Title Page:

A comprehensive catalogue of regulatory variants in the cattle transcriptome

Shuli Liu[1,4†], Yahui Gao[1,5†], Oriol Canela-Xandri[3†], Sheng Wang[6†], Ying Yu[4†], Wentao Cai[7], Bingjie Li[8], Erola Pairo-Castineira[2,3], Kenton D'Mellow[3], Konrad Rawlik[2], Charley Xia[2], Yuelin Yao[3], Xiujin Li[9], Ze Yan[4], Congjun Li[1], Benjamin D. Rosen[1], Curtis P. Van Tassell[1], Paul M. Vanraden[1], Shengli Zhang[4], Li Ma[5], John B. Cole[1], George E. Liu[1*], Albert Tenesa[2,3*], Lingzhao Fang[1,2,3*]

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

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

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

[4]College of Animal Science and Technology, China Agricultural University, Beijing 100193, China

[5]Department of Animal and Avian Sciences, University of Maryland, College Park, Maryland 20742, USA

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

[7]Institute of Animal Science, Chinese Academy of Agricultural Science, Beijing 100193, China

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

[9]Guangdong Provincial Key Laboratory of Waterfowl Healthy Breeding, College of Animal 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.

64 **Introduction**

65 Genome-wide association studies (GWAS) have identified thousands of genetic
66 variants associated with complex traits in human and livestock populations[1-3]. As the
67 majority of these variants are non-coding, the characterization of the molecular mechanisms
68 by which such variants affect complex traits has been extremely challenging. Indeed, in
69 human genetics, it would have been impossible without projects such as the Genotype-Tissue
70 Expression (GTEx) project that has characterized genetic effects on the human transcriptome
71 and paved the way to understanding the molecular mechanisms of human variation[4].

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

82 There has been a recent exponential growth in the number of RNA-Seq samples
83 made publicly available in cattle (Fig. S1a), but these data have never before been gathered
84 in one collection and processed uniformly. Here, we present a pipeline to uniformly integrate
85 11,642 public RNA-Seq datasets and identify both *cis*- and *trans*-expression and splicing
86 quantitative trait loci (eQTLs and sQTLs) for 24 important cattle tissues. The latter is
87 facilitated by calling variants directly from the RNA-Seq reads and imputing to sequence
88 level with data from the 1000 Bull Genomes Project[7], as previously done with human data[8].
89 Next, we conducted *in silico* analyses, annotating eQTLs and sQTLs with a variety of
90 publicly available omics data, including DNA methylation, chromatin state, and chromatin
91 conformation characteristics. Finally, we integrated gene expression results with a large

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

97 **Results**

98 **Data summary**

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

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

115 **Variation in gene expression across individuals and tissues**

116 As expected, the number of expressed genes (Transcripts per Kilobase Million, TPM >
117 0.1) increased with the number of clean reads. However, we observed a plateau at 50 million

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

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

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

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

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

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

150 **Tissue specificity of gene expression**

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

167 **Discovery of expression and splicing QTLs**

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

170 significant *cis*-eQTLs) ranged from 172 in ileum to 10,174 in blood, with 19,559 (83% of all
171 23,523 tested genes) eGenes in at least one tissue (Table S4). The number of sGenes (genes
172 with significant *cis*-sQTLs) ranged from four in the salivary gland to 7,913 in macrophages,
173 with 15,376 (70.8%) sGenes in at least one tissue (Table S4). Genes with no *cis*-eQTLs or -
174 sQTLs in any of the tissues were significantly enriched in hormone activity, regulation of
175 receptor activity, neuropeptide signaling pathway, and reproduction. In general, the larger the
176 number of samples, the larger the number of *cis*-e/sGenes detected (Pearson's $r = 0.85$ for
177 *cis*-eQTLs with $P < 1.32\times10^{-7}$, $r = 0.63$ for *cis*-sQTLs with $P < 9.1\times10^{-4}$), (Fig.3a-b). As
178 expected, with larger sample size, we had more power to detect *cis*-eQTLs with smaller
179 effect sizes (Fig. S7a-b). Consistent with findings in human[12], significant variants (eVariants)
180 centered around the transcript start sites (TSS) of the measured genes (Fig. S7c-d). Across 24
181 tissues, an average of 46% (range 25.5 - 76.6%) of eVariants were found around TSS ± 100
182 kb of target genes. In non-eGenes, there was also an enrichment of top SNPs around TSS,
183 suggesting a lack of power to detect such associations (Fig. S7c). Furthermore, we fine-
184 mapped eGenes to assess whether the identified signals were attributed to one or more causal
185 SNPs. We found that 46% (range 14.5 - 73.9%) of eGenes had more than one independent
186 SNPs associated with expression (Fig. 3c), indicating the widespread allele heterogeneity of
187 gene expression. In addition to discover the SNPs with the smallest effects within a locus, we
188 needed to adjust for the SNPs with the biggest effects (Fig. 3d), and the biggest effects
189 tended to be closer to the TSS (Fig. S7e). Allele-specific expression (ASE) analyses found
190 that associated genetic variants were significantly overrepresented in *cis*-eQTLs, and their
191 effect sizes were significantly correlated ($r = 0.25$, $P < 2.2\times10^{-16}$) (Fig. 3e, Fig. S7f),
192 verifying our *cis*-eQTLs. To further validate our results, we examined eQTLs that were
193 previously reported with independent datasets in cattle liver (n = 35), muscle (n = 41) and
194 intramuscular fat (n = 192)[13,14], and found that the effects of top variants in eGenes
195 previously reported were significantly correlated with the same genes in corresponding
196 tissues in the current study (Fig. S7g).

197 To investigate whether eQTLs are conserved among sub-species, we conducted *cis*-
198 eQTL mapping for muscle samples from *Bos indicus* (n = 51), *Bos taurus* (n = 505), and
199 their hybrids (n = 108) separately, yielding 86, 2,766, and 800 eGenes, respectively. The
200 effect sizes of SNPs with the largest effects were more similar across muscle tissues of
201 different sub-species than among tissues within the same species (Fig. 3f). About 60.0% of
202 eGene-eVariant pairs detected in the combined muscle samples were reproducible in separate
203 subspecies or the hybrids. Similarly, 72.3%, 79.2% and 56.4% of eGene-eVariant pairs
204 detected in *Bos indicus*, *Bos taurus*, and their hybrids could be validated in the combined
205 analyses of all samples (Fig. S7h). In addition, 51.3% of eGene-eVariant pairs detected in
206 *Bos indicus* were also significant in *Bos taurus*. For example, the expression of *NMRAL1*
207 was consistently and significantly regulated by a *cis*-eQTL among *Bos indicus*, *Bos taurus*,
208 and their hybrids (Fig. 3g). Combining the summary statistics of each sub-species in a meta-
209 analysis found that the results were significantly correlated (Pearson's $r = 0.56$, $P < 2.2\times10^{-}$
210 ^{16}) with those of the analyses of the combined samples (Fig. 3h). These results suggest that
211 eGene-eVariant associations identified in one sub-species are potentially transferable to other
212 sub-species, particularly for eQTLs with large effect sizes (Fig. 3i-j). Combining samples
213 from different sub-species and breeds will increase statistical power for detecting shared
214 eQTLs, but decreased power will almost certainly result from eQTL present in only some of
215 the sub-species. We also noted the variability in eQTLs among sub-species. For instance, the
216 expression of an immune-related gene, *SSNA1*, was regulated by a *cis*-eQTL in *Bos indicus*
217 ($P = 3.93\times10^{-11}$) but not in *Bos taurus* or the hybrids (Fig. 3k).

218 Patterns and biological mechanisms underlying tissue-specificity/similarity of *cis*-QTLs
219 provide insights into pleiotropic regulatory effects on phenotypes[4]. We therefore conducted a
220 non-model-based pairwise analysis using significant eGene-eVariant pairs in one tissue to
221 estimate the proportion of non-null associations of these pairs in another tissue (Methods).
222 We found that tissues with similar functions (e.g., hypothalamus and pituitary) tended to
223 have a higher proportion of non-null associations, suggesting a higher similarity in their

224 genetic regulatory effects on both gene expression and splicing (Fig. S8a, b). To investigate
225 whether similarity of genetic regulatory effects reflected the correlations of gene expression
226 across tissues, we estimated the proportion of non-null associations for values of tissue-pairs
227 based on the correlation of gene expression (Fig. S8c). These associations were significantly
228 correlated with those obtained from non-model-based pairwise analysis of *cis*-eQTLs (r =
229 0.48; $P < 2.2\times10^{-16}$) and *cis*-sQTLs (r = 0.61; $P < 2.2\times10^{-16}$). This result suggests that the
230 similarity of gene expression across tissues might be partially mediated through sharing of
231 genetic regulatory effects. Further analysis on expression of eGenes across tissues revealed
232 that effect sizes of eVariants decreased with the increasing number of tissues where target
233 eGenes were expressed (Pearson's r = -0.27, $P < 2.2\times10^{-16}$), indicating that, on average,
234 tissue-specific genes might have larger genetic regulatory effects than widely-expressed
235 genes (Fig. S8d).

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

251 **Functional annotation of QTLs**

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

258 Hypomethylated regions across 13 tissues were enriched in *cis*-e/sQTLs compared to
259 other tissues (Fig. 5d-e). For instance, the liver exhibited the highest enrichment of *cis*-
260 e/sQTL in liver-specific hypomethylated regions. Consistent with the brain having distinct
261 patterns of alternative splicing, related to the development of the nervous system[12], *cis*-
262 sQTLs but not *cis*-eQTLs in the hypothalamus and pituitary had the highest enrichments in
263 their specific hypomethylated regions (Fig. 5e).

264 Topologically associated domains (TADs) enable chromatin interactions between distal
265 regulatory regions and target promoters[15]. By examining Hi-C data of lung tissue in cattle,
266 we obtained TADs and significant Hi-C contacts, which were conserved across tissues[15]. By
267 comparing with random eGene-SNP pairs with matched distances, we observed significantly
268 higher percentages of eGene-eVariant pairs within TADs across the majority of tissues (Fig.
269 5f). For instance, *APCS* and its *cis*-eQTL peak (144 kb upstream of the TSS) were
270 encompassed by a TAD and linked by a significant Hi-C contact, which allowed the
271 regulation of its expression by distal (> 2kb from TSS) eVariants (Fig. 5g).

272 **eQTLs and complex trait associations**

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

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

283 We then detected 496 candidate gene-tissue pairs for 43 agronomic traits in cattle *via*
284 TWAS, representing 337 unique genes (Table S6). Out of 319 previously fine-mapped
285 genes[16,17], we verified 54, including linking expression of *DGAT1* in liver and mammary
286 gland, and expression of *MGST1* in milk cells, as well as expression of *CLN3* in liver to milk
287 yield (Fig. 6d). The expression of *ZNF613* in hypothalamus was the top significant
288 association for many reproduction and body conformation traits, including daughter-still-
289 birth and stature, supporting our previous finding that *ZNF613* was significantly associated
290 with gestation length might be through affecting embryonic development[18]. Overall, TWAS
291 enhanced our ability to detect causal genes and to better understand the biological
292 underpinnings of these traits.

293 **Discussion**

294 The cGTEx atlas represents the most comprehensive reference resource of the cattle
295 transcriptome to date. It provides a detailed characterization of genetic control of gene
296 regulation across 24 tissues. Compared to the human GTEx, we identified a large number of
297 *trans*-eQTLs even with modest sample sizes. This result is likely explained by the distinct
298 LD properties between the two species, as factors like domestication, breed formation, and
299 strong artificial selection for production traits (e.g., polygenic selection of milk yield) have
300 led to long-distance LD within chromosomes and LD even among non-syntenic
301 chromosomes. Reinforcing this hypothesis, we detected the largest number of *trans*-eQTLs
302 in the mammary gland, which might be the top tissue under selection in dairy cattle[19], even

303 with a relatively small sample size compared to other tissues.

304 Three types of QTLs (*cis*-e/sQTLs and *trans*-eQTLs) detected in this study provide a rich
305 set of candidate functional variants for agronomic traits in cattle, as we found that top GWAS
306 associations for a trait were significantly enriched with these regulatory QTLs in their
307 relevant target tissues. Our data will aid future colocalization analyses for significant GWAS
308 results and inform candidate variants/genes for functional follow-up. Further integration of
309 our QTLs with functional annotations of a range of tissues from the on-going Functional
310 Annotation of Animal Genomes (FAANG) project will provide valuable opportunities to
311 understand transcriptional/post-transcriptional regulatory mechanisms underpinning GWAS
312 hits for agronomic traits[20]. Our study will enable exploring the molecular mechanisms
313 underlying the extensive pleiotropic effects identified in livestock[21]. This information will
314 allow the understanding of mechanisms of response to intended selection as well as
315 disentangling unintended and unfavorable correlated responses to this same selection (e.g.
316 increasing mastitis or deteriorating fertility when selection for increased milk production).
317 This resource will assist in the development of genome selection methods to improve
318 targeted responses in animal health and wellbeing. Moreover, a better understanding of the
319 genetic architecture underpinning agronomic traits will benefit genetic improvement
320 programs by incorporating biological knowledge into genomic prediction models[7,22], and
321 may improve prediction accuracy across generations and populations.

322 **Online Methods**

323 **Quantification of gene expression**

324 We downloaded 11,642 RNA-Seq runs (by July, 2019) from SRA (n = 11,513,
325 https://www.ncbi.nlm.nih.gov/sra/) and BIGD databases (n = 129,
326 https://bigd.big.ac.cn/bioproject/). We merged multiple runs from single samples, yielding
327 8,536 unique RNA-Seq samples. We applied a stringent and uniform pipeline to filter and
328 analyze all the data. Briefly, we first removed adaptors and low quality reads using

329 Trimmomatic (v0.39)[23] with parameters: adapters/TruSeq3-SE.fa:2:30:10 LEADING:3
330 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. We filtered out samples with clean
331 read counts ≤ 500K, resulting in 7,680 samples, and mapped clean reads to the ARS-UCD1.2
332 cattle reference genome[24] using single or paired mapping modules of STAR (v2.7.0) with
333 parameters of outFilterMismatchNmax 3, outFilterMultimapNmax 10 and
334 outFilterScoreMinOverLread 0.66. We kept 7,264 samples with uniquely mapping rates ≥
335 60% (mean, 91.07%; range, 60.44%-100%; mapping details in Table S1). We then obtained
336 normalized expression (TPM) of 27,608 Ensembl (v96) annotated genes using Stringtie
337 (v2.1.1)[25], and extracted raw read counts of them with featureCounts (v1.5.2)[26] as well. We
338 finally clustered 7,264 samples based on $\log_2$(TPM +1) using a hierarchical clustering
339 method, implemented in R package *dendextend*, with distance = (1-*r*), where *r* is the Pearson
340 correlation coefficient. We excluded samples with obvious clustering errors (e.g., samples
341 labeled as liver that were not clustered with other liver samples), resulting in 7,180 samples
342 for subsequent analysis.

343 **Quantification of alternative splicing**

344 We used Leafcutter (v0.2.9)[27] to identify and quantify variable alternative splicing events
345 of genes by leveraging information of junction reads (i.e., reads spanning introns) that were
346 obtained from the STAR alignment. The Leafcutter enables the identification of splicing
347 events without relying on existing annotations that are typically incomplete, especially in the
348 setting of large genes or individual- and/or population-specific isoforms[27]. We first converted
349 bam files from STAR alignment into junction files using the script "bam2junc.sh", and then
350 performed intron clustering using the script "leafcutter_cluster.py" with default settings of 50
351 reads per cluster and a maximum intron length of 500 kb. We employed the
352 "prepare_genotype_table.py" script in Leafcutter to calculate intron excision ratios and to
353 remove introns used in less than 40% of individuals or with no variation. Ultimately, we
354 standardized and quantile normalized intron excision ratios as Percent Spliced-In (PSI)
355 values across samples. We clustered 7,180 samples based on PSI using the same method as

356 used in gene expression.

357 **Genotyping and imputation**

358 We called genotypes of known genomic variants in the 1000 Bull Genomes Projects[7] for
359 7,180 high-quality RNA-Seq samples individually, following the recommended best
360 practices pipeline in Genome Analysis Toolkit (GATK) (v4.0.8.1)[28] with default settings. We
361 filtered low quality SNPs using --filter-expression "FS > 30.0 & QD < 2.0". We then
362 imputed the filtered SNPs on autosomes to sequence level using Beagle (v5.1)[29] based on a
363 multiple-breed reference population consisted of 3,103 individuals from run7 of the 1000
364 Bull Genomes Project[7] and 207 public individuals from *Bos taurus* (n = 101), *Bos indicus*
365 (zebu, n = 20), and *Bos grunniens* (yak, n = 86) (Table S3). Finally, we obtained 6,123
366 samples that were genotyped and imputed successfully. We filtered out variants with MAF <
367 0.05 and dosage R-squared (DR^2) < 0.8, resulting in 3,824,444 SNPs used for QTL mapping.
368 To evaluate the accuracy of imputation, we called genotypes (~6 M SNPs) from WGS
369 (average read depth = 10×) of four Holstein individuals, which had RNA-Seq data from 13
370 tissues as well. We then measured the genotype concordance rates between WGS-SNPs and
371 RNA-Seq/imputed SNPs across 13 tissues. We then extracted 153,913 LD-independent SNPs
372 using plink (v1.90)[30] (--indep-pairwise 1000 5 0.2), and conducted PCA analysis for all
373 6,123 samples using these SNPs in EIGENSOFT (v7.2.1)[31]. We also calculated the identity-
374 by-state distance (IBS distance) among samples by using these independent SNPs to remove
375 duplicate individuals (i.e., IBS distance > 0.85). IBS distance = (IBS2 + 0.5*IBS1) / (IBS0 +
376 IBS1 + IBS2), where IBS0 is the number of IBS 0 non-missing variants, IBS1 is the number
377 of IBS 1 non-missing variants and IBS2 is the number of IBS 2 non-missing variants. We set
378 an IBS distance cutoff of 0.85 to deem two samples as duplicates and randomly kept one of
379 them. When conducting QTL mapping, we removed an average of 43 duplicate samples
380 within each tested tissue (ranging from 1 for salivary gland and leukocyte to 132 for muscle).

381 **Allele specific expression (ASE)**

382 We conducted ASE analysis using the GATK ASEReadCounter tool (v4.0.8.1) with the
383 following settings: --U ALLOW_N_CIGAR_READS -minDepth 10 –minMappingQuality
384 255 --minBaseQuality 10. SNPs for ASE detection fulfilled the following criteria:
385 heterozygous in at least five samples, at least 10 reads per allele, and at least 2% of all reads
386 supporting the minor allele. We then calculated a binominal *P*-value by comparing to the
387 expected ratio under the null hypothesis, followed by multiple-test correction with the
388 Benjamini–Hochberg (BH) approach. SNPs with FDR < 0.05 were considered as significant
389 ASE. We calculated the effect size of an ASE variant in a tissue as median log-ratio:

390 $$s(x, y) = median[log_2\left(\frac{x_i}{y_i}\right)]$$

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

394 **DNA methylation analysis of WGBS data**

395 We downloaded (in July, 2019) 144 publicly available WGBS cattle data from 21 distinct
396 tissues (Table S2). We first used FastQC (v0.11.2) and Trim Galore v0.4.0 (--max_n 15 --
397 quality 20 --length 20 -e 0.1) to determine read quality and to filter reads with low quality,
398 respectively. We then mapped clean reads to the same cattle reference genome (ARS-
399 UCD1.2) using Bismark software (v0.14.5)[32] with default parameters. After deduplication of
400 reads, we extracted methylation levels of cytosines using the *bismark_methylation_extractor*
401 (--ignore_r2 6) function. The coverages of all WGBS data were calculated on the basis of
402 clean reads, ranging from 5 to 47-fold coverage with an average of 27.6-fold coverage.
403 Ultimately, we kept CpG sites that were represented by at least five reads for subsequent
404 analyses. We clustered samples based on DNA methylation levels of shared CpGs using both
405 hierarchical clustering and t-SNE approaches[33].

406 **Identification of TAD and significant Hi-C contacts**

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

408 identified TADs and Hi-C contacts from Hi-C data from lung tissue in cattle that was
409 retrieved from NCBI Sequence Read Archive (SRA) under accessions: SRR5753600,
410 SRR5753603, and SRR5753606. We used Trim Galore (v0.4.0) to trim adapter sequences
411 and low-quality reads (--max_n 15 --quality 20 --length 20 -e 0.1), resulting ~820 million
412 clean reads. We then mapped clean reads to the cattle reference genome (ARS-UCD1.2[24])
413 using BWA[34]. We applied HiCExplorer v3.4.1[35] to build a Hi-C contact matrix with 10kb
414 resolution and identified TAD with hicFindTAD. We kept TADs with FDR less than 0.01 to
415 link eQTLs to eGenes. We further employed HiC-Pro (v2.11.4)[36] to call Hi-C contacts with
416 10 kb resolution from Hi-C data. Briefly, HiC-Pro aligned clean reads to the cattle reference
417 genome with Bowtie2 (v2.3.5)[34]. After building a contact matrix, HiC-Pro generated intra-
418 and inter-chromosomal maps and normalized them using the original ICE normalization
419 algorithm. We considered Hi-C contacts with FDR < 0.05 as significant.

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

421 To quantify tissue-specific expression of genes, we computed a *t*-statistics for each gene
422 in each of the 114 tissues. We grouped 114 tissues into 13 categories (Table S1). We scaled
423 the $\log_2$-transformed expression (i.e., $\log_2$TPM) of genes to have a mean of zero and variance
424 of one within each tissue. We then fitted a linear model as described in[37] for each gene in
425 each tissue using least-squares. When constructing the matrix of dummy variables (i.e.,
426 design matrix) for tissues, we denoted samples of the target tissue/cell type (e.g., CD4 cells)
427 as '1', while samples outside the target category (e.g., non-blood/immune tissues) as '-1'. We
428 excluded samples within the same category (e.g., CD8 cells and lymphocytes) to detect
429 genes with specific expression in each particular category, even if they were not specific to
430 the target tissue within this category. We obtained *t*-statistics for each gene to measure its
431 expression specificity in a given tissue. We considered the top 5% of genes ranked by largest
432 *t*-statistics as genes with high tissue-specific expression.

433 To detect tissue-specific alternative splicing, we used leafcutter to analyze the differential
434 intron excision by comparing the samples from target tissue to the remaining tissues[27].

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

437 For DNA methylation, we focused on DNA methylation levels of gene promoters (from
438 upstream 1500bp to downstream 500bp of TSS) and body regions (from TSS to TES), which
439 were calculated with a weighted methylation method using the roimethstat function in
440 MethPipe (v3.4.3)[38]. We computed a *t*-statistic for the promoter of each gene using the same
441 method as in tissue-specific expression analysis. We considered the bottom 5% of genes
442 ranked by *t*-statistics as genes with tissue-specific promoter hypomethylation. We also
443 detected tissue-specific methylation regions in a genome-wide mode using SMART2[39] with
444 parameters of -t DeNovoDMR -MR 0.5 -AG 1.0 -MS 0.5 -ED 0.2 -SM 0.6 -CD 500 -CN 5 -
445 SL 20 -PD 0.05 -PM 0.05.

446 **Covariate analysis for QTL discovery**

447 To account for hidden batch effects and other technical/biological sources of
448 transcriptome-wide variation in gene expression, we estimated latent covariates in each
449 tissue using the Probabilistic Estimation of Expression Residuals (PEER) method[40]. In each
450 tissue, we estimated 75 PEER factors first. The posterior variances of factor weights
451 dramatically decreased and reached or nearly reached plains when 10 PEER factors were
452 included (Fig. S6a). Therefore, we used 10 PEER covariates to account for effects of
453 confounding variables on gene expression in all following QTL analyses. For instance, the
454 variance of gene expression among samples in liver captured by 9 out of 10 PEER factors
455 were significantly (FDR < 0.05) correlated with known technical and biological covariates
456 like clean data size, mapping rate, project, breeds, sub-species, sex and age. (Fig. S6b). To
457 further control the effect of population structure on the discovery of QTLs, we included
458 genotype PCs based on sample size bins: three PCs for tissues with < 150 samples, five PCs
459 for tissues with ≥ 150 and < 250 samples, and ten PCs for tissues with ≥ 250 samples.

460 ***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 ≥ 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 FastQTL[41], 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 $\geqslant$ 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) tools[43]. We further applied the statistical fine-mapping method, dap-g[44], 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

488 detect representative/independent eQTLs for the target eGene. To analyze pairwise tissue
489 similarity in cis-eQTL, we calculated π_1 value, defined as proportion of true positive cis-
490 eQTLs identified in first tissue amongst all tested gene-variant pairs in second tissue, using
491 the Storey and Tibshirani qvalue approach, as described in[12].

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

493 Data from muscle samples were available from three sub-species: *Bos indicus* (n = 51),
494 *Bos taurus* (n = 505), and their crosses (n = 108). To explore the similarity and variability of
495 *cis*-eQTLs among sub-species, we associated expression of target genes with their local
496 SNPs (1Mb around TSS) using muscle samples from each of sub-species separately. We then
497 conducted a meta-analysis to integrate *cis*-eQTL results from three sub-species using the
498 Metal tool[45]. We obtained Z-scores (the sum of weighted effect sizes) of SNPs from the
499 meta-analysis. Weights were proportional to the square-root of the number of individuals in
500 each sub-species[45].

501 ***cis*-sQTL mapping**

502 In each of the 24 tissues, we applied a linear regression model, implemented in
503 FastQTL[41], to test for associations of genotypes (MAF > 1%) within 1 Mb up- and down-
504 stream of target intron clusters and their corresponding intron excision ratios. We used the
505 first five genotype PCs to account for the effect of ancestry, and 10 PEER factors to adjust
506 for the effect of unknown confounding variables. We applied the permutation pass mode
507 (1,000 permutations) in FastQTL[41] to obtain beta approximated permutation *p* values,
508 followed by multiple test correction with the FDR method. We considered sQTL-intron
509 cluster pairs with FDR < 0.05 as significant, and defined sGene as genes containing a
510 significant sQTL in any introns.

511 ***trans*-eQTL mapping**

512 We conducted *trans*-eQTLs for 13 tissues with at least 100 samples each. We filtered
513 genomic variants using a more stringent threshold than *cis*-eQTL mapping to partially

514 account for the reduction in statistical power. We obtained mappability of variants based on
515 k-mer length of 36 and 75 following the procedure described in
516 https://wiki.bits.vib.be/index.php/Create_a_mappability_track. We excluded any variants
517 within repeats (Repeatmasker and simple repeats), and further removed variants with
518 mappability < 1, based on k-mer length of 75. After filtering, we kept SNPs with MAF > 5%
519 and at least 10 minor alleles within each tissue for association testing. We mapped *trans*-
520 eQTLs for protein-coding genes with average mappability ≥ 0.8 based on k-mer length of 36
521 using MatrixQTL[46], by associating expression of target genes with genotypes on other
522 autosomal chromosomes, while adjusting for the same covariates as in *cis*-eQTL analysis.
523 We adjusted *P*-values for multiple testing using Benjamini-Hochberg method to obtain FDR.
524 We considered gene-variant pairs with FDR < 0.05 as significant.

525 **WGCNA co-expression network analysis and estimation of π_1 statistics**

526 We applied the Weighted Gene Co-Expression network (WGCNA) analysis [47] to obtain
527 gene co-expression network in each of 24 tissues used in eQTL mapping. We estimated the
528 sharing of *cis*-eQTLs and *cis*-sQTLs among tissues using the π_1 statistics, as described in
529 human GTEx Consortium (2015)[12]. Briefly, given a "discovery" tissue A and a "Validation"
530 tissue B, we extracted *p*-values for all significant *cis*-eQTLs in A from B, and then computed
531 a Storey's π_1 value based on these *P*-values. We also computed the π_1 statistics of tissue-pairs
532 based on gene expression correlation to quantify the sharing of gene expression among
533 tissues.

534 **TWAS analyses**

535 To associate gene expression in a tissue with complex traits, we conduced TWAS
536 analysis using S-PrediXcan[48] by prioritizing GWAS summary statistics for 43 agronomic
537 traits of economic importance in cattle, including reproduction (n = 11), production (milk-
538 relevant; n = 6), body type (n = 17), and health (immune/metabolic-relevant; n = 9). For
539 body conformation (type), reproduction, and production traits, we conducted a single-marker
540 GWAS by fitting a linear mixed model in 27,214 U.S. Holstein bulls as described

541 previously[16]. For health traits, we conducted GWAS using the same method in a subset
542 (ranging from 11,880 for hypocalcemia to 24,699 for livability) of the 27,214 available
543 bulls[17]. We constructed a Nested Cross Validated Elastic Net prediction model using
544 genotype and expression data. We included sub-species, 10 PEER factors and corresponding
545 genotype PCs (details in *cis*-eQTL mapping part) in the model to adjust for unknown
546 confounding variables and underlying population structure. For each trait, we conducted
547 TWAS in each of the same 24 tissues as in *cis*-eQTL mapping. We considered genes with
548 Bonferroni-corrected P (q-value) < 0.05 as significant. We visualized the Manhattan plots of
549 P-values of all tested genes using ggplot2 (v3.3.2) in R (v3.4.1).

550 **Other downstream bioinformatic analysis**

551 We used Genomic Association Tester (GATv1.3.4)[49] 10,000 permutations to estimate
552 the functional enrichment of QTLs in particular genomic regions, chromatin states,
553 methylation elements, and WGCNA co-expression modules. We considered enrichments with
554 FDR (adjusted P-values with Benjamini-Hochberg method) < 0.05 as significant. We used
555 the R package, ClusterProfiler[50], to annotate function of genes based on the Gene ontology
556 database from Bioconductor (org.Bt.eg.db v3.11.4). We considered GO terms with FDR <
557 0.05 as significant. We obtained complex traits/diseases that were associated with a
558 particular gene in humans using the Region PheWAS function in GeneATLAS database
559 (http://geneatlas.roslin.ed.ac.uk/) with set region of ±50kb and P-value threshold of 1×10^{-8}.

560 **Acknowledgements**

561 This work was supported in part by AFRI grant numbers 2013-67015-20951, 2016-
562 67015-24886, and 2019-67015-29321 from the USDA National Institute of Food and
563 Agriculture (NIFA) Animal Genome and Reproduction Programs and BARD grant number
564 US-4997-17 from the US-Israel Binational Agricultural Research and Development (BARD)
565 Fund. A. Tenesa acknowledged funding from the BBSRC through programme grants
566 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

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

594 **Completing interests statement**

595 The authors declare no competing interests.

596 **Data availability statement**

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

600 **References**

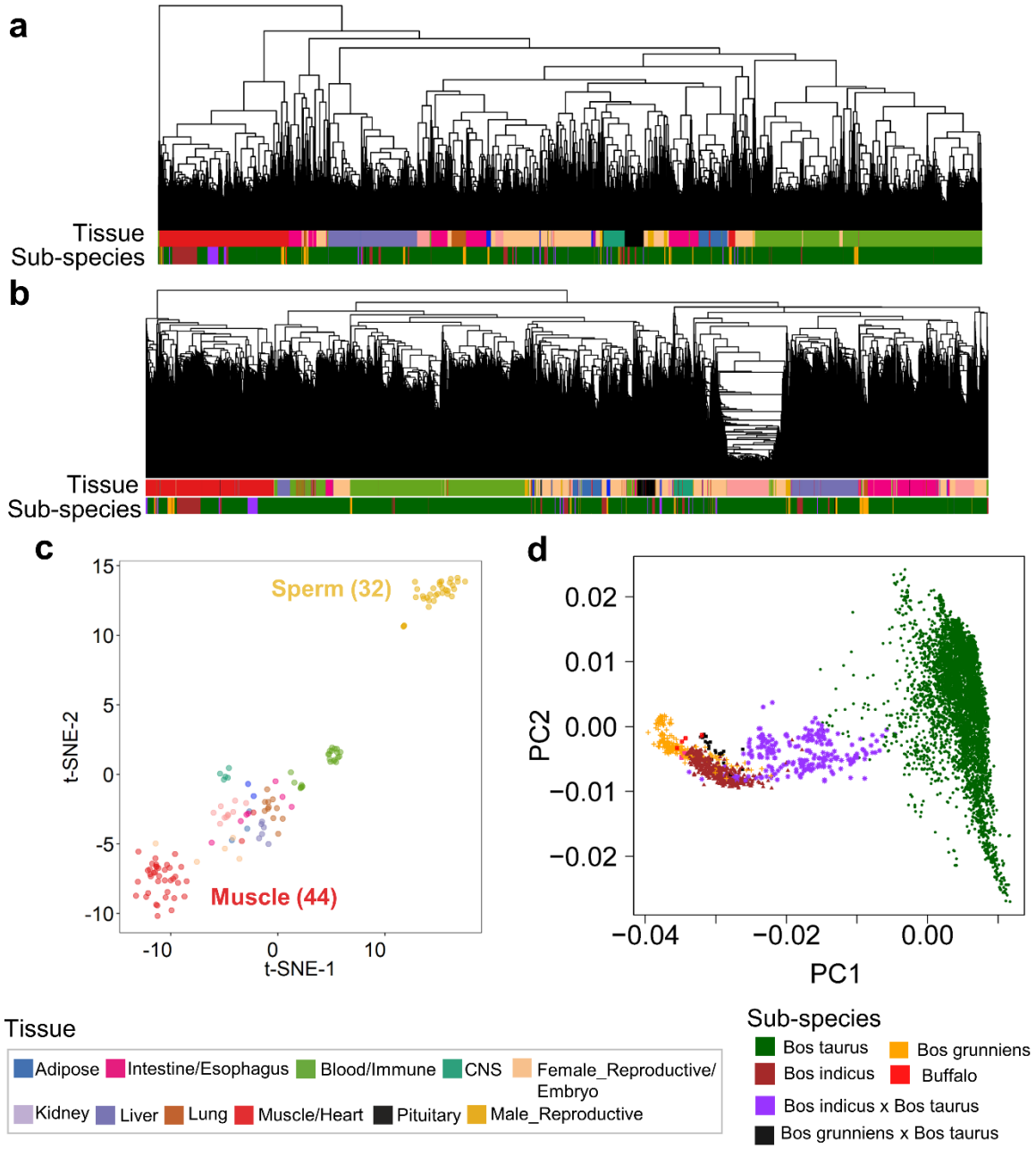
601 1. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide
602 association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**,
603 D1005-D1012 (2019).
604 2. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK
605 Biobank. *Nat Genet* **50**, 1593-1599 (2018).
606 3. Hu, Z.L., Park, C.A. & Reecy, J.M. Building a livestock genetic and genomic
607 information knowledgebase through integrative developments of Animal QTLdb and
608 CorrDB. *Nucleic Acids Res* **47**, D701-D710 (2019).
609 4. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across
610 human tissues. *Science* **369**, 1318-1330 (2020).
611 5. Fang, L. *et al.* Comprehensive analyses of 723 transcriptomes enhance genetic and
612 biological interpretations for complex traits in cattle. *Genome Res* **30**, 790-801 (2020).
613 6. Xiang, R. *et al.* Quantifying the contribution of sequence variants with regulatory and
614 evolutionary significance to 34 bovine complex traits. *Proc Natl Acad Sci U S A* **116**,
615 19398-19408 (2019).
616 7. Hayes, B.J. & Daetwyler, H.D. 1000 Bull Genomes Project to Map Simple and
617 Complex Genetic Traits in Cattle: Applications and Outcomes. *Annu Rev Anim Biosci*
618 **7**, 89-102 (2019).
619 8. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables
620 identification of genetic variants that affect gene-expression levels. *Genome Med* **7**, 30
621 (2015).
622 9. Didychuk, A.L., Butcher, S.E. & Brow, D.A. The life of U6 small nuclear RNA, from
623 cradle to grave. *RNA* **24**, 437-460 (2018).
624 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

666 10 1-11 10 33 (2013).
667 29. Browning, B.L., Zhou, Y. & Browning, S.R. A One-Penny Imputed Genome from
668 Next-Generation Reference Panels. *Am J Hum Genet* **103**, 338-348 (2018).
669 30. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and
670 richer datasets. *Gigascience* **4**, 7 (2015).
671 31. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS*
672 *Genet* **2**, e190 (2006).
673 32. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for
674 Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-2 (2011).
675 33. Maaten, L.v.d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning*
676 *research* **9**, 2579-2605 (2008).
677 34. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
678 transform. *Bioinformatics* **25**, 1754-60 (2009).
679 35. Ramirez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome
680 organization in flies. *Nat Commun* **9**, 189 (2018).
681 36. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.
682 *Genome Biol* **16**, 259 (2015).
683 37. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies
684 disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
685 38. Song, Q. *et al.* A reference methylome database and analysis pipeline to facilitate
686 integrative and comparative epigenomics. *PLoS One* **8**, e81148 (2013).
687 39. Liu, H. *et al.* Systematic identification and annotation of human methylation marks
688 based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific
689 hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res* **44**, 75-94
690 (2016).
691 40. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation
692 of expression residuals (PEER) to obtain increased power and interpretability of gene
693 expression analyses. *Nat Protoc* **7**, 500-7 (2012).
694 41. Ongen, H., Buil, A., Brown, A.A., Dermitzakis, E.T. & Delaneau, O. Fast and efficient
695 QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-85
696 (2016).
697 42. Consortium, G.T. *et al.* Genetic effects on gene expression across human tissues.
698 *Nature* **550**, 204-213 (2017).
699 43. Mohammadi, P., Castel, S.E., Brown, A.A. & Lappalainen, T. Quantifying the
700 regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome*
701 *Res* **27**, 1872-1884 (2017).
702 44. Wen, X., Lee, Y., Luca, F. & Pique-Regi, R. Efficient Integrative Multi-SNP
703 Association Analysis via Deterministic Approximation of Posteriors. *Am J Hum Genet*
704 **98**, 1114-1129 (2016).
705 45. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of
706 genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

46. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).
47. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
48. 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).
49. Heger, A., Webber, C., Goodson, M., Ponting, C.P. & Lunter, G. GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046-8 (2013).
50. Yu, G., Wang, L.G., Han, Y. & He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-7 (2012).

720 **Figure Legends**



721

722 **Fig. 1. Hierarchical clustering and principal component analysis of samples based on**
723 **gene expression levels, alternative splicing, genotypes and DNA methylation levels. (a)**
724 Sample hierarchical clustering based on gene expression levels of all transcribed genes
725 (TPM > 0.1). **(b)** Sample hierarchical clustering based on alternative splicing value (Percent
726 Spliced-In, PSI) of spliced introns. **(c)** Principal component analysis of samples based on
727 imputed genotypes. **(d)** Sample clustering using *t*-distributed SNE coordinates based on
728 DNA methylation levels of CpG sites (⩾ five reads).

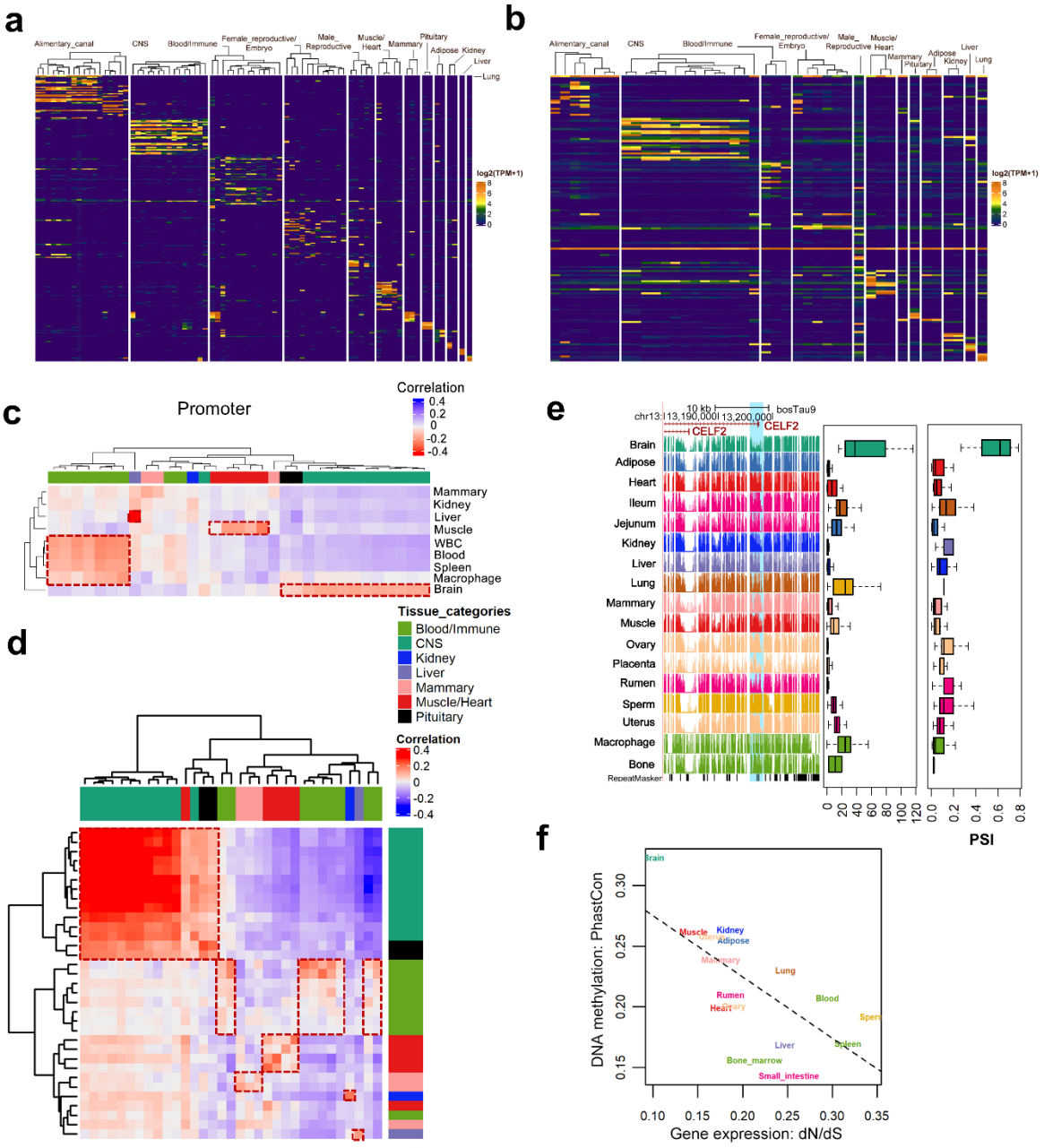


Fig. 2. Tissue-specificity of gene expression, alternative splicing and DNA methylation. **(a)** Median expression profile of tissue-specific expressed genes across tissues. Each row represents a gene, and each column represents an individual tissue. Color represents median $\log_2$(TPM+1) of genes among 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 = 3.6×10^{-3}).

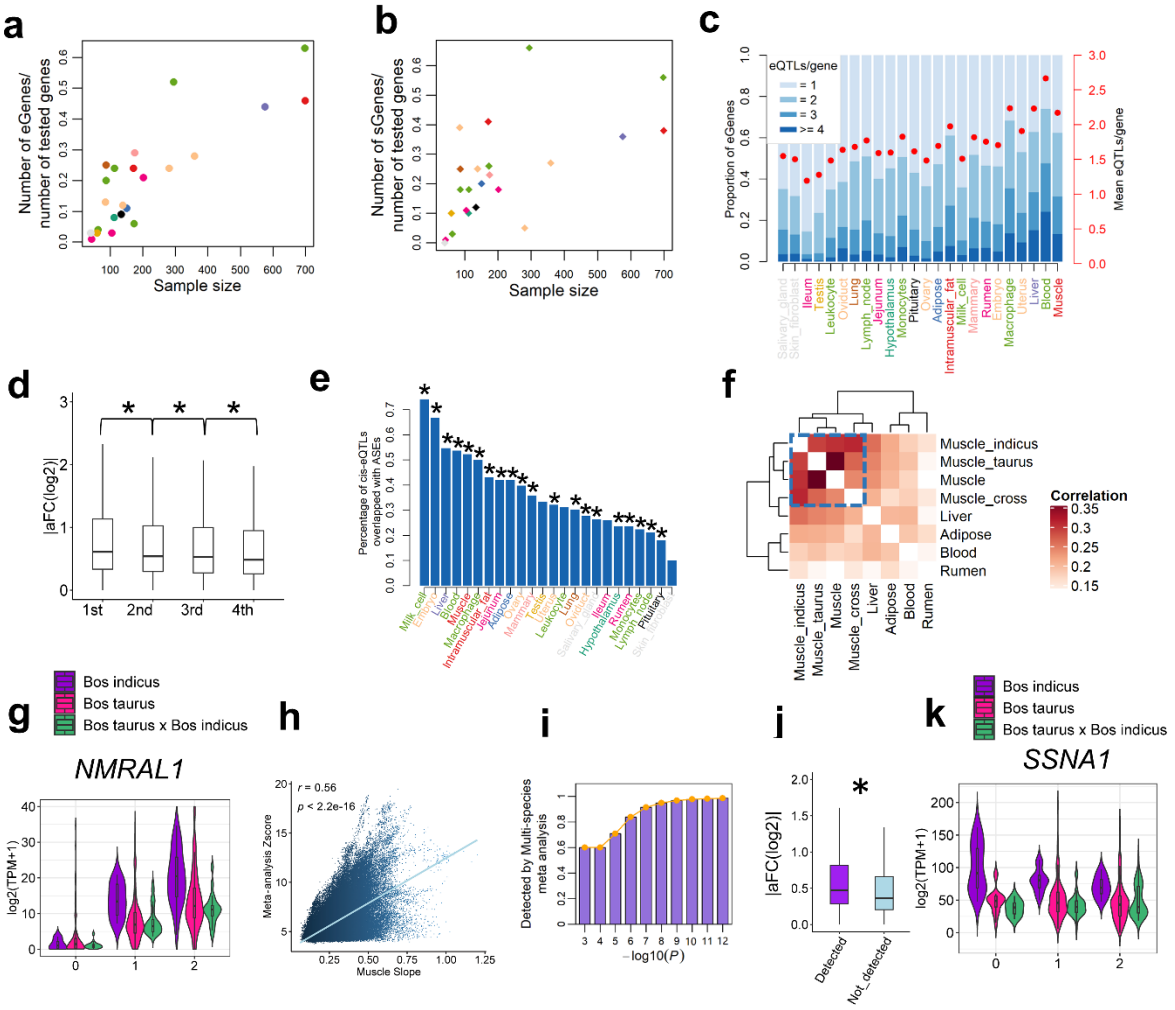


Fig. 3. Discovery and characterization of *cis*-eQTLs and *cis*-sQTLs. (a) Relationship between the percentage of eGenes over all tested genes and sample size (Pearson r = 0.83; P < 2.2×10^{-16}) across 24 tissues. **(b)** Relationship between the percentage of sGenes over tested genes and sample size (Pearson r = 0.63; P < 2.2×10^{-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 ($P < 1\times10^{-14}$, 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(log2)|) 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 *SSNA1*. 0, 1 and 2 indicate genotypes GG, AG, and AA, respectively.

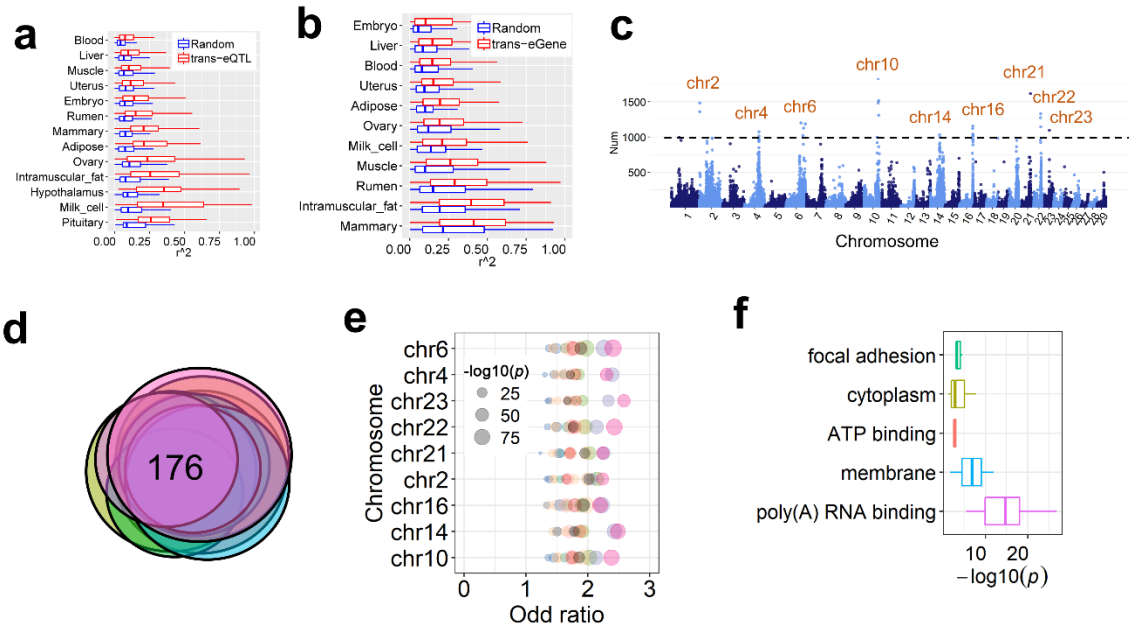


Fig. 4. Discovery and characterization of *trans*-eQTLs. (a) Comparison of linkage disequilibrium (r^2) between *cis*-eQTLs *vs. trans*-eQTLs with shared genes and *cis*-eQTLs *vs.* random SNPs on matched chromosomes. LD levels between the *cis*-eQTLs and their mediated *trans*-eQTLs are significantly ($P < 0.05$, 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

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

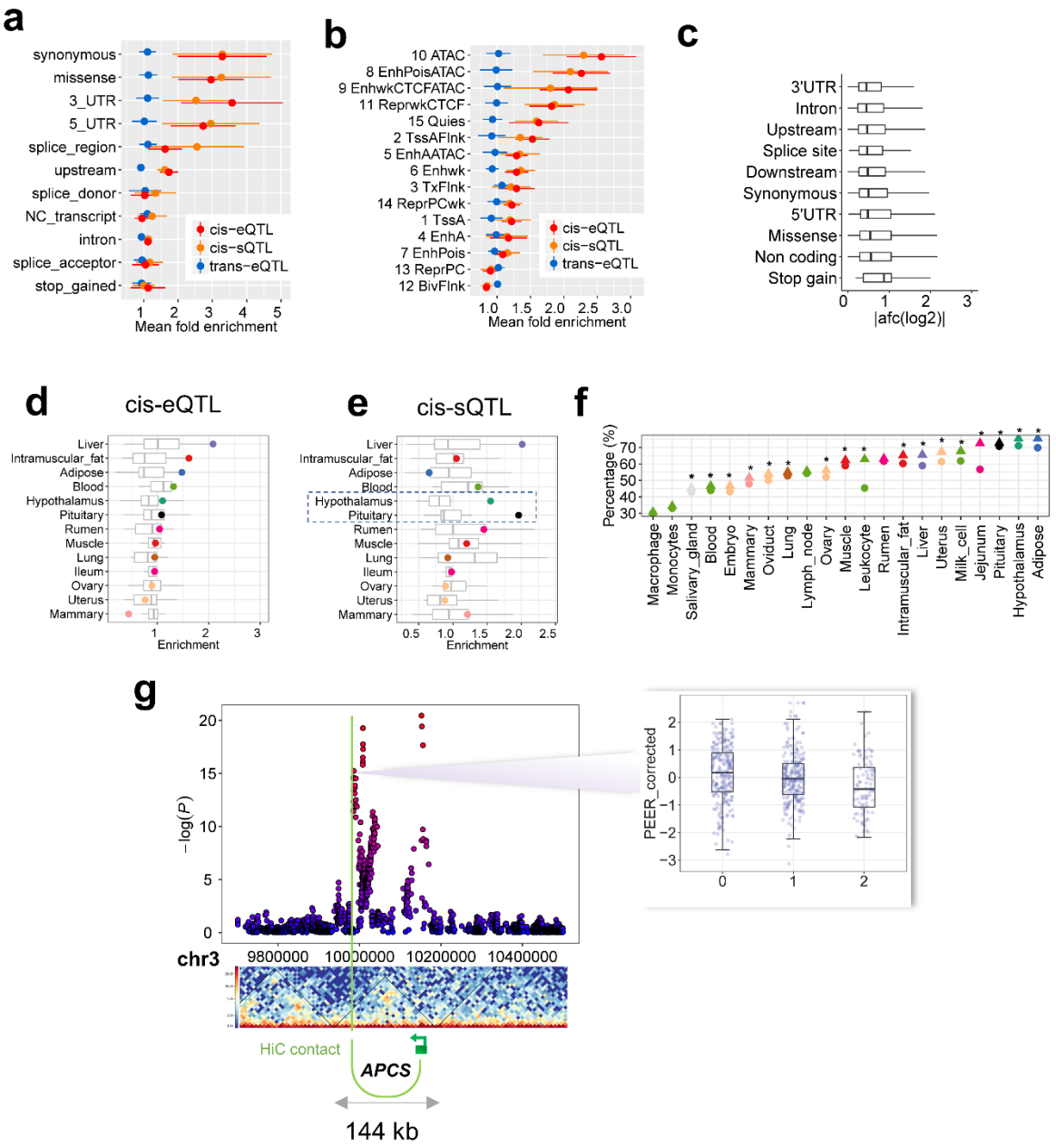


Fig. 5. Functional annotation of QTLs. (a) Enrichment of *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in sequence ontology. **(b)** Enrichment of *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in

787 15 chromatin states predicted from Rumen tissue in Holstein. **(c)** Effect sizes (measured as
788 log_2aFC) 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.

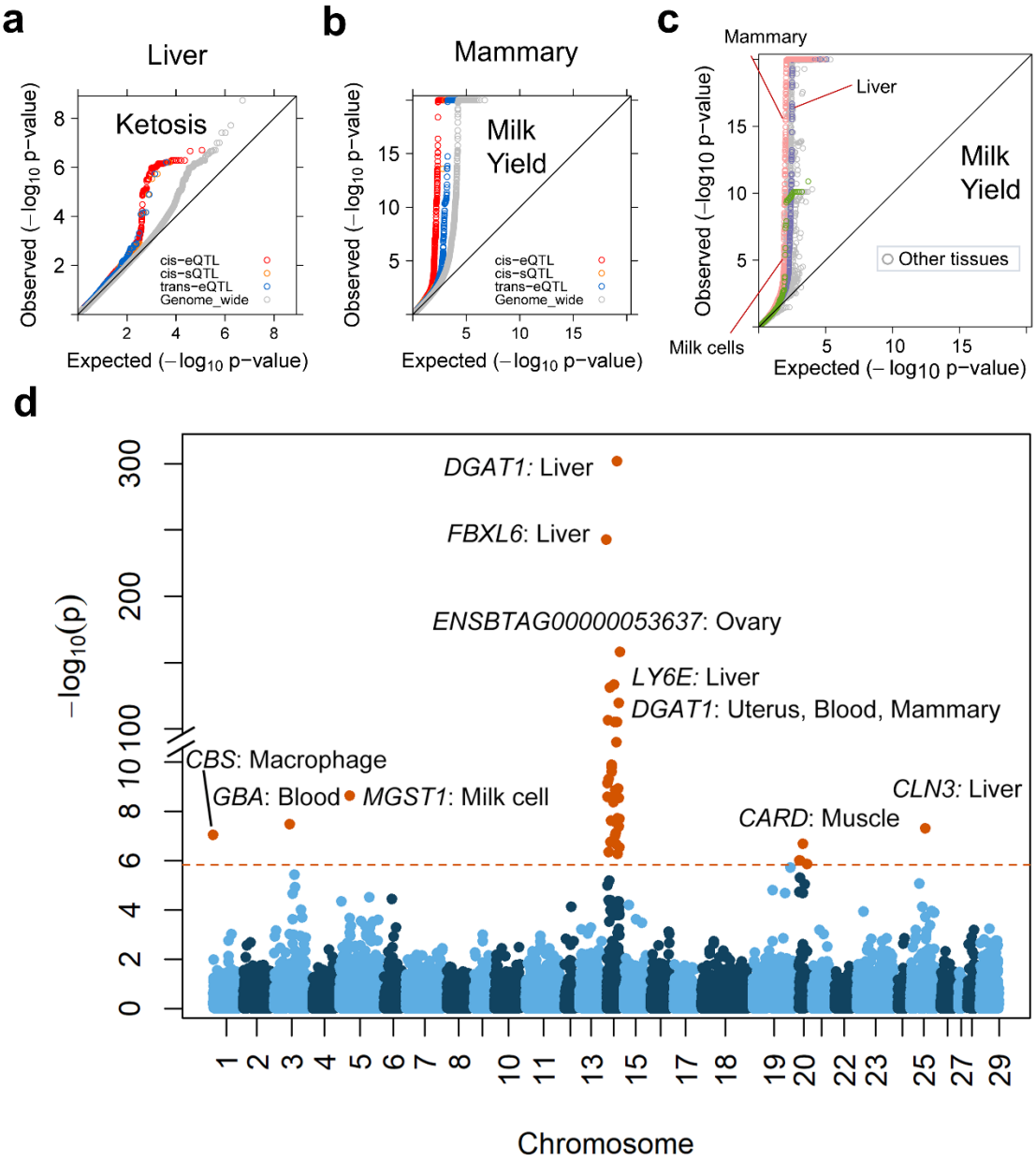


Fig. 6. Relationship between QTLs and livestock traits of economic value. (a) *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in liver show higher enrichments for SNPs associated with ketosis compared to genome-wide SNPs (shown in grey). **(b)** *cis*-eQTLs, *cis*-sQTLs and *trans*-eQTLs in mammary gland 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.