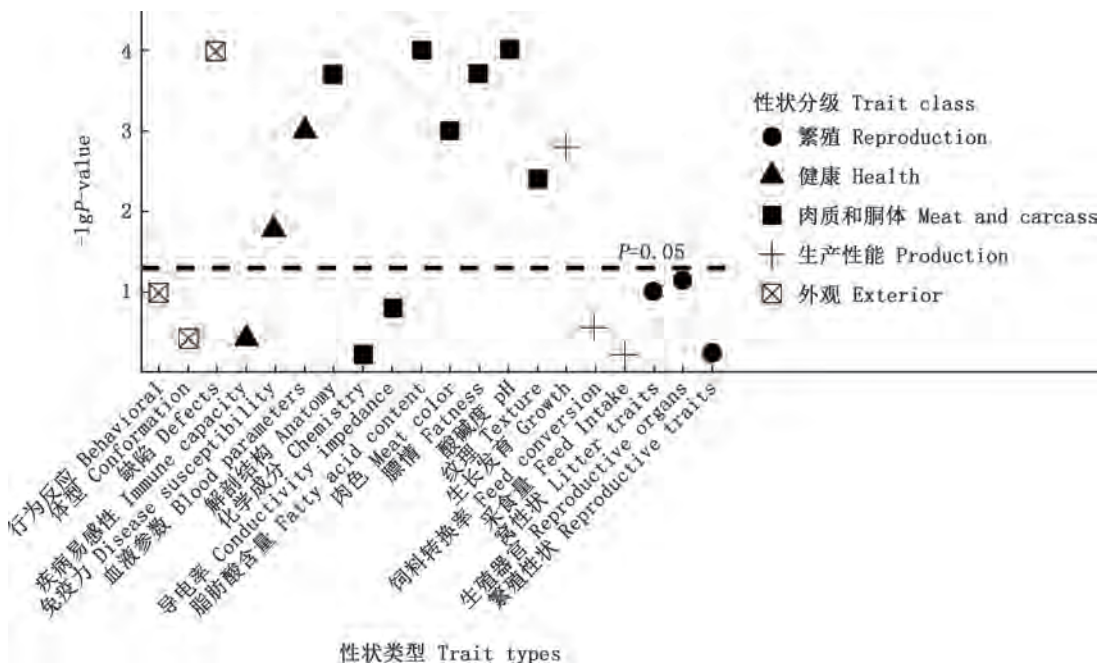
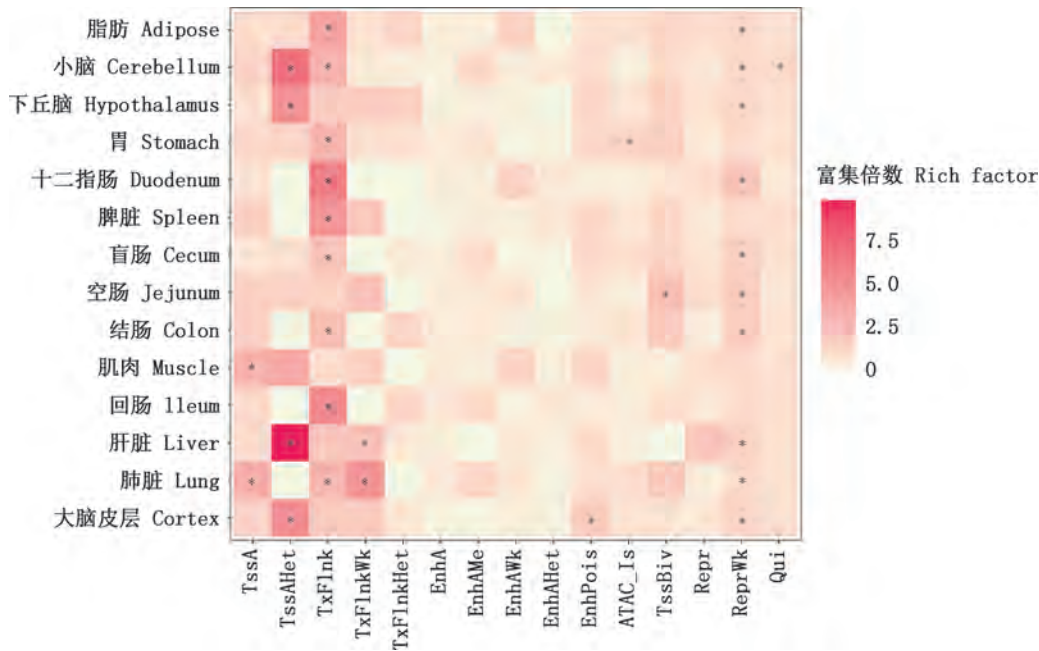


图 4 中国地方猪显著受选择位点的 QTL 区域富集分析散点图

Fig. 4 Scatter plot of enriched QTL regions for significantly selected sites in Chinese indigenous pigs



*, 差异显著 ($P<0.05$)

*, Significant difference ($P<0.05$)

图 5 中国地方猪显著受选择位点的染色质状态富集分析热图

Fig. 5 Heatmap of chromatin state enrichment analysis for significantly selected sites in Chinese indigenous pigs

表 1 中国地方猪候选基因的 GO 功能与 KEGG 通路富集分析

Table 1 GO function and KEGG pathway enrichment analysis for candidate genes in Chinese indigenous pigs

类型 Categories	条目 Terms	P 值 P-value	基因 Genes
GO 功能 GO function			
细胞组分 Cell component	GO:0048471:细胞质核周区(perinuclear region of cytoplasm)	0.004287	<i>FXR1</i> 、 <i>FBXW8</i> 、 <i>PRKCE</i> 、ENSSSCG00000032080、ENSSSCG00000007599
生物过程 Biological process	GO:0006637:酰基-CoA 代谢过程(acyl-CoA metabolic process)	0.004287	<i>ACSM4</i> 、 <i>ACSM5</i> 、ENSSSCG00000007858
	GO:0006633:脂肪酸生物合成过程(fatty acid biosynthetic process)	0.013097	<i>ACSM4</i> 、 <i>ACSM5</i> 、ENSSSCG00000007858
分子功能 Molecular function	GO:0004321:脂肪酰-CoA 合成酶活性(fatty-acyl-CoA synthase activity)	9.39116E-5	<i>ACSM4</i> 、 <i>ACSM5</i> 、ENSSSCG00000007858
	GO:0015645:脂肪酸连接酶活性(fatty acid ligase activity)	9.39116E-5	<i>ACSM4</i> 、 <i>ACSM5</i> 、ENSSSCG00000007858
	GO:0000981:RNA 聚合酶 II 转录因子活性, 序列特异性 DNA 结合(RNA polymerase II transcription factor activity, sequence-specific DNA binding)	0.032998	<i>ZNF502</i> 、 <i>FOXA2</i> 、 <i>LHX9</i> 、 <i>ZKSCAN7</i> 、 <i>BHLHA15</i> 、 <i>ZNF35</i> 、 <i>ZNF660</i> 、ENSSSCG00000020988
KEGG 通路 KEGG pathway	ssc00650: 丁酸代谢(butanoate metabolism)	0.004449	<i>ACSM4</i> 、 <i>ACSM5</i> 、ENSSSCG00000007858

3 讨 论

中国地方猪品种分布广泛,在华南、华中、华北甚至高原部分地区都有所分布。不同地区的气候、地形等生态条件对地方猪进化选择有巨大影响,因此形成了若干体型差异的地方猪品种。本研究基于《中国畜禽遗传资源志·猪志》^[5]中 4 个重要生长性状数据并结合欧氏距离标准化分析对涉及的试验品种进行了统一分类,将包括滇南小耳猪和德保猪在内的 10 个地方猪种作为小体型地方猪组,将包括东北民猪和内江猪在内的 13 个地方猪种作为大体型地方猪组。采用群体 F_{ST} 检验和 XP-EHH 分析确定比较组间的受选择信号区域,两种选择信号检测方法互为补充,可获得到更为准确的选择信号。

本研究共确定了 633 个潜在受选择区域。将潜在受选择区域与 Animal QTL 数据库^[31]进行比对分析发现,与潜在受选择区域发生重叠的 QTL 大多与猪的平均日增重(182 处、1 360 次报道)、滴水损失(95 处、2 252 次报道)、脂肪雄烯酮水平(84 处、338 次报道)、平均背膘厚(57 处、347 次报道)及其他肉质性状相关,而这一结果在大体型和小体型地方猪育肥期平均日增重数据中有所体现,如东北民猪(507.05 ± 2.00) g、内江猪(410.0 ± 11.65) g、巴马香猪(294 ± 14.00) g、香猪(180.0 ± 29.10) g^[5]。389 QTL 富集分析结果显示,潜在受选择区域主要富集

在肉质和胴体性状分级的 QTL 上,同时染色质状态富集分析显示,潜在受选择区域主要富集于 TssAHet 和 TxFlnk 等活性较强的染色体状态。表明不同体型的地方猪遗传调控更有可能是通过调控组织特异性基因的启动子来影响基因表达,从而影响体型大小。

对显著受选择位点的上、下游 200 kb 进行基因注释和通路富集分析后,结合 PigGTE_x 数据库^[13]中家猪 34 个组织的转录组水平图谱发现,4 个基因(*ACSM5*、*ACSM4*、*FXR1*、*FOXA2*)与机体的生长发育、物质代谢存在直接或间接的关联,在代谢相关的通路中多次富集或是在家猪肌肉、脂肪等组织中高表达。其中,*ACSM5* 和 *ACSM4* 属于同一基因家族酰基-CoA 中链家族成员。*ACSM5* 基因编码一种丙酮酸激酶,在脂肪酸代谢中起着重要的作用,*ACSM5* 基因表达水平与猪肉中特定脂肪酸的含量和组成显著相关,即高表达 *ACSM5* 基因的猪肉样品倾向于具有更高的单不饱和脂肪酸含量,揭示了 *ACSM5* 基因与猪肉质特性有关^[32]。作为人类肌肉特异性基因,*FXR1* 基因的编码蛋白在真核生物中可作为 mRNA 转运、稳定性和翻译的因子^[33-35]。McClure 等^[36]强调了 *FXR1* 基因编码蛋白的 X1 与 X3 亚型在人类头颈部鳞状细胞癌(HNSCC)中作为致癌驱动因子的潜在作用,且抑制 *FXR1* 基因表达

会导致细胞衰老增加。先天性高胰岛素血症(CHI)是一种罕见的胰岛素释放失调导致低血糖的疾病,是新生儿期和幼儿期持续严重低血糖的最常见原因。FOXA2被认为是参与调节胰岛β细胞分泌胰岛素的关键基因,其突变可以改变CHI的分子机制^[37];同时,FOXA2和多巴胺(dopamine,DA)神经元的发育和维持息息相关^[38]。

本研究中的候选基因大多在生长发育、代谢、繁殖中起关键作用。综合以上结果,大、小体型中国地方猪曾在生长性状功能上受到选择,从而导致体型差异的形成。不同体型的中国地方猪之间的差异不仅体现在生长发育上,也体现在免疫力和繁殖力上。

4 结 论

本研究通过对大、小体型中国地方猪进行选择信号检测,共获得633个潜在受选择区域,注释到133个基因。其中,受选择区域主要富集在生长发育相关的性状QTL上,以及与胃肠道组织有关的高活性染色质状态中,如TssAHet和TxFlnk。筛选到了与猪体型差异形成相关的候选功能基因,如ACSM4、ACSM5、FXR1和FOXA2。

参考文献(References):

[1] MACHUGH D E, LARSON G, ORLANDO L. Taming the past: Ancient DNA and the study of animal domestication[J]. *Annual Review of Animal Biosciences*, 2017, 5: 329-351.

[2] FRANTZ L A, SCHRAIBER J G, MADSEN O, et al. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes[J]. *Nature Genetics*, 2015, 47(10): 1141-1148.

[3] LARSON G, DOBNEY K, ALBARELLA U, et al. Worldwide phylogeography of wild boar reveals multiple centers of pig domestication[J]. *Science*, 2005, 307(5715): 1618-1621.

[4] ZHANG J, ZHUANG Y, JI H, et al. Pig weight and body size estimation using a multiple output regression convolutional neural network: A fast and fully automatic method[J]. *Sensors*, 2021, 21(9): 3218.

[5] 国家畜禽遗传资源委员会. 中国畜禽遗传资源志——猪志[M]. 北京: 中国农业出版社, 2011.

CHINA NATIONAL COMMISSION OF ANIMAL GENETIC RESOURCES. Animal Genetic Resources 390 in China: Pigs[M]. Beijing: China Agriculture Press,

2011. (in Chinese)

[6] REIMER C, RUBIN C J, SHARIFI A R, et al. Analysis of porcine body size variation using re-sequencing data of miniature and large pigs[J]. *BMC Genomics*, 2018, 19(1): 687.

[7] KIM H, SONG K D, KIM H J, et al. Exploring the genetic signature of body size in Yucatan Miniature pig[J]. *PLoS One*, 2015, 10(4): e0121732.

[8] SIMIANER H, KOEHN F. Genetic management of the Gottingen Mini pig population[J]. *Journal of Pharmacological and Toxicological Methods*, 2010, 62(3): 221-226.

[9] MA Z, CHANG Y, BRITO L F, et al. Multitrait Meta-analyses identify potential candidate genes for growth-related traits in Holstein heifers[J]. *Journal of Dairy Science*, 2023, 106(12): 9055-9070.

[10] MIAO Y, ZHAO Y, WAN S S, et al. Integrated analysis of genome-wide association studies and 3D epigenomic characteristics reveal the BMP2 gene regulating loin muscle depth in Yorkshire pigs[J]. *PLoS Genetics*, 2023, 19(6): e1010820.

[11] MATSUSHITA M, KITOH H, MISHIMA K, et al. Phase 1b study on the repurposing of meclizine hydrochloride for children with achondroplasia[J]. *PLoS One*, 2023, 18(7): e0283425.

[12] LIU P, CHE L, YANG Z, et al. A maternal high-energy diet promotes intestinal development and intrauterine growth of offspring[J]. *Nutrients*, 2016, 8(5): 258.

[13] TENG J, GAO Y, YIN H, et al. A compendium of genetic regulatory effects across pig tissues[J]. *Nature Genetics*, 2024, 56(1): 112-123.

[14] ANDERSSON L, ARCHIBALD A L, BOTTEMA C D, et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the functional annotation of animal genomes project[J]. *Genome Biology*, 2015, 16(1): 57.

[15] ZENG H, ZHANG W, LIN Q, et al. PigBiobank: A valuable resource for understanding genetic and biological mechanisms of diverse complex traits in pigs[J]. *Nucleic Acids Research*, 2023, 52(D1): D980-D989.

[16] ZHOU P, YIN C, WANG Y, et al. Genomic association analysis of growth and backfat traits in Large White pigs[J]. *Genes (Basel)*, 2023, 14(6): 1258.

[17] XU P, LI D, WU Z, et al. An imputation-based genome-wide association study for growth and fatness

traits in Sujiang pigs[J]. *Animal*, 2022, 16(8): 100591.

[18] POKLUKAR K, MESTRE C, SKRLEP M, et al. A Meta-analysis of genetic and phenotypic diversity of European local pig breeds reveals genomic regions associated with breed differentiation for production traits[J]. *Genetics Selection Evolution*, 2023, 55(1): 88.

[19] PURCELL S, NEALE B, TODD-BROWN K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses[J]. *American Journal Human Genetics*, 2007, 81(3): 559-575.

[20] YANG J, LEE S H, GODDARD M E, et al. GCTA: A tool for genome-wide complex trait analysis[J]. *American Journal Human Genetics*, 2011, 88(1): 76-82.

[21] GINESTET C. ggplot2: Elegant graphics for data analysis[J]. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 2011, 174(1): 245-246.

[22] SABETI P C, VARILLY P, FRY B, et al. Genome-wide detection and characterization of positive selection in human populations[J]. *Nature*, 2007, 449(7164): 913-918.

[23] WEIR B S, COCKERHAM C C. Estimating *F*-statistics for the analysis of population structure[J]. *Evolution*, 1984, 38(6): 1358-1370.

[24] SZPIECH Z A, HERNANDEZ R D. Selscan: An efficient multithreaded program to perform EHH-based scans for positive selection[J]. *Molecular Biology Evolution*, 2014, 31(10): 2824-2827.

[25] DANECEK P, AUTON A, ABECASIS G, et al. The variant call format and VCFtools[J]. *Bioinformatics*, 2011, 27(15): 2156-2158.

[26] HU Z L, PARK C A, REECY J M. Bringing the animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services[J]. *Nucleic Acids Research*, 2022, 50(D1): D956-D961.

[27] MALINVERNI R, CORUJO D, GEL B, et al. regioneReloaded: Evaluating the association of multiple genomic region sets[J]. *Bioinformatics*, 2023, 39(11): btad704.

[28] SHEFFIELD N C, BOCK C. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor[J]. *Bioinformatics*, 2015, 32(4): 587-589.

[29] HUNT S E, MCLAREN W, GIL L, et al. Ensembl variation resources[J]. *Database (Oxford)*, 2018, 2018: bay119.

[30] SHERMAN B T, HAO M, QIU J, et al. DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update)[J]. *Nucleic Acids Research*, 2022, 50(W1): W216-W221.

[31] HU Z L, PARK C A, REECY J M. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB[J]. *Nucleic Acids Research*, 2019, 47(D1): D701-D710.

[32] PUIG-OLIVERAS A, REVILLA M, CASTELLO A, et al. Expression-based GWAS identifies variants, gene interactions and key regulators affecting intramuscular fatty acid content and composition in porcine meat[J]. *Scientific Reports*, 2016, 6: 31803.

[33] KIRKPATRICK L L, MCILWAIN K A, NELSON D L. Alternative splicing in the murine and human *FXR1* genes[J]. *Genomics*, 1999, 59(2): 193-202.

[34] DUBE M, HUOT M E, KHANDJIAN E W. Muscle specific fragile X related protein 1 isoforms are sequestered in the nucleus of undifferentiated myoblast[J]. *BMC Genetics*, 2000, 1: 4.

[35] KHANDJIAN E W, BARDONI B, CORBIN F, et al. Novel isoforms of the fragile X related protein FXR1P are expressed during myogenesis[J]. *Human Molecular Genetics*, 1998, 7(13): 2121-2128.

[36] MCCLURE J J, PALANISAMY V. Muscle-specific FXR1 isoforms in squamous cell cancer[J]. *Trends in Cancer*, 2019, 5(2): 82-84.

[37] GIRI D, HAWTON K, SENNIAPPAN S. Congenital hyperinsulinism: Recent updates on molecular mechanisms, diagnosis and management[J]. *Journal of Pediatric Endocrinology Metabolism*, 2022, 35(3): 279-296.

[38] KITTAPPA R, CHANG W W, AWATRAMANI R B, et al. The *FOXA2* gene controls the birth and spontaneous degeneration of dopamine neurons in old age[J]. *PLoS Biology*, 2007, 5(12): e325.

(责任编辑 晋大鹏)