1 **Title Page:**

2 **A comprehensive catalogue of regulatory variants in the cattle transcriptome**

3 Shuli Liu^{1,4†}, Yahui Gao^{1,5†}, Oriol Canela-Xandri^{3†}, Sheng Wang^{6†}, Ying Yu^{4†}, Wentao Cai⁷,
4 Bingjie Li⁸, Erola Pairo-Castineira^{2,3}, Kenton D'Mellow³, Konrad Rawlik², Charley Xia²,
5 Yuelin Yao³, Xiujin Li⁹, Ze Yan⁴, Congjun Li¹, Benjamin D. Rosen¹, Curtis P. Van Tassell¹,
6 Paul M. Vanraden¹, Shengli Zhang⁴, Li Ma⁵, John B. Cole¹, George E. Liu^{1*}, Albert
7 Tenesa^{2,3*}, Lingzhao Fang^{1,2,3*}

8 ¹Animal Genomics and Improvement Laboratory, Henry A. Wallace Beltsville Agricultural
9 Research Center, Agricultural Research Service, Agricultural Research Service, USDA,
10 Beltsville, Maryland 20705, USA

11 ²The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of
12 Edinburgh, Midlothian EH25 9RG, UK

13 ³MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine, The
14 University of Edinburgh, Edinburgh EH4 2XU, UK

15 ⁴College of Animal Science and Technology, China Agricultural University, Beijing 100193,
16 China

17 ⁵Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland
18 20742, USA

19 ⁶State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology,
20 Chinese Academy of Sciences, Kunming, Yunnan 650223, China

21 ⁷Institute of Animal Science, Chinese Academy of Agricultural Science, Beijing 100193,
22 China

23 ⁸Scotland's Rural College (SRUC), Roslin Institute Building, Midlothian EH25 9RG, UK

24 ⁹Guangdong Provincial Key Laboratory of Waterfowl Healthy Breeding, College of Animal
25 Science & Technology, Zhongkai University of Agriculture and Engineering, Guangzhou,

Guangdong 510225, China

[†] Equal contribution

*Corresponding Authors:

GEL: Animal Genomics and Improvement Laboratory, USDA-ARS, Building 306, Room 111, BARC-East, Beltsville, MD 20705, USA.

E-mail: George.Liu@usda.gov, Voice Phone: +1-301-504-9843, Fax: +1-301-504-8414.

AT: The Roslin Institute, Royal (Dick) School of Veterinary Studies, The University of Edinburgh, Midlothian EH25 9RG, UK.

E-mail: Albert.Tenesa@ed.ac.uk

LF: MRC Human Genetics Unit at the Institute of Genetics and Molecular Medicine, The University of Edinburgh, Edinburgh EH4 2XU, UK.

E-mail: Lingzhao.fang@igmm.ed.ac.uk

Emails: shuliliu1991@gmail.com; gyhalvin@gmail.com; Oriol.canela-xandri@igmm.ed.ac.uk; wangsheng@mail.kiz.ac.cn; yuying@cau.edu.cn; caiwentao@caas.cn; Bingjie.Li@sruc.ac.uk; erola.pairo-castineira@igmm.ed.ac.uk; Kenton.D'Mellow@ed.ac.uk; Konrad.Rawlik@roslin.ed.ac.uk; Xiachi.Xin@ed.ac.uk; lixiujin996@126.com; Y.Yao-38@sms.ed.ac.uk; zeyan.ms@gmail.com; Congjun.Li@usda.gov; Ben.Rosen@usda.gov; Curt.VanTassell@usda.gov; Ben.Rosen@usda.gov; Paul.Vanraden@usda.gov; zhangslcau@cau.edu.cn; lima@umd.edu; John.Cole@usda.gov; George.Liu@usda.gov; Albert.Tenesa@ed.ac.uk; Lingzhao.fang@igmm.ed.ac.uk

Running Title: Cattle Genotype-Tissue Expression Atlas

Keywords: Cattle; expression QTLs; GWAS; RNA-Seq; TWAS

Abstract

Characterization of genetic regulatory variants acting on the transcriptome of livestock 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 (cGTEx, <http://cgtex.roslin.ed.ac.uk/>) for the research community based on 11,642 RNA sequences from publicly available datasets representing over 100 cattle tissues. We describe the landscape of the transcriptome across tissues and report hundreds of thousands of *cis*- and *trans*- genetic associations with gene expression and alternative splicing for 24 major tissues. We evaluate the specificity/similarity of these genetic regulatory effects across tissues, and functionally annotate them using a combination of multi-omics data. Finally, we link gene expression in different tissues to 43 economically important traits using a large transcriptome-wide association study (TWAS) to provide novel biological insights into the molecular regulatory mechanisms underpinning agronomic traits in cattle.

Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex traits in human and livestock populations¹⁻³. As the majority of these variants are non-coding, the characterization of the molecular mechanisms by which such variants affect complex traits has been extremely challenging. Indeed, in human genetics, it would have been impossible without projects such as the Genotype-Tissue Expression (GTEx) project that has characterized genetic effects on the human transcriptome and paved the way to understanding the molecular mechanisms of human variation⁴.

However, livestock genetics lags behind human genetics, and to date no study has systematically explored the regulatory landscape of the transcriptome across a wide range of tissues. GWAS signals of agronomic traits are significantly enriched in regulatory regions of genes expressed in trait-relevant tissues in cattle^{5,6}, but attempts to dissect variation in gene expression have generally been small, both in terms of number of individuals and tissues. Here, we describe the largest and most comprehensive study of the regulatory landscape of any livestock species by analyzing 11,642 publicly available cattle RNA-Seq runs, representing over 100 different tissues and cells types. We combined all of these data and make the results freely and easily accessible to the research community through a web portal (<http://cgtex.roslin.ed.ac.uk/>).

There has been a recent exponential growth in the number of RNA-Seq samples made publicly available in cattle (Fig. S1a), but these data have never before been gathered in one collection and processed uniformly. Here, we present a pipeline to uniformly integrate 11,642 public RNA-Seq datasets and identify both *cis*- and *trans*-expression and splicing quantitative trait loci (eQTLs and sQTLs) for 24 important cattle tissues. The latter is facilitated by calling variants directly from the RNA-Seq reads and imputing to sequence level with data from the 1000 Bull Genomes Project⁷, as previously done with human data⁸. Next, we conducted *in silico* analyses, annotating eQTLs and sQTLs with a variety of publicly available omics data, including DNA methylation, chromatin state, and chromatin conformation characteristics. Finally, we integrated gene expression results with a large

GWAS of 27,214 dairy bulls from 43 cattle traits *via* a transcriptome-wide association study (TWAS) and identified 442 previously unknown genes associated with 43 traits. The cattle Genotype-Tissue Expression (cGTEx) atlas will serve as a primary source of reference for cattle genomics, cattle breeding, adaptive evolution, comparative genomics, and veterinary medicine.

Results

Data summary

We analyzed 11,642 publicly available RNA-Seq runs from 8,653 samples using a uniform quality control pipeline, yielding ~200 billion cleaned reads (Table S1). Summary distributions of sequence platform, single/paired reads, clean read number, read length, sex, age, and mapping rate across samples show that the quality of publicly available data is acceptable for the analyses shown here (Fig. S1b-h). We kept 7,180 samples with clean reads number > 500,000 and mapping rates > 60% for subsequent analyses, representing 114 tissues from 47 breeds and breed combinations (Fig. S1i, Table S1). Holsteins were the most represented population (35.5% of all samples), reflecting its global economic value. A total of 1,831 samples (21%) had no breed identified. We grouped the 114 tissues into 13 categories based on trait phylogeny and the 47 breeds into seven major sub-species, with *Bos taurus* representing 87% of all samples (Table S1).

We also analyzed 144 publicly available whole-genome bisulfite sequencing (WGBS) data sets from 21 cattle tissues to investigate the tissue-specificity of DNA methylation and to functionally annotate eQTLs and sQTLs. These data were even more heavily dominated by Holsteins with 90 (62%) samples representing that single breed. These data included ~73 billion good quality reads with an average mapping rate of 71% (Table S2).

Variation in gene expression across individuals and tissues

As expected, the number of expressed genes (Transcripts per Kilobase Million, TPM > 0.1) increased with the number of clean reads. However, we observed a plateau at 50 million

clean reads (Fig. S2a) where we only detected about ~60% of 27,607 annotated genes. This observation suggests that tissues express only a certain proportion of genes and that there is little benefit in increasing read depth beyond this plateau. Only 61 genes were not expressed in any of the samples. About half of those (54.10%) were located in unplaced scaffolds, with significantly ($P < 0.05$, 1-sided) shorter gene length, fewer exons, higher CG density, and lower sequence conservation than expressed genes (Fig. S2b-f).

Similarly, we detected more alternative splicing events with increasing numbers of clean reads (Fig. S2g). More genes (874) had no splicing events than were not expressed (61) in any sample. These genes also exhibited significantly shorter gene length, fewer exons, lower expression, and lower sequence constraints than spliced ones (Fig. S2h-k). Furthermore, 27% of them were housekeeping RNAs and included snRNA, snoRNA, snRNAs, snoRNAs and rRNAs known to play important roles in RNA splicing^{9,10} (Fig. S2l). Genes without splicing events were significantly engaged in the integral component of membrane and G-protein coupled receptor signaling pathway (Fig. S2m).

We found that about 25% of CpG sites were not covered in any of the WGBS samples, even if these had more than 300 million clean reads, partially due to sequence technique problems (Fig. S3a). These CpG sites were enriched in gene desert regions (e.g., telomeres) with significantly higher CG density than the CpG sites captured by the WGBS (Fig. S3b-d).

We called a median of 21,623 SNPs from RNA-Seq data (Fig. S4a), and then imputed each sample up to 3,824,444 SNPs using a multi-breed reference population of 3,310 bulls (Methods, Table S3). We validated the imputation accuracy by comparing imputed and RNA-Seq-derived SNPs with those directly called from whole-genome sequence (WGS) in the same individuals (Fig. S4b, c), and the concordance rate was over 98%. After removing duplicated samples (Methods), we obtained 4,889 individuals for subsequent QTL mapping (Fig. S4d).

We found that clusters of samples based on both gene expression and alternative

splicing accurately recapitulated tissue types (Fig. 1a, b), reinforcing the quality and utility of the data for our purposes. Tissue type was the primary driver of expression/splicing differences, followed by sub-species. For instance, all 831 muscle samples from 46 projects clustered together. Similar to expression and splicing, DNA methylation profiles also recapitulated tissue types (Fig. 1c). But for imputed genotypes, as expected, samples clustered by sub-species (Fig. 1d).

Tissue specificity of gene expression

Tissue-specificity of gene expression was conserved across cattle and human (Fig. 2a-b), and the function of genes with tissue-specific expression accurately reflected the known biology of the tissues (Fig. S5a). For instance, brain-specific genes were significantly associated with synapse and neuron function, and testis-specific genes with spermatogenesis and reproduction. We also calculated tissue-specificity of promoter DNA methylation and alternative splicing (Methods). We found that, based on tissue-specificity, gene expression level was significantly ($FDR < 0.05$) and negatively correlated with DNA methylation level in promoter (Fig. 2c), and positively correlated with splicing ratios (Fig. 2d). For example, *CELF2*, a brain-related gene, had significantly higher expression, lower methylation, and more splicing events in the brain than in the other tissues considered (Fig. 2e). Similarly, the function of genes with tissue-specific promoter hypomethylation and splicing reflected the known biology of the tissues (Fig. S5b, c). Tissue-specific genes exhibited distinct patterns of sequence constraints (Fig. S5d), supporting the hypothesis of tissue-driven genome evolution^{5,11}. We found that while brain-specific genes evolve slowly, blood or testis-specific ones evolve rapidly. This trend was also observed within tissue-specific hypomethylated regions (Fig. 2f; Fig. S5e).

Discovery of expression and splicing QTLs

We identified *cis*-e/sQTLs for 24 tissues with 40 or more individuals, while accounting for relevant confounding factors (Methods, Fig. S6a, b). The number of eGenes (genes with

significant *cis*-eQTLs) ranged from 172 in ileum to 10,174 in blood, with 19,559 (83% of all 23,523 tested genes) eGenes in at least one tissue (Table S4). The number of sGenes (genes with significant *cis*-sQTLs) ranged from four in the salivary gland to 7,913 in macrophages, with 15,376 (70.8%) sGenes in at least one tissue (Table S4). Genes with no *cis*-eQTLs or -sQTLs in any of the tissues were significantly enriched in hormone activity, regulation of receptor activity, neuropeptide signaling pathway, and reproduction. In general, the larger the number of samples, the larger the number of *cis*-e/sGenes detected (Pearson's $r = 0.85$ for *cis*-eQTLs with $P < 1.32 \times 10^{-7}$, $r = 0.63$ for *cis*-sQTLs with $P < 9.1 \times 10^{-4}$), (Fig.3a-b). As expected, with larger sample size, we had more power to detect *cis*-eQTLs with smaller effect sizes (Fig. S7a-b). Consistent with findings in human¹², significant variants (eVariants) centered around the transcript start sites (TSS) of the measured genes (Fig. S7c-d). Across 24 tissues, an average of 46% (range 25.5 - 76.6%) of eVariants were found around TSS ± 100 kb of target genes. In non-eGenes, there was also an enrichment of top SNPs around TSS, suggesting a lack of power to detect such associations (Fig. S7c). Furthermore, we fine-mapped eGenes to assess whether the identified signals were attributed to one or more causal SNPs. We found that 46% (range 14.5 - 73.9%) of eGenes had more than one independent SNPs associated with expression (Fig. 3c), indicating the widespread allele heterogeneity of gene expression. In addition to discover the SNPs with the smallest effects within a locus, we needed to adjust for the SNPs with the biggest effects (Fig. 3d), and the biggest effects tended to be closer to the TSS (Fig. S7e). Allele-specific expression (ASE) analyses found that associated genetic variants were significantly overrepresented in *cis*-eQTLs, and their effect sizes were significantly correlated ($r = 0.25$, $P < 2.2 \times 10^{-16}$) (Fig. 3e, Fig. S7f), verifying our *cis*-eQTLs. To further validate our results, we examined eQTLs that were previously reported with independent datasets in cattle liver ($n = 35$), muscle ($n = 41$) and intramuscular fat ($n = 192$)^{13,14}, and found that the effects of top variants in eGenes previously reported were significantly correlated with the same genes in corresponding tissues in the current study (Fig. S7g).

To investigate whether eQTLs are conserved among sub-species, we conducted *cis*-eQTL mapping for muscle samples from *Bos indicus* (n = 51), *Bos taurus* (n = 505), and their hybrids (n = 108) separately, yielding 86, 2,766, and 800 eGenes, respectively. The effect sizes of SNPs with the largest effects were more similar across muscle tissues of different sub-species than among tissues within the same species (Fig. 3f). About 60.0% of eGene-eVariant pairs detected in the combined muscle samples were reproducible in separate subspecies or the hybrids. Similarly, 72.3%, 79.2% and 56.4% of eGene-eVariant pairs detected in *Bos indicus*, *Bos taurus*, and their hybrids could be validated in the combined analyses of all samples (Fig. S7h). In addition, 51.3% of eGene-eVariant pairs detected in *Bos indicus* were also significant in *Bos taurus*. For example, the expression of *NMRAL1* was consistently and significantly regulated by a *cis*-eQTL among *Bos indicus*, *Bos taurus*, and their hybrids (Fig. 3g). Combining the summary statistics of each sub-species in a meta-analysis found that the results were significantly correlated (Pearson's $r = 0.56$, $P < 2.2 \times 10^{-16}$) with those of the analyses of the combined samples (Fig. 3h). These results suggest that eGene-eVariant associations identified in one sub-species are potentially transferable to other sub-species, particularly for eQTLs with large effect sizes (Fig. 3i-j). Combining samples from different sub-species and breeds will increase statistical power for detecting shared eQTLs, but decreased power will almost certainly result from eQTL present in only some of the sub-species. We also noted the variability in eQTLs among sub-species. For instance, the expression of an immune-related gene, *SSNA1*, was regulated by a *cis*-eQTL in *Bos indicus* ($P = 3.93 \times 10^{-11}$) but not in *Bos taurus* or the hybrids (Fig. 3k).

Patterns and biological mechanisms underlying tissue-specificity/similarity of *cis*-QTLs provide insights into pleiotropic regulatory effects on phenotypes⁴. We therefore conducted a non-model-based pairwise analysis using significant eGene-eVariant pairs in one tissue to estimate the proportion of non-null associations of these pairs in another tissue (Methods). We found that tissues with similar functions (e.g., hypothalamus and pituitary) tended to have a higher proportion of non-null associations, suggesting a higher similarity in their

genetic regulatory effects on both gene expression and splicing (Fig. S8a, b). To investigate whether similarity of genetic regulatory effects reflected the correlations of gene expression across tissues, we estimated the proportion of non-null associations for values of tissue-pairs based on the correlation of gene expression (Fig. S8c). These associations were significantly correlated with those obtained from non-model-based pairwise analysis of *cis*-eQTLs ($r = 0.48$; $P < 2.2 \times 10^{-16}$) and *cis*-sQTLs ($r = 0.61$; $P < 2.2 \times 10^{-16}$). This result suggests that the similarity of gene expression across tissues might be partially mediated through sharing of genetic regulatory effects. Further analysis on expression of eGenes across tissues revealed that effect sizes of eVariants decreased with the increasing number of tissues where target eGenes were expressed (Pearson's $r = -0.27$, $P < 2.2 \times 10^{-16}$), indicating that, on average, tissue-specific genes might have larger genetic regulatory effects than widely-expressed genes (Fig. S8d).

We applied *trans*-eQTL mapping for 13 tissues with over 100 individuals, yielding an average of 2,120 *trans*-eGenes (ranging from 1,397 in hypothalamus to 14,474 in mammary), with 17,372 *trans*-eGenes in at least one tissue (Table S5). We speculated that the large number of *trans*-eQTLs in cattle could be due to the high selection intensity for economically important traits and the modest effective population size. This has led to a complex inter-chromosomal pattern of linkage disequilibrium (LD) and gene co-expression. We showed this might be the case, as we found significantly higher LD for *cis*-eQTL & *trans*-eQTL pairs than *cis*-eQTL & random-SNP pairs (on matched chromosomes) across all tissues (Fig. 4a). Similarly, the expression correlation between *cis*-eGenes and *trans*-eGenes with shared eQTLs were significantly higher than *cis*-eGene and random-gene pairs across all tissues (Fig. 4b). Among 29 autosomes, we observed *trans*-eQTL hotspots in nine chromosomes (Fig. 4c). Genes associated with these *trans*-eQTLs hotspots were significantly enriched in gene co-expression modules (Fig. 4d, e), and played key roles in RNA biology and energy metabolism (Fig. 4f). After accounting for fine-mapped *cis*-eQTLs, the number of *trans*-eGenes slightly decreased in each tissue (Fig. S9, Table S5).

Functional annotation of QTLs

We employed multiple layers of biological data to better define the molecular mechanisms of genetic regulatory effects. As expected, *cis*-e/sQTLs were significantly enriched in functional elements, such as 3'UTR and open chromatin regions (Fig. 5a-b). As expected, *cis*-sQTLs had a significantly higher enrichment in splice regions than *cis*-eQTLs. The *cis*-eQTLs associated with stop gain had significantly larger effect sizes than other *cis*-eQTLs (Fig. 5c).

Hypomethylated regions across 13 tissues were enriched in *cis*-e/sQTLs compared to other 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 distinct patterns of alternative splicing, related to the development of the nervous system¹², *cis*-sQTLs but not *cis*-eQTLs in the hypothalamus and pituitary had the highest enrichments in their specific hypomethylated regions (Fig. 5e).

Topologically associated domains (TADs) enable chromatin interactions between distal regulatory regions and target promoters¹⁵. By examining Hi-C data of lung tissue in cattle, we obtained TADs and significant Hi-C contacts, which were conserved across tissues¹⁵. By comparing with random eGene-SNP pairs with matched distances, we observed significantly higher percentages of eGene-eVariant pairs within TADs across the majority of tissues (Fig. 5f). For instance, *APCS* and its *cis*-eQTL peak (144 kb upstream of the TSS) were encompassed by a TAD and linked by a significant Hi-C contact, which allowed the regulation of its expression by distal (> 2kb from TSS) eVariants (Fig. 5g).

eQTLs and complex trait associations

The primary goal of this study is to provide a resource for elucidating the genetic and biological bases of agronomic traits in cattle. We thus evaluated e/sQTLs detected in each tissue for associations with four distinct agronomic traits, i.e., ketosis, somatic cell score in milk (SCS), age at first calving (AFC), and milk yield (MY). The top SNPs associated with

ketosis from GWAS were significantly enriched within liver *cis*-e/sQTLs and *trans*-eQTLs (Fig. 6a, Fig. S10a). Similarly, MY-associated SNPs were overrepresented in mammary gland *cis*-e/sQTLs and *trans*-eQTLs (Fig. 6b, Fig. S10b). Compared to other tissues, mammary gland, milk cells and liver were top tissues regarding the enrichment of *cis*-eQTLs with MY-associated SNPs (Fig. 6c). Additionally, AFC-associated SNPs were most enriched for monocytes *cis*-eQTLs, and SCS for mammary gland (Fig. S10c-d).

We then detected 496 candidate gene-tissue pairs for 43 agronomic traits in cattle *via* TWAS, representing 337 unique genes (Table S6). Out of 319 previously fine-mapped genes^{16,17}, we verified 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). The expression of *ZNF613* in hypothalamus was the top significant association for many reproduction and body conformation traits, including daughter-still-birth and stature, supporting our previous finding that *ZNF613* was significantly associated with gestation length might be through affecting embryonic development¹⁸. Overall, TWAS enhanced our ability to detect causal genes and to better understand the biological underpinnings of these traits.

Discussion

The cGTEx atlas represents the most comprehensive reference resource of the cattle transcriptome to date. It provides a detailed characterization of genetic control of gene regulation across 24 tissues. Compared to the human GTEx, we identified a large number of *trans*-eQTLs even with modest sample sizes. This result is likely explained by the distinct LD properties between the two species, as factors like domestication, breed formation, and strong artificial selection for production traits (e.g., polygenic selection of milk yield) have led to long-distance LD within chromosomes and LD even among non-syntenic chromosomes. Reinforcing this hypothesis, we detected the largest number of *trans*-eQTLs in the mammary gland, which might be the top tissue under selection in dairy cattle¹⁹, even

with a relatively small sample size compared to other tissues.

Three types of QTLs (*cis*-eQTLs and *trans*-eQTLs) detected in this study provide a rich set of candidate functional variants for agronomic traits in cattle, as we found that top GWAS associations for a trait were significantly enriched with these regulatory QTLs in their relevant target tissues. Our data will aid future colocalization analyses for significant GWAS results and inform candidate variants/genes for functional follow-up. Further integration of our QTLs with functional annotations of a range of tissues from the on-going Functional Annotation of Animal Genomes (FAANG) project will provide valuable opportunities to understand transcriptional/post-transcriptional regulatory mechanisms underpinning GWAS hits for agronomic traits²⁰. Our study will enable exploring the molecular mechanisms underlying the extensive pleiotropic effects identified in livestock²¹. This information will allow the understanding of mechanisms of response to intended selection as well as disentangling unintended and unfavorable correlated responses to this same selection (e.g. increasing mastitis or deteriorating fertility when selection for increased milk production). This resource will assist in the development of genome selection methods to improve targeted responses in animal health and wellbeing. Moreover, a better understanding of the genetic architecture underpinning agronomic traits will benefit genetic improvement programs by incorporating biological knowledge into genomic prediction models^{7,22}, and may improve prediction accuracy across generations and populations.

Online Methods

Quantification of gene expression

We downloaded 11,642 RNA-Seq runs (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/>). We merged multiple runs 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 adaptors and low quality reads using

Trimmomatic (v0.39)²³ 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 genome²⁴ using single or paired mapping modules of STAR (v2.7.0) with parameters of outFilterMismatchNmax 3, outFilterMultimapNmax 10 and outFilterScoreMinOverLread 0.66. We kept 7,264 samples with uniquely mapping rates $\geq 60\%$ (mean, 91.07%; range, 60.44%-100%; mapping details in Table S1). We then obtained normalized expression (TPM) of 27,608 Ensembl (v96) annotated genes using Stringtie (v2.1.1)²⁵, and extracted raw read counts of them with featureCounts (v1.5.2)²⁶ as well. We finally clustered 7,264 samples based on $\log_2(\text{TPM} + 1)$ using a hierarchical clustering method, implemented in R 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 (v0.2.9)²⁷ to identify and quantify variable alternative splicing events of genes by leveraging information of junction reads (i.e., reads spanning introns) that were obtained from the STAR alignment. The 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 isoforms²⁷. 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 less 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 Projects⁷ for 7,180 high-quality RNA-Seq samples individually, following the recommended best practices pipeline in Genome Analysis Toolkit (GATK) (v4.0.8.1)²⁸ with default settings. We filtered 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 (v5.1)²⁹ based on a multiple-breed reference population consisted of 3,103 individuals from run7 of the 1000 Bull Genomes Project⁷ and 207 public individuals from *Bos taurus* (n = 101), *Bos indicus* (zebu, n = 20), and *Bos grunniens* (yak, n = 86) (Table S3). Finally, we obtained 6,123 samples that were genotyped and imputed successfully. We filtered out variants with MAF < 0.05 and dosage R-squared (DR^2) < 0.8, resulting in 3,824,444 SNPs used for QTL mapping. To evaluate the accuracy of imputation, we called genotypes (~6 M SNPs) from WGS (average read depth = $10 \times$) of four Holstein individuals, which had RNA-Seq data from 13 tissues as well. We then measured the genotype concordance rates between WGS-SNPs and RNA-Seq/imputed SNPs across 13 tissues. We then extracted 153,913 LD-independent SNPs using plink (v1.90)³⁰ (--indep-pairwise 1000 5 0.2), and conducted PCA analysis for all 6,123 samples using these SNPs in EIGENSOFT (v7.2.1)³¹. We also calculated the identity-by-state distance (IBS distance) among samples by using these independent SNPs to remove duplicate individuals (i.e., IBS distance > 0.85). IBS distance = $(IBS2 + 0.5 \cdot IBS1) / (IBS0 + IBS1 + IBS2)$, where IBS0 is the number of IBS 0 non-missing variants, IBS1 is the number of IBS 1 non-missing variants and IBS2 is the number of IBS 2 non-missing variants. We set an IBS distance cutoff of 0.85 to deem two samples as duplicates and randomly kept one of them. When conducting QTL mapping, we removed an average of 43 duplicate samples within each tested tissue (ranging from 1 for salivary gland and leukocyte to 132 for muscle).

Allele specific expression (ASE)

We conducted ASE analysis using the GATK ASEReadCounter tool (v4.0.8.1) 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 10 reads per allele, and at least 2% of all reads supporting the minor allele. We then calculated a binominal *P*-value by comparing to the expected ratio under the null hypothesis, followed by multiple-test correction with the Benjamini–Hochberg (BH) approach. SNPs with FDR < 0.05 were considered as significant ASE. We calculated the effect size of an ASE variant in a tissue as median log-ratio:

$$s(x, y) = \text{median}[\log_2 \left(\frac{x_i}{y_i} \right)]$$

where: *s*(*x*,*y*) is the effect size for a tissue, *x_i* is the number of RNA-Seq reads aligned to alternative allele in individual *i* in a given tissue, *y_i* is the number of RNA-Seq reads aligned to reference allele in individual *i* in the same tissue.

DNA methylation analysis of WGBS data

We downloaded (in July, 2019) 144 publicly available WGBS cattle data from 21 distinct tissues (Table S2). We first used FastQC (v0.11.2) and Trim Galore v0.4.0 (--max_n 15 --quality 20 --length 20 -e 0.1) to determine read quality and to filter reads with low quality, respectively. We then mapped clean reads to the same cattle reference genome (ARS-UCD1.2) using Bismark software (v0.14.5)³² 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 on the basis of clean reads, ranging from 5 to 47-fold coverage with an average of 27.6-fold coverage. Ultimately, we kept CpG sites that were represented by at least five reads for subsequent analyses. We clustered samples based on DNA methylation levels of shared CpGs using both hierarchical clustering and t-SNE approaches³³.

Identification of TAD and significant Hi-C contacts

To find potential chromatin interactions between distal eVariants and target eGenes, we

identified TADs and Hi-C contacts from Hi-C data from lung tissue in cattle that was retrieved from NCBI Sequence Read Archive (SRA) under accessions: SRR5753600, SRR5753603, and SRR5753606. We used Trim Galore (v0.4.0) to trim adapter sequences and low-quality reads (`--max_n 15 --quality 20 --length 20 -e 0.1`), resulting ~820 million clean reads. We then mapped clean reads to the cattle reference genome (ARS-UCD1.2²⁴) using BWA³⁴. We applied HiCExplorer v3.4.1³⁵ to build a Hi-C contact matrix with 10kb resolution and identified TAD with hicFindTAD. We kept TADs with FDR less than 0.01 to link eQTLs to eGenes. We further employed HiC-Pro (v2.11.4)³⁶ to call Hi-C contacts with 10 kb resolution from Hi-C data. Briefly, HiC-Pro aligned clean reads to the cattle reference genome with Bowtie2 (v2.3.5)³⁴. After building a contact matrix, HiC-Pro generated intra- and inter-chromosomal maps and normalized them using the original ICE normalization algorithm. We considered Hi-C contacts with $FDR < 0.05$ as significant.

Tissue-specificity analysis of gene expression, alternative splicing and DNA methylation

To quantify tissue-specific expression of genes, we computed a *t*-statistics for each gene in each of the 114 tissues. We grouped 114 tissues into 13 categories (Table S1). We scaled the log₂-transformed expression (i.e., log₂TPM) of genes to have a mean of zero and variance of one within each tissue. We then fitted a linear model as described in³⁷ for each gene in each tissue using least-squares. When constructing the matrix of dummy variables (i.e., design matrix) for tissues, we denoted samples of the target tissue/cell type (e.g., CD4 cells) as '1', while samples outside the target category (e.g., non-blood/immune tissues) as '-1'. We excluded samples within the same category (e.g., 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 largest *t*-statistics as genes with high tissue-specific expression.

To detect tissue-specific alternative splicing, we used leafcutter to analyze the differential intron excision by comparing the samples from target tissue to the remaining tissues²⁷.

Samples from tissues of the same category with the target tissue were excluded. We used Benjamini-Hochberg method (FDR) to control multiple testing.

For DNA methylation, we focused on DNA methylation levels of gene promoters (from upstream 1500bp to downstream 500bp of TSS) and body regions (from TSS to TES), which were calculated with a weighted methylation method using the roimethstat function in MethPipe (v3.4.3)³⁸. 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 SMART2³⁹ 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) method⁴⁰. In each tissue, we estimated 75 PEER factors first. The posterior variances of factor weights dramatically decreased and reached or nearly reached plains when 10 PEER factors were included (Fig. S6a). Therefore, we used 10 PEER covariates to account for effects of confounding variables on gene expression in all following QTL analyses. For instance, the variance of gene expression among samples in liver captured by 9 out of 10 PEER factors were significantly (FDR < 0.05) correlated with known technical and biological covariates like clean data size, mapping rate, project, breeds, sub-species, sex and age. (Fig. S6b). To further control the effect of population structure on the discovery of QTLs, we included genotype 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 24 tissues with at least 40 individuals each, while adjusting for corresponding PEER factors and genotype PCs. Detailed information about these 24 tissues is in Table S4. We kept genes with TPM > 0.1 in $\geq 20\%$ 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 (TMM) across samples. The *cis*-eQTL mapping was done using a linear regression model, implemented in FastQTL⁴¹, to test associations of normalized expression level of genes (TMM) with genetic variants in 1Mb of transcriptional start sites (TSS) of target genes. We only considered imputed variants with MAF ≥ 0.01 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, which was calculated with 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_t was defined as the empirical *P*-value of the gene closest to the 0.05 FDR threshold^{4,42}. We then calculated the nominal threshold as $F^{-1}(p_t)$, where F^{-1} is the binominal inverse cumulative distribution, of which parameters for genes were obtained from the above permutation mode of FastQTL analysis. We considered variants with nominal *P*-values below the nominal threshold as significant, and included them into the list of eGene-eVariant pairs. We calculated the allele fold change (aFC), defined as the ratio of 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 (v0.3) tools⁴³. We further applied the statistical fine-mapping method, dap-g⁴⁴, to infer multiple independent casual *cis*-eQTLs of a gene in a tissue. The dap-g approach employed a Bayesian variable selection model, using a signal-level posterior inclusion probability (SPIP) to measure the strength of each association signal (SNPs in LD). We set a cutoff of 0.1 (i.e., SPIP > 0.9) as the inclusion threshold to

detect representative/independent eQTLs for the target eGene. To analyze pairwise tissue similarity in cis-eQTL, we calculated π_1 value, defined as proportion of true positive cis-eQTLs identified in first tissue amongst all tested gene-variant pairs in second tissue, using the Storey and Tibshirani qvalue approach, as described in¹².

Meta-analysis of cis-eQTLs of muscle samples from three sub-species

Data from muscle samples were available from three sub-species: *Bos indicus* (n = 51), *Bos taurus* (n = 505), and their crosses (n = 108). To explore the similarity and variability of cis-eQTLs among sub-species, we associated expression of target genes with their local SNPs (1Mb around TSS) using muscle samples from each of sub-species separately. We then conducted a meta-analysis to integrate cis-eQTL results from three sub-species using the Metal 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 sub-species⁴⁵.

cis-sQTL mapping

In each of the 24 tissues, we applied a linear regression model, implemented in FastQTL⁴¹, to test for associations of genotypes (MAF > 1%) 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 10 PEER factors to adjust for the effect of unknown confounding variables. We applied the permutation pass mode (1,000 permutations) in FastQTL⁴¹ to obtain beta approximated permutation *p* values, followed by multiple test correction with the FDR method. We considered sQTL-intron cluster pairs with FDR < 0.05 as significant, and defined sGene as genes containing a significant sQTL in any introns.

trans-eQTL mapping

We conducted trans-eQTLs for 13 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 length of 36 and 75 following the procedure described in https://wiki.bits.vib.be/index.php/Create_a_mappability_track. We excluded any variants within repeats (Repeatmasker and simple repeats), and further removed variants with mappability < 1, based on k-mer length of 75. After filtering, we kept SNPs with MAF > 5% and at least 10 minor alleles within each tissue for association testing. We mapped *trans*-eQTLs for protein-coding genes with average mappability ≥ 0.8 based on k-mer length of 36 using MatrixQTL⁴⁶, by associating expression of target genes with genotypes on other autosomal chromosomes, while adjusting for the same covariates as in *cis*-eQTL analysis. We adjusted *P*-values for multiple testing using Benjamini-Hochberg method to obtain FDR. We considered gene-variant pairs with FDR < 0.05 as significant.

WGCNA co-expression network analysis and estimation of π_1 statistics

We applied the Weighted Gene Co-Expression network (WGCNA) analysis⁴⁷ to obtain gene co-expression network in each of 24 tissues used in eQTL mapping. We estimated the sharing of *cis*-eQTLs and *cis*-sQTLs among tissues using the π_1 statistics, as described in human GTEx Consortium (2015)¹². Briefly, given a “discovery” tissue A and a “Validation” tissue B, we extracted *p*-values for all significant *cis*-eQTLs in A from B, and then computed a Storey’s π_1 value based on these *P*-values. We also computed the π_1 statistics of tissue-pairs based on gene expression correlation to quantify the sharing of gene expression among tissues.

TWAS analyses

To associate gene expression in a tissue with complex traits, we conducted TWAS analysis using S-PrediXcan⁴⁸ by prioritizing GWAS summary statistics for 43 agronomic traits of economic importance in cattle, 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 bulls as described

previously¹⁶. 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 bulls¹⁷. We constructed a Nested Cross Validated Elastic Net prediction model using genotype and expression data. We included sub-species, 10 PEER factors and corresponding genotype PCs (details in *cis*-eQTL mapping part) in the model to adjust for unknown confounding variables and underlying population structure. For each trait, we conducted TWAS in each of the same 24 tissues as in *cis*-eQTL mapping. We considered genes with Bonferroni-corrected P (q -value) < 0.05 as significant. We visualized the Manhattan plots of P -values of all tested genes using ggplot2 (v3.3.2) in R (v3.4.1).

Other downstream bioinformatic analysis

We used Genomic Association Tester (GATv1.3.4)⁴⁹ 10,000 permutations to estimate the functional enrichment of QTLs in particular genomic regions, chromatin states, methylation elements, and WGCNA co-expression modules. We considered enrichments with FDR (adjusted P -values with Benjamini-Hochberg method) < 0.05 as significant. We used the R package, ClusterProfiler⁵⁰, to annotate function of genes based on the Gene ontology database from Bioconductor (org.Bt.eg.db v3.11.4). We considered GO terms with FDR < 0.05 as significant. We obtained complex traits/diseases that were associated with a particular gene in humans using the Region PheWAS function in GeneATLAS database (<http://geneatlas.roslin.ed.ac.uk/>) with set region of ± 50 kb and P -value threshold of 1×10^{-8} .

Acknowledgements

This work was supported in part by AFRI grant numbers 2013-67015-20951, 2016-67015-24886, and 2019-67015-29321 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. A. Tenesa acknowledged funding from the BBSRC through programme grants BBS/E/D/10002070 and BBS/E/D/30002275, MRC research grant MR/P015514/1, and

HDR-UK award HDR-9004. L. Fang. was partially funded through HDR-UK award HDR-9004 and the Marie Skłodowska-Curie grant agreement No [801215]. O. Canela-Xandri was supported by MR/R025851/1.

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 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. The authors have not stated any conflicts of interest.

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., W.C., B.J., C.X., Y.Yao, Z.Y. and X.L., performed bioinformatics analyses. O.C.-X., L.F., E.P.-C., K. D., K.R., C.-J.L., B.D.R., C.P.V.T., P.M.V., J.B.C., Y.Yu, S.Z. and A.T. contributed to the resource generation. S.L., Y.G., L.F., G.E.L. and A.T. wrote the

manuscript. All authors read, edited and approved the final manuscript.

Completing interests statement

The authors declare no competing interests.

Data availability statement

All raw sequencing data analyzed in this study are publicly available in NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>). All processed results and scripts codes can be found in <https://cgtex.roslin.ed.ac.uk/>.

References

1. 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).
2. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat Genet* **50**, 1593-1599 (2018).
3. 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).
4. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
5. 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).
6. Xiang, R. *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* **116**, 19398-19408 (2019).
7. 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).
8. 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).
9. Didychuk, A.L., Butcher, S.E. & Brow, D.A. The life of U6 small nuclear RNA, from cradle to grave. *RNA* **24**, 437-460 (2018).
10. Hombach, S. & Kretz, M. Non-coding RNAs: Classification, Biology and Functioning.

625 *Adv Exp Med Biol* **937**, 3-17 (2016).

626 11. Gu, X. & Su, Z. Tissue-driven hypothesis of genomic evolution and sequence-
627 expression correlations. *Proc Natl Acad Sci U S A* **104**, 2779-84 (2007).

628 12. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot
629 analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).

630 13. Xiang, R. *et al.* Genome variants associated with RNA splicing variations in bovine
631 are extensively shared between tissues. *BMC Genomics* **19**, 521 (2018).

632 14. Cesar, A.S.M. *et al.* Identification of putative regulatory regions and transcription
633 factors associated with intramuscular fat content traits. *BMC Genomics* **19**, 499 (2018).

634 15. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis
635 of chromatin interactions. *Nature* **485**, 376-80 (2012).

636 16. Jiang, J. *et al.* Functional annotation and Bayesian fine-mapping reveals candidate
637 genes for important agronomic traits in Holstein bulls. *Commun Biol* **2**, 212 (2019).

638 17. Freebern, E. *et al.* GWAS and fine-mapping of livability and six disease traits in
639 Holstein cattle. *BMC Genomics* **21**, 41 (2020).

640 18. Fang, L. *et al.* Genetic and epigenetic architecture of paternal origin contribute to
641 gestation length in cattle. *Commun Biol* **2**, 100 (2019).

642 19. Lemay, D.G. *et al.* The bovine lactation genome: insights into the evolution of
643 mammalian milk. *Genome Biol* **10**, R43 (2009).

644 20. Andersson, L. *et al.* Coordinated international action to accelerate genome-to-phenome
645 with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* **16**,
646 57 (2015).

647 21. Xiang, R., van den Berg, I., MacLeod, I.M., Daetwyler, H.D. & Goddard, M.E. Effect
648 direction meta-analysis of GWAS identifies extreme, prevalent and shared pleiotropy
649 in a large mammal. *Commun Biol* **3**, 88 (2020).

650 22. Daetwyler, H.D., Pong-Wong, R., Villanueva, B. & Woolliams, J.A. The impact of
651 genetic architecture on genome-wide evaluation methods. *Genetics* **185**, 1021-31
652 (2010).

653 23. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
654 sequence data. *Bioinformatics* **30**, 2114-20 (2014).

655 24. Rosen, B.D. *et al.* De novo assembly of the cattle reference genome with single-
656 molecule sequencing. *Gigascience* **9**(2020).

657 25. Pertea, M., Kim, D., Pertea, G.M., Leek, J.T. & Salzberg, S.L. Transcript-level
658 expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.
659 *Nat Protoc* **11**, 1650-67 (2016).

660 26. Liao, Y., Smyth, G.K. & Shi, W. featureCounts: an efficient general purpose program
661 for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-30 (2014).

662 27. Li, Y.I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat*
663 *Genet* **50**, 151-158 (2018).

664 28. Van der Auwera, G.A. *et al.* From FastQ data to high confidence variant calls: the
665 Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11

10 1-11 10 33 (2013).

29. 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).

30. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

31. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet* **2**, e190 (2006).

32. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-2 (2011).

33. Maaten, L.v.d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).

34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60 (2009).

35. Ramirez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**, 189 (2018).

36. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).

37. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).

38. Song, Q. *et al.* A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8**, e81148 (2013).

39. 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).

40. 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**, 500-7 (2012).

41. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-85 (2016).

42. Consortium, G.T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

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**, 2190-1 (2010).

- 707 46. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
708 *Bioinformatics* **28**, 1353-8 (2012).
- 709 47. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
710 analysis. *BMC Bioinformatics* **9**, 559 (2008).
- 711 48. Barbeira, A.N. *et al.* Exploring the phenotypic consequences of tissue specific gene
712 expression variation inferred from GWAS summary statistics. *Nat Commun* **9**, 1825
713 (2018).
- 714 49. Heger, A., Webber, C., Goodson, M., Ponting, C.P. & Lunter, G. GAT: a simulation
715 framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046-8
716 (2013).
- 717 50. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing
718 biological themes among gene clusters. *OMICS* **16**, 284-7 (2012).

719

Fig. 2. Tissue-specificity of gene expression, alternative splicing and DNA methylation.

(a) Median expression profile of tissue-specific expressed genes across tissues. Each row represents a gene, and each column represents an individual tissue. Color represents median $\log_2(\text{TPM}+1)$ of genes among samples in a tissue. **(b)** Median expression profile of homologous genes across tissues in human GTEx (v8)⁴. **(c)** Pearson correlation of tissue-specificity between mRNA and promoter DNA methylation levels. X-axis is for tissues in RNA-Seq data, and y-axis is for tissues in DNA methylation data. Red boxes indicate the high negative correlation between matched tissues. **(d)** Pearson correlation of tissue-

specificity between gene expression (TPM, x-axis) and alternative splicing (PSI, y-axis). **(e)** *CELF2* shows specific high expression, low DNA methylation levels in splice sites and high PSI value of spliced introns (chr13:13034717-13197300) in brain tissue compared to the rest of tissues. Red boxes indicate the high positive correlation between matched tissues. **(f)** The relationship between PhastCons scores of tissue-specific DNA methylation regions and dN/dS ratio of tissue-specific expressed genes across tissues (Pearson's $r = -0.68$, $P = 3.6 \times 10^{-3}$).

Fig. 3. Discovery and characterization of cis-eQTLs and cis-sQTLs. (a) Relationship between the percentage of eGenes over all tested genes and sample size (Pearson $r = 0.83$; $P < 2.2 \times 10^{-16}$) across 24 tissues. (b) Relationship between the percentage of sGenes over tested genes and sample size (Pearson $r = 0.63$; $P < 2.2 \times 10^{-16}$) across 24 tissues. (c) Distribution and average number of conditionally independent eQTLs per gene across tissues. Tissues are ordered by sample size. (d) Effect sizes of the first four fitted eQTLs in the fine-mapping analyses. Effect sizes, measured as $\log_2$ allelic fold change ($|aFC(\log_2)|$), decrease from 1st to 4th eQTL. * indicates significant differences ($P < 0.05$) with Student t -test. (e) cis-eQTLs

were significantly ($P < 1 \times 10^{-14}$, denoted as *, Fisher Exact test) overrepresented in the ASEs. The Y-axis indicates the percentage of *cis*-eQTLs overlapped with ASEs compared to all tested SNPs for ASE detection. **(f)** Correlation of effect size ($|aFC(\log_2)|$) of all tested SNPs across tissues and sub-species. Liver, adipose, blood and rumen tissues were measured in *Bos taurus* **(g)** A shared *cis*-eQTL (rs208377990) for *NMRAL1* in muscle across *Bos indicus*, *Bos taurus* and their crosses. 0, 1 and 2 indicate genotypes AA, AG, and GG, respectively. **(h)** Pearson's correlation between the regression slopes of *cis*-eQTLs in the combined muscle population and the Z-score from the multi-species meta-analysis. **(i)** Percentage of *cis*-eQTLs in the combined muscle data that are replicated in multi-species meta-analysis as a function of different *P*-value cutoffs. The *cis*-eQTLs with higher significant levels in the combined population are more likely to be replicated in multi-species meta-analysis. **(j)** Effects sizes ($|aFC(\log_2)|$) of *cis*-eQTLs replicated in the multi-species meta-analysis are larger than those that didn't replicate. **(k)** A *Bos indicus*-specific *cis*-eQTL (rs110492559) for *SSNAI1*. 0, 1 and 2 indicate genotypes GG, AG, and AA, respectively.

Fig. 4. Discovery and characterization of *trans*-eQTLs. **(a)** Comparison of linkage disequilibrium (r^2) between *cis*-eQTLs vs. *trans*-eQTLs with shared genes and *cis*-eQTLs vs. random SNPs on matched chromosomes. LD levels between the *cis*-eQTLs and their mediated *trans*-eQTLs are significantly ($P < 0.05$, Student *t*-test) higher than those between *cis*-eQTLs and random SNPs. **(b)** Comparison of expression correlation (r^2) between *cis*-eGenes and *trans*-eGenes with shared eQTLs; and between *cis*-eGenes and random genes on matched chromosomes. Expression correlations of *cis*-eGenes vs. *trans*-eGenes are significantly higher than those of *cis*-eGenes vs. random genes ($P < 0.05$, Student *t*-test). **(c)** Manhattan plot for the number of genes with associated *trans*-eQTLs across the genome. The

778 nine *trans*-eQTLs peaks marked explained variance of more than 1,000 different genes **(d)**
 779 176 genes are shared across all *trans*-eQTLs in the nine peaks. **(e)** Genes mediated by *trans*-
 780 eQTLs are significantly ($P < 0.05$, 10,000 times Permutation test) enriched in tissue-specific
 781 gene co-expression modules. Each color corresponds to gene co-expression modules of a
 782 tissue. The average odd ratio is 1.76. **(f)** GO enrichment of genes mediated by *trans*-eQTLs
 783 in each of the nine peaks. P values are obtained from the hypergeometric test.

784
 785 **Fig. 5. Functional annotation of QTLs.** (a) Enrichment of *cis*-eQTLs, *cis*-sQTLs and *trans*-
 786 eQTLs in sequence ontology. (b) Enrichment of *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in

787 15 chromatin states predicted from Rumen tissue in Holstein. **(c)** Effect sizes (measured as
788 $\log_2\text{aFC}$) of *cis*-eQTLs across sequence ontology. Enrichment of *cis*-eQTLs **(d)** and *cis*-
789 sQTLs **(e)** in tissue-specific hypomethylated regions. Higher enrichments are observed for
790 matched tissues (highlighted dots). Two brain related tissues, hypothalamus and pituitary,
791 show specifically higher enrichments for *cis*-sQTLs compared to *cis*-eQTLs. **(f)** Percentages
792 of eGene-eVariant pairs within topologically associating domains (TADs) are significantly (P
793 < 0.05 , Fisher Exact test) higher than those of random eGene-SNP pairs matched by distance.
794 Triangle represents eGene-eVariant pairs; circle represents random matching eGene-SNP
795 pairs. **(g)** A significant HiC-contact links an eVariant rs136092944 to its distal regulated gene
796 *APCS*. This eGene-eVariant pair is also within a TAD. Expression of *APCS* is significantly
797 associated with rs136092944. 0, 1 and 2 indicate genotypes AA, AG, and GG, respectively.

Fig. 6. Relationship between QTLs and livestock traits of economic value. (a) *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in liver show higher enrichments for SNPs associated with ketosis compared to genome-wide SNPs (shown in grey). **(b)** *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in mammary gland show higher enrichments for SNPs associated with milk yield compared to genome-wide SNPs. **(c)** Enrichment of *cis*-eQTLs for milk yield associations is tissue-dependent. The *cis*-eQTLs in mammary gland, milk cells and liver exhibit higher enrichments for milk yield associations compared to those in other 10 tissues. **(d)** Manhattan plots of TWAS for milk yield.