An integrated multi-tissue atlas of epigenomic landscapes and regulatory elements in the bovine genome

Dailu Guan[1,†], Jennifer Jessica Bruscadin[1,2†], Wenjing Yang[1,†], Claire Prowse-Wilkins[1,†], Junjian Wang[3], Bruna Petry[4], Houcheng Li[5], Huicong Zhang[5], Xiaoqin (Shally) Xu[1], Ying Wang[1], Zhangyuan Pan[1], Yuri Utsunomiya[6], Gordon Murdoch[7], Kimberley Davenport[7], Guosong Wang[8], Yue Xing[9], Christian Maltecca[3], Li Ma[10], Gonzalo Rincon[11], Sarah Corum[11], Pengcheng Lyu[12] Joseph (Nader) Deeb[6], Zhiping Weng[13], Jicai Jiang[3], Lingzhao Fang[5], Honglin Jiang[12], Brenda M. Murdoch[14], Stephanie D. McKay[15], Monique Rijnkels[8], Claire A. Gill[8], Timothy P. L. Smith[16], Zhihua Jiang[17], Wansheng Liu[18], Jim Reecy[4], Juan F. Mendrano[1], James E. Koltes[4], Pablo Ross[1], Huaijun Zhou[1]

[1]Department of Animal Science, University of California, Davis, CA, 95616, USA
[2]Center of Biological and Health Sciences, Federal University of Sao Carlos, Sao Carlos, Brazil
[3]Department of Animal Science, North Carolina State University, Raleigh, NC 27695, United States
[4]Department of Animal Science, Iowa State University, Ames, Iowa 50011, USA
[5]Center for Quantitative Genetics and Genomics (QGG), Aarhus University, Aarhus, Denmark
[6]STgenetics, 22515 Highway 6 South, Navasota, Texas 77868
[7]Department of Animal, Veterinary & Food Sciences, University of Idaho, Moscow, ID, USA
[8]Department of Veterinary Integrative Biosciences, School of Veterinary Medicine, Texas A&M University, College Station, TX, 77843, USA
[9]Department of Animal Science, Texas A&M University, College Station, TX 77843, USA
[10]Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA
[11]Zoetis Inc.
[12]School of Animal Sciences, Virginia Tech, Blacksburg, VA 24060, USA
[13]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA
[14]Department of Animal, Veterinary & Food Sciences, University of Idaho, Moscow, ID, USA
[15]Division of Animal Sciences, University of Missouri, Columbia, MO, 65211, USA
[16]USDA, ARS, US Meat Animal Research Center, Clay Center, Nebraska, USA
[17]Department of Animal Sciences and Center for Reproductive Biology, Washington State University, Pullman WA 99164-7620, USA.
[18]Department of Animal Science, Center for Reproductive Biology and Health (CRBH), College of Agricultural Sciences, The Pennsylvania State University, University Park, PA, USA

†These authors contributed equally.

Corresponding authors*:

Prof. Huaijun Zhou: Department of Animal Science, University of California-Davis, Davis, CA, 95616, USA

E-mail: hzhou@ucdavis.edu

Emails for all authors:
dguan@ucdavis.edu;
jennifer.j.bruscadin@gmail.com;
wenyang@ucdavis.edu;
c.prowsewilkins@gmail.com;
bpetry@iastate.edu;
jwang259@ncsu.edu;
houcheng.li@qgg.au.dk;
huicongz@qgg.au.dk;
xqxu@ucdavis.edu;
bpetry@iastate.edu;
yuri.utsunomiya@stgen.com;
ucywang@ucdavis.edu;
zhypan01@163.com;
yuri.utsunomiya@stgen.com
gordon.murdoch@wsu.edu
Sarah.Corum@zoetis.com;
Pengcheng@vt.edu;
joseph.deeb@stgen.com
zhipingweng@gmail.com
jicai_jiang@ncsu.edu;
lingzhao.fang@qgg.au.dk;
hojiang@vt.edu;
bmurdoch@uidaho.edu;
stephanie.mckay@uvm.edu;
mrijnkels@cvm.tamu.edu;
clare.gill@ag.tamu.edu;
tim.smith2@usda.gov;
jiangz@wsu.edu;
wul12@psu.edu;
jreecy@iastate.edu;
jekoltes@iastate.edu;
pross@ucdavis.edu;

77 hzhou@ucdavis.edu

Abstract

Deciphering the regulatory grammar of the genome is essential to understand the genetic and molecular architecture of complex traits, as most trait-associated variants lie in non-coding regions. Yet, functional annotation of the bovine genome remains limited, hindering our ability to unravel the mechanisms underpinning complex traits of economic and ecological importance in cattle. Here, we present a comprehensive epigenetic atlas comprising 1,147 genome-wide epigenetic profiles, including chromatin accessibility, six histone modifications, CTCF transcription factor binding sites, DNA methylation, chromatin conformation, and transcriptomes across 53 adult tissues, 6 fetal tissues, and 7 primary cell types. This atlas covers 99.8% of the bovine genome with at least one epigenomic feature, and annotates 45% of the genome as putative regulatory elements exhibiting tissue- or cell-specific regulatory activity. Leveraging sequence-to-function deep learning models, we discovered 301 sequence motifs and predicted the functional impact of genetic variants through *in silico* mutagenesis, thereby facilitating the decoding of the regulatory syntax of the cattle genome and fine-mapping of GWAS loci for 22 complex traits. Cross-species analysis further revealed evolutionarily conserved features of regulatory architecture and provided evolutionary insights into complex traits and diseases in humans. Together, this atlas offers a foundational resource for advancing cattle functional genomics, sustainable breeding, and studies of regulatory evolution.

Introduction

Cattle are a key agricultural species, providing meat, milk, and leather, while also serving as a valuable model for studying embryo development, epigenetics and imprinting, immunology, and the genetics of complex traits[1,2]. Advances in high-throughput sequencing and genotyping technologies have substantially enhanced our understanding of the genetic architecture of economically important traits in cattle and accelerated the pace of genetic improvement[3]. For instance, incorporation of single nucleotide polymorphisms (SNPs) into dairy breeding programs has doubled the rate of genetic gain for key production traits[4]. As of the latest update, the CattleQTLdb[5] catalogs approximately 193,453 quantitative trait loci (QTL), the majority of which reside in non-coding genomic regions[6]. This strongly suggests that many QTL act through regulatory mechanisms, modulating transcriptional activity of nearby or distal gene regulatory elements[7]. Comprehensive annotation of the non-coding bovine genome is therefore essential to uncover mechanistic insights into GWAS signals and pinpoint causal variants. Such insights may ultimately enable precision breeding strategies and the development of "ultimate genotypes"[8] through genome editing[9] and synthetic biology[10], while further enhancing its utility as a biomedical model.

Initial efforts to characterize bovine regulatory landscape have provided important foundations. For example, Kern *et al.*[11] performed genome-wide annotation of regulatory elements by analyzing chromatin immunoprecipitation followed by sequencing (ChIP-seq) for four histone modifications (H3K4me3, H3K27ac, H3K4me1, H3K27me3) and the DNA-binding protein CTCF, as well as Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), across eight core tissues (cortex, cerebellum, hypothalamus, liver, lung, spleen, muscle, adipose). Subsequent studies extended these annotations to rumen tissues [12,13] and rumen epithelial primary cells[12], and incorporated additional assays such as Cap Analysis Gene Expression (CAGE) sequencing[14], as well as broader genomic feature mapping[15]. Although these studies parallel large-scale efforts like ENCODE[16] and Roadmap Epigenomics[17] in humans towards annotating the non-coding genome, the bovine non-coding genome remains incompletely annotated. A systematic, large-scale effort spanning diverse tissues, cell types, and developmental stages (hereafter referred to as "tissue contexts") is essential to fully characterize context-specific regulatory mechanisms.

Complementary transcriptomic resources, including the Cattle Genotype-Tissue Expression (CattleGTEx)[18,19] and large-scale allele-specific (AS) analysis[20], have identified millions of regulatory variants influencing gene expression and splicing. Yet, these resources have not yet advanced to systematic identification of causal regulatory variants underlying GWAS associations of complex traits in cattle. Meanwhile, the emergence of sequence-to-function (S2F) deep learning models offers powerful approaches to decode the regulatory syntax of the genome and to prioritize functional variants. Tools like SpliceAI[21] and Pangolin[22] have demonstrated utility in predicting splicing effects in cattle[23], but broader application of S2F models for interpreting regulatory elements and context-specific variants remains largely unexplored.

In this study, we constructed a comprehensive bovine epigenomic atlas by integrating 540 newly generated datasets from 29 previously unprofiled tissues and assays with public available data[12,13,15,24,25]. In total, our resource comprises 1,147 genome-wide epigenomic profiles including 158 RNA-seq, 204 ATAC-seq, 91 whole genome bisulfite sequencing (WGBS), 682 ChIP-seq (six histone marks and CTCF), and 12 chromatin conformation (Hi-C) datasets. We used this data atlas to annotate chromatin states, link enhancers to their target genes, decode

regulatory syntax via deep learning approaches, fine-map GWAS loci, and perform multi-tissue cross-species comparisons (**Fig. 1a**). The resulting resource provides a powerful foundation for mechanistic dissection of complex traits, prioritization of functional variants, and comparative regulatory biology.

Results

Near-complete bovine epigenomic landscape

We have compiled a comprehensive atlas of 1,147 genome-wide epigenomic maps through uniform processing and quality control (**Fig. 1a-b**, **Supplementary Figures S1-S3**, **Table S1-S4**, see **Methods**). These include 158 RNA-seq maps to quantify gene expression (**Supplementary Figure S4, Table S1**), 204 ATAC-seq datasets to profile open chromatin regions (OCRs) (**Supplementary Figure S5-S6**, **Table S2**), 91 WGBS datasets to assess DNA methylation (**Supplementary Figure S7-S8**, **Table S3**), and 682 ChIP-seq datasets capturing histone modifications and transcription factor (TF) binding (**Table S4**). The ChIP-seq datasets encompass H3K4me1 (n = 131), marking enhancer regions; H3K4me3 (n = 132), associated with promoter activity; H3K27ac (n = 125), indicative of active enhancer and promoter regions; H3K27me3 (n = 123), marking Polycomb-mediated repression; H3K9me3 (n = 20), linked to heterochromatin; H3K36me3 (n = 27), associated with actively transcribed regions; and CTCF (n = 124), marking insulator elements[26] (**Supplementary Figure S9**). Among these, 540 ChIP-seq datasets (47.5%) from 29 tissues were newly generated in this study, and 75 datasets were computationally imputed using the ChromImpute pipeline[27], including H3K4me1 (n = 11), H3K4me3 (n = 10), H3K27ac (n = 12), H3K27me3 (n = 22), and CTCF (n = 20), with genome-wide Pearson correlation ranging from 0.56 (H3K27me3) to 0.91 (H3K4me3), comparable to those observed in human datasets[27] (**Extended data Fig. 1, Supplementary Figure S10-S11**). Peak calling from experimental ChIP-seq datasets yielded 895,108 H3K4me1 (52.7% of the autosomal genome), 737,720 H3K4me3 (18.2%), 802,966 H3K27ac (26.9%), 1,038,083 H3K27me3 (47.8%), 442,773 H3K9me3 (7.2%), 567,480 H3K36me3 (18.4%), 831,902 CTCF (19.8%) non-overlapping peaks. ATAC-seq analysis identified 953,179 non-overlapping open chromatin regions (OCRs) covering 21% of the genome, while WGBS profiling identified approximately 1.8 million methylation blocks (i.e. regions with ≥4 continuous CpGs), covering 73.4% of the genome. Additionally, we uniformly processed 12 public Hi-C datasets from 3 tissue contexts, yielding on average 4,979 topologically associating domains (TADs) per sample at 10 kb resolution (**Table S5**). Altogether, these datasets cover 99.8% of the bovine genome with at least one epigenomic feature (71% when excluding DNA methylation and TADs), providing a comprehensive resource for exploring the regulatory landscape of the cattle genome.

Nearly half of the bovine genome annotated as regulatory elements

Leveraging the atlas-level maps above, we defined a 10-state "core" chromatin model across 72 complete bovine epigenomes representing 33 tissue contexts (**Fig. 1c**) (**Fig. 2a**, **Supplementary Figure S12**). Here, a "complete epigenome" is defined as a biosample with all four core histone marks (i.e., H3K4me1, H3K4me3, H3K27ac, and H3K27me3). These states included three promoter-like states (PLS): TssA (active transcription start site [TSS], covering 2.2% of the genome across tissue context), TssFlnk (flanking TSS, 3%), and TssWk (weak flanking TSS,

3.5%); three enhancer-like states (ELS): EnhA (active enhancer, 7.6%), EnhWk (weak enhancer, 11.7%), and EnhPos (poised enhancer, 12.5%); two bivalent states: BivTss (bivalent TSS, 0.98%) and BivEnh (bivalent enhancer, 0.68%); as well as ReprPC (repressed Polycomb, 4.6%) and Quies (quiescent/low signal) **(Fig. 2a)**. These chromatin states, as expected[26,28,29], exhibited distinct patterns of gene expression (**Extended data Fig. 2a**), chromatin accessibility (**Extended data Fig. 1a**), DNA methylation (**Extended data Fig. 2b**), evolutionary conservation (**Fig. 2a, Extended data Fig. 2d, Supplementary Figure S13**), and transposable elements enrichment (**Supplementary Figure S14-15**). Of note, we found ELS from fetal epigenomes were significantly enriched for CR1 and L2 subfamilies (**Supplementary Figure S15**), and a high number of ELS overlapped with LINE (e.g. L1 and L2) and SINE subfamilies (**Supplementary Figure S16**).

Compared to the previous annotation of eight core tissues by Kern et al[11], our atlas incorporates 25 additional tissues (16 adult, 5 fetal and 4 primary cell types), resulting in an additional 106 Mb of annotated sequence (19% increase). For example, we annotated 146,445 non-overlapping EnhA, of which 66% were novel[11] (**Fig. 2b**). Across 65 complete bovine epigenomes with two additional marks (OCRs, CTCF), we further constructed a 15-state "expanded" model. This model introduced additional states such as TssFlnD (downstream flanking TSS), EnhAHet (active enhancer lacking ATAC signal), EnhCtcf (enhancer with CTCF), CtcfIsl (CTCF island), AtacIsl (ATAC island), and BivTssCtcf (bivalent TSS with CTCF) (**Extended data Fig. 3**).

To link enhancers to their target genes, moreover, we employed the Activity-by-Contact (ABC) approach[30], and successfully linked 25,567 unique enhancers to 26,100 genes, resulting in a total of 182,596 enhancer-gene pairs (**Fig. 2c-e, Supplementary Figure S17**). On average, each enhancer was associated with 7.1 genes (**Fig. 2c**), and each gene with 7 enhancers (**Fig. 2d**), the mean enhancer–gene distance was 168 kb (**Fig. 2e**). Furthermore, using ROSS software[31], we predicted an average of 1,242 super-enhancers across 32 tissues (covering 18.6% of the genome), ranging from 6 in liver to 3,128 in cerebral cortex (**Supplementary Figure S18-19**). Altogether, 45% of the bovine genome (excluding Quies) was annotated as regulatory elements, representing the most comprehensive regulatory annotation to date in cattle.

Dynamic epigenomic signatures across tissues and cell types

To investigate context-specific regulation of gene expression, we systematically analyzed chromatin state dynamics across tissues, developmental stages, and primary cell types. Tissue-specific states were first identified (**Supplementary Figure S20** and **Table S6)**. Focusing on active enhancer (EnhA), we observed the highest number of tissue-specific EnhA elements in muscle tissues, including muscle, heart, fetal heart, fetal muscle, and myoblasts (**Table S6**). In contrast, digestive tissues such as omasum, reticulum, and duodenum (excluding rumen) showed relatively few unique EnhA elements (**Fig. 2f**). These tissue-specific EnhA elements were strongly enriched for biologically relevant functional terms (**Supplementary Figure S21, Table S7**) and TF motifs[32] (**Fig. 2g, Table S8**). For instance, MYOG, MYF5, and MYOD motifs were enriched in muscle tissues, NEUROG2, NEUROD1, and NF1 in central nervous system (CNS) tissues, JUN, FOS and MYB73 in myoblasts (**Fig. 2g**), consistent with their established roles in cell proliferation and lineage-specific differentiation[33,34]. Comparisons between adult, fetal, and primary cells states, revealed additional layers of regulatory dynamics (**Supplementary Figure S22-24, Table S9-S10**). As illustrated in **Fig. 2h**, we identified 27,255 adult-specific (30%),

13,425 fetal-specific (15%), and 29,946 myoblast-specific (34%) EnhA elements. A representative example involves the myogenic regulators *MYF5* and *MYF6*. While *MYF6* was highly expressed in adult and fetal muscle, but not in myoblasts, *MYF5* shows high expressions in fetal muscle and myoblasts, but not in adult muscle (**Fig. 2i, Supplementary Figure S24**). Chromatin accessibility data mirrored these expression patterns, with myoblasts- and fetal muscle-specific enhancers flanking the *MYF5* locus (**Fig. 2i**). These findings align with the established role of *MYF5* in early myogenic differentiation and *MYF6* in terminal myogenic differentiation[35], mediated by a combinatorial network of enhancer elements during development[36].

We next explored DNA methylation dynamics using 91 newly generated WGBS datasets (**Supplementary Figure S7-S8**), identifying 207,436 differentially methylated blocks (DMBs) (see **Methods**, **Table S11-S12**) across approximately 1.8 million tested regions (**Table S11**). Fetal liver exhibited the highest number of DMBs (196,231). To construct a bovine tissue-specific hypomethylated atlas, we selected the top 50 differentially unmethylated regions per tissue (for tissues with >50 DMBs), resulting in a total of 1,120 DMBs (**Extended Data Fig. 4**). These DMBs exhibited low methylation level (average 18%) and strong co-enrichment with chromatin accessibility and histone modifications, exemplified by the cerebellum (**Extended Data Fig. 5a**). Motif enrichment analysis using HOMER[32] further revealed tissue-relevant TFs (**Table S13**, **Extended Data Fig. 5b**), such as *NEUROD1* in cerebellum, *MYB81* in muscle, and *MYF5* in myoblast (**Extended Data Fig. 5b**). Notably, *MYF5* encodes a critical regulator of skeletal myogenesis[37]. The tissue-specific patterns for chromatin accessibility are shown in **Supplementary Fig. 25,** and for gene expression in **Supplementary Fig. 26.** Finally, integration of tissue-specific OCRs with GWAS data revealed significant enrichment of complex trait loci in context-matched regulatory landscapes (**Extended Data Fig. 6**), underscoring the functional relevance of these dynamic epigenomic signatures.

Deep learning reveals regulatory sequence syntax across tissue contexts

Gene expression is initiated and regulated through the binding of TF to specific DNA sequence, known as TF binding site (TFBS)[38]. Recent advances in S2F deep learning models have demonstrated strong performance in decoding regulatory syntax of DNA sequences and predicting TFBS[7,39]. To explore the sequence-level grammar of regulatory elements in cattle, we trained ChromBPNet[40], a convolutional neural network (CNN), to model the relationship between genomic sequence and chromatin accessibility across 55 tissue contexts. For each tissue, model performance, evaluated as the Pearson correlation between predicted and observed log-transformed ATAC-seq read counts on held-out peaks (**Methods**), achieved a median of 0.76 (**Extended Data Fig. 7a, Table S14**), comparable to reports in human datasets[40,41]. Importantly, Pearson correlations were stable across distances to the nearest transcription start site (TSS) (**Extended Data Fig. 7b**). To assess generalizability, we applied the models to independent ATAC-seq data from Yuan et al.[24] across 19 overlapping tissues. For example, in abomasum, our model achieved a Pearson correlation of 0.84 on our datasets (**Extended Data Fig. 7c**) and 0.75 on Yuan et al.'s dataset[24] (**Extended Data Fig. 7d**), with minor differences potentially reflecting breed or age effects[24]. A representative locus at *DGAT1* (chromosome 14: 603,067-604,765) across tissues in both datasets, achieved a Pearson correlation of 0.69, with bone marrow showing the highest chromatin accessibility (**Extended Data Fig. 7e**). Additionally, the model could also predict tissue-specific accessibility peaks, as demonstrated in **Extended Data Fig. 7f**.

We next used high-performing tissue models (Pearson's $r \geq 0.72$, corresponding to the lower quantile of model performances across tissues) for *de novo* motif discovery (44 tissue models retained, **Extended Data Fig. 7a**). Using DeepLIFT [42] to score nucleotide contribution and TF-MoDISco[43] to cluster predictive subsequences, we identified a non-redundant set of 301 *de novo* motifs across 44 selected tissues (average: 16 per tissue, range: 9–23; **Fig. 3a**). The canonical CTCF motif was consistently recovered in all tissue contexts, with contribution weight matrices (CWM) strongly resembling the JASPAR positional weight matrix (PWM) (**Fig. 3b**), and closely matching motifs identified in humans[41]. We also recovered some well-known TF motifs (e.g., NEUROD1 in cerebellum and MEF2B in heart) (**Fig. 3c**) as well as motifs with limited similarity to entries in the JASPAR database (q-value < 0.05), suggesting under-characterized or potentially novel TFBSs (**Fig. 3d**). Scanning across all OCRs revealed the motif for KLF15, a zinc-finger transcription factor with key roles in metabolic processes[44], as the most frequently occurring one across tissues (**Fig. 3e**). Collectively, our findings demonstrate that ChromBPNet generalizes well to cattle, enabling accurate prediction of chromatin accessibility and discovery of conserved as well as novel sequence motifs. This supports its utility in decoding *cis*-regulatory logic and prioritizing functional variants in livestock.

An atlas of allele-specific regulatory variations

To assess the impact of genetic variants on regulatory activity, we performed allele-specific (AS) analyses across about 10 million SNPs identified from two individuals (see **Methods**). Out of 3,568,140 heterozygous SNPs overlapping epigenomic peaks, 91,125 (2.55%) exhibited significant allelic imbalance (adjusted $P < 0.05$, **Fig. 4a, Table S15**). AS SNPs were enriched in gene-dense and regulatory regions, showing positive correlation with gene density (Pearson's $r = 0.23$, $P = 3.92\times10^{-32}$), ChIP-seq peak density (Pearson's $r = 0.15$, $P = 8.7\times10^{-15}$) and OCRs (Pearson's $r = 0.18$, $P = 3.04\times10^{-20}$) (**Fig. 4a**). The majority (70%) of AS SNPs were located in intronic or intergenic regions (**Fig. 4b**), with significant enrichment within 5 kb upstream of TSS (**Fig. 4c**). AS effects on gene expression correlated strongly with those affecting active epigenomic marks (**Fig. 4d**), highlighting coordinated regulatory roles. Although effect sizes of RNA-seq AS SNPs showed limited concordance with eQTLs from the multi-breed cattle GTEx dataset[18,45] (**Supplementary Figure S27a**), they were highly and significantly correlated with AS SNPs previously reported by Prowse-Wilkins et al.[46] (**Fig. 4e**). Furthermore, effect sizes of epigenomic AS SNPs identified in our analysis showed high concordance with the Prowse-Wilkins dataset[46] (**Fig. 4f**).

To evaluate their functional relevance to complex traits, we intersected AS SNPs with trait-associated SNPs curated in the CattleQTLdb[5], identifying 439 putative variants (**Supplementary Figure S27b, Table S16**). These AS SNPs exhibited strong assay- and tissue-specific epigenomic effects (**Fig. 4g**). For example, SNP 3:15345227 in the *PKLR* promoter, a gene associated with milk traits[47] and highly expressed in brain-related tissues, exhibits a strong H3K27ac AS signal in jejunum ($\log_2$FC = 3.42), coinciding with a tissue-specific BivTssCtcf state (bivalent transcription start site with CTCF binding) (**Fig. 4h-i**). Another variant, SNP 29:42051788 in the *RTN3* promoter, showed the largest AS fold change in muscle, affecting H3K4me3 signal (**Supplementary Figure S27c-d**). *RTN3* has been associated with tenderness score[48] and lipid metabolism[49], suggesting potential causal functionality.

Together, these findings establish an atlas of allele-specific regulatory variants in cattle, linking genetic variation to tissue-specific regulatory effects and providing candidate functional variants underlying complex traits.

In silico prioritization of regulatory variants

S2F modeling enables quantitative prediction of the regulatory impact of genetic variants through *in silico* mutagenesis[39]. To systematically prioritize functional variants across tissue contexts, we employed deltaSVM[50], which balances context coverage (e.g. tissue and assay types) with computational efficiency (**Extended Data Fig. 8a**). Models were trained using epigenomic peaks as positives and GC-matched controls as negatives across 206 assay–tissue combinations (55 OCRs, 29 CTCF, 32 H3K27ac, 27 H3K27me3, 30 H3K4me1, 33 H3K4me3). Model performance was high, with average the receiver operating characteristic curves (AUCs) of 0.82–0.95 across assays (**Extended Data Fig. 8b**). Robustness was confirmed using OCR datasets from dairy cattle[24] with a median AUC of 0.95.

We then scored 28 million SNPs derived from 3,530 whole-genome sequences across diverse cattle breeds. We observed that deltaSVM scores were positively correlated with effect sizes of AS SNPs across epigenomic assays (**Fig. 5a, Supplementary Figure 28**), validating their use as proxies for regulatory impact of allelic changes. Following a strategy used in humans[51], we then defined a "regulatory magnitude" for each SNP as its maximum deltaSVM score across all contexts. This score was weakly but significantly correlated with both context specificity (i.e., the number of contexts in which the SNP overlapped with regulatory peaks) and genomic proximity to TSS (i.e., the distance of a SNP to the nearest gene TSS) (**Fig. 5b-c**), as well as with conservation metrics (phyloP and GERP), but not with the FAETH score[52]. We further categorized SNPs by regulatory context overlap: 1) Null: no overlap; 2) Specific: overlapping peaks in a single context; 3) Multiple: overlapping peaks in more than one but fewer than 80% of contexts; and 4) Shared: overlapping peaks in more than 80% of contexts. Stratification of SNPs by regulatory overlap revealed that tissue-shared variants were enriched near TSS and associated with promoter-like states, whereas tissue-specific variants overlapped primarily with enhancers (**Fig. 5d-e**). Moreover, both eQTL and trait-associated QTLs were significantly enriched for variants with high regulatory magnitude scores (**Supplementary Figure 28c**).

To evaluate their utility for fine mapping, we incorporated regulatory magnitude–based priors into Bayesian analyses of 22 complex traits (**Table S17**). This approach significantly reduced credible set sizes compared to unweighted analyses (**Fig. 5f**). For protein percentage, for example, regulatory weighting identified an additional putative causal SNP (6:84709163) (**Fig. 5g**). This variant is located in an intronic of *MGC152010*, which is highly expression in liver (**Supplementary Figure 28d**, **Fig. 5g, Table S17**), overlaps a liver-specific enhancer, and is predicted to disrupt high-mobility group factor binding (**Fig. 5g**). These results demonstrate that incorporating predicted regulatory effects into GWAS fine-mapping enhance causal-variant resolution, providing a framework for functional prioritization in livestock.

Comparative analysis of bovine and human epigenomes

To explore the evolutionary conservation of *cis*-regulatory elements (CREs), we compared chromatin states between cattle and human across 22 matched tissues. For humans, we analyzed

1,089 Roadmap Epigenomics datasets[17] spanning six regulatory marks (OCR, H3K27ac, H3K27me3, H3K4me1, H3K4me3, and CTCF). Human chromatin states were re-annotated using our pipeline to ensure methodological consistency, producing comparable emission profiles and state numbers in both species (**Fig. 6a**). We classified human CREs into four conservation categories: (1) sfCREs: conserved in both sequence and function; (2) sdCREs: conserved in sequence but divergent in function; (3) soCREs: conserved in sequence only; and (4) ssCREs: human-specific elements (**Fig. 6a**). Approximately 30% of human CREs were sdCREs, indicating dynamic functional divergence despite sequence conservation. PLS were more often functionally conserved (sfCREs) than ELS (**Fig. 6b**). Among active enhancers (EnhA), sequence-conserved types (sfCREs, sdCREs, and soCREs) exhibited significantly higher phyloP scores than ssCREs (q-value < 10^{-300}), with sfCREs being the most evolutionarily constrained (**Fig. 6c**). Notably, ssCREs were more proximal to TSS, whereas sequence-conserved elements, particularly sfCREs, were enriched at distal regions (**Fig. 6d**).

To assess trait relevance, we performed heritability partitioning using Linkage Disequilibrium Score Regression (LDSC)[53] across 25 human GWAS traits (**Table S18**). The analysis identified 46 significant tissue-trait associations (FDR < 0.05), including enrichment of heart failure heritability in myoblasts and fetal heart (**Fig. 6e**). Remarkably, sfCREs and ssCREs contributed disproportionately to these enrichments, accounting for 41.3% and 37.0% of all significant tissue-trait pairs, respectively (**Fig. 6f**). These findings highlight that both evolutionarily conserved and lineage-specific enhancers play prominent roles in shaping the genetic architecture of complex traits.

Discussion

Here we present a comprehensive bovine epigenomic atlas that integrates 1,147 datasets across a broad range of tissues, developmental stages, and primary cell types. By combining RNA-seq, ATAC-seq, WGBS, ChIP-seq, and Hi-C data, we mapped chromatin states and regulatory elements that collectively cover approximately 45% of the bovine genome, representing a ~20% increase over previous annotations[11]. This expanded regulatory landscape displayed extensive context specificity, supported by enriched tissue-relevant TFBS motifs such as *NEUROD1* in neural tissues and *MYF5* in myoblasts. Leveraging this atlas, we applied S2F models, including ChromBPNet[40] and deltaSVM[50], to decode regulatory grammar and predict variant effects. These models captured base-resolution accessibility patterns, recovered canonical (e.g. CTCF) and novel sequence motifs, and assigned quantitative regulatory scores to millions of variants. Integration with AS analyses and GWAS fine-mapping, highlighted putative causal variants. Cross-species comparisons further revealed that both evolutionarily conserved and lineage-specific enhancers contribute significantly to genetic architecture of complex traits: conserved elements (sfCREs) were distal and under strong sequence constraint, whereas species-specific elements (ssCREs), though often proximal to TSS, explained a large proportion of heritability for agricultural and biomedical traits. Together, these findings establish cattle as a powerful model for dissecting regulatory biology and complex traits, with relevance extending beyond agriculture to comparative and evolutionary genomics.

Despite these advances, several limitations remain. First, reliance on bulk tissue datasets limits resolution, masking cell-type–specific activity. Incorporating single-cell epigenomic and transcriptomic profiles will be critical to capture cellular heterogeneity and dynamic transitions

during development or disease progression[54]. Second, the small number of individuals profiled restricts detection of population-level regulatory variation. Population-scale efforts will enable discovery of molecular QTLs, construction of gene regulatory networks, and more precise genotype–phenotype links[10,18,45,55,56]. Third, deeper sequencing and enhanced data quality will improve sensitivity for detecting weak or transient regulatory elements, annotation and enable assessment of complex variants, such as structural variants[57], which are currently underexplored in functional contexts. Fourth, while our S2F models[58] perform robustly, further refinements—especially for those leveraging multimodal data and single-cell resolution—are needed to improve prediction of subtle or combinatorial variant effects. Fifth, comparative epigenomics across additional species will provide deeper insights into regulatory evolution and shared principles of trait heritability. Finally, functional validation remains essential. Massively parallel reporter assays[59], STARR-seq[60], and CRISPR-based perturbation strategies will be key for functionally confirming enhancer activity and directly testing variant effects in their native chromatin context.

In summary, this study provides the first atlas-scale annotation of regulatory elements in cattle, offering a foundational resource for mechanistic interpretation of non-coding variation. By integrating functional genomics with predictive modeling and comparative analysis, our work supports applications in precision breeding, evolutionary biology, and functional genomics. As future efforts expand to single-cell resolution and population scale, this resource lays essential groundwork for a comprehensive understanding of gene regulation and complex trait variation in mammals.

Acknowledgments:

We thank all of the people that contributed to the breeding and collection of tissues used in this study. We also thank all the researchers who have contributed to the publicly available data used in this research.

Funding:

This study was supported by Agriculture and Food Research Initiative Competitive Grant no. 2018-67015-27500 (H.Z., P.R. etc.) and sample collection was supported by no. 2015-67015-22940 (H.Z. and P.R.) from the USDA National Institute of Food and Agriculture. D.G. was supported by Agriculture and Food Research Initiative Competitive grants nos. 2018-67015-27500 (H.Z.) from the USDA National Institute of Food and Agriculture. J.B. received scholarships financed by Foundation to Support the Research of São Paulo State (FAPESP) (Grants Numbers: 2020/01369-1 and 2022/12631-4). H.Z. acknowledges funding from the USDA National Institute of Food and Agriculture, Multistate Research Project NRSP8 and NC1170 and the California Agricultural Experimental Station.

Author contribution statement:

H.Z., L.F. and D.G. conceived the manuscript. D.G. led all the data analysis. J.B. and D.G. analyzed RNA-seq and linked enhancer to target genes, J.B., D.G., C.A.G, G.W. and Y.X. processed WGBS data, D.G. and W.Y. annotated super-enhancers. W.Y., J.B., D.G., B.P., and

C.P.W. performed allele-specific analysis, C.P.W., D.G. analyzed chromatin accessibility data, D.G., J.W. and J.J. conducted complex trait integration analysis, S.C., X.X., Y.Wang, Y. Z. and D.G. led new data generation and quality control, H.L., H.Z. and L.F. provided eQTL resources used in this study, Y.U. and J.D. provide GWAS summary statistics from beef cattle, L.M., C.M., J.W. and J.J. provided GWAS summary statistics from dairy cattle. G.W. and Y.X. analyzed ATAC-Seq. P.L. and H.J. generated ChIP-seq data and primary cells. G.R., C.G., T.P.L.S., H.J., Z.J., W.L., J.R., J.F.M., J.D., P.R., J.E.K., H.Z. contributed to funding acquisition. D.G. prepared the initial manuscript draft with input from L.F. and H.Z. D.G., L.F. and H.Z., and revised and finalized the manuscript. All authors read, edited and approved the final manuscript.

Competing interests:

Y.T.U., V.M.P.R. and J.D. are employed by STgenetics. P.R.S acquired the funding for this research while affiliated to the University of California, Davis, and is currently employed by STgenetics. S.C. and G.R. is from Zoetis Inc.

Data and materials availability:

We have deposited newly generated RNA-Seq, ChIP-Seq, ATAC-seq WGBS in FAANG portal (https://data.faang.org/), NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra/), and European Nucleotide Archive. Accession IDs will be available upon the acceptance of the manuscript. Accession IDs for public data includes: PRJEB41939, PRJNA672996, PRJNA531208, E-MTAB-11825 and E-MTAB-11826 (ENA). Metadata details are available in Supplementary Tables S1-S4. All processed data are available at Zenodo database (https://zenodo.org/10.5281/zenodo.12216791, available upon the acceptance). Cattle traits QTLs retrieved from CattleQTLdb (https://www.animalgenome.org/cgi-bin/QTLdb/BT/index), and eQTL data retrieved from https://cattlecell.kiz.ac.cn/. Other data sets used in this study, published previously, are described in the Methods. All the computational scripts and codes are available at the GitHub website, including https://github.com/guandailu/BovineFAANG, https://github.com/guandailu/ChIPQC, and https://github.com/guandailu/SnakeHiC.

Online Methods

Sample collection

Samples used in this study were collected by the Functional Annotation of Animal Genomes (FAANG) community. The sampling procedure has been previously described in detail by Tixier-Boichard et al[61]. Briefly, tissue samples for beef cattle were collected from two male and two female Hereford cattle within 1–2 hours post-euthanasia, flash-frozen in liquid nitrogen, and stored at -80°C until further processing. The same protocol was applied to collect mammary gland tissues from Holstein cows at different physiological stages. All animal experiments were conducted in accordance with Protocol for Animal Care and Use No. 18464, approved by the Institutional Animal Care and Use Committee at the University of California, Davis.

Library preparation and sequencing

For RNA-seq experiments, frozen tissue samples were pulverized in liquid nitrogen using a mortar and pestle. Total RNA was then extracted with TRIzol reagent (Thermo Fisher Scientific, #15596026) and treated with DNase I (Thermo Fisher Scientific, #EN0521). RNA integrity was assessed using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and samples with RNA integrity number (RIN) > 8 were used for library preparation, followed by sequencing on an Illumina Hiseq 4000 platform (Illumina, San Diego, CA, USA). For ChIP-seq (targeting H3K4me3, H3K4me1, H3K27ac, H3K27me3, and CTCF) experiments, samples were processed as previously described[28,29]. ATAC-seq libraries were prepared using a modified protocol through incorporating cryopreserved nuclei workflows, as described in [62,63]. Both ATAC-seq and ChIP-seq libraries were sequenced on an Illumina HiSeq 4000 platform using 50-bp paired-end and single-end reads, respectively. For whole-genome bisulfite sequencing (WGBS), genomic DNA of tissue samples was extracted, fragmented, and sequencing libraries were prepared using methylation sequencing adaptors. Libraries were then treated with sodium bisulfite, which converted unmethylated cytosines to uracils while preserving methylated cytosines. The bisulfite-converted libraries were then PCR-amplified and subjected to Illumina paired-end 150 bp sequencing to assess genome-wide DNA methylation at single-base resolution.

Raw data processing and quality control

For ChIP-seq sequencing data, we processed them using a uniform pipeline available at https://github.com/guandailu/ChIPQC, after gathering newly generated sequencing data in this study and previously published datasets (details provided in **Table S1**). Briefly, sequencing adapters were trimmed using Trim Galore (v0.6.7). Cleaned reads were aligned to the ARS_UCD1.2 reference genome (Ensembl release v105) using BWA-MEM[64]. PCR duplicated reads were removed prior to peak calling. Deduplicated BAM files were used to identify peaks with MACS2[65] with the options of "--q 0.01" for narrow marks (i.e. H3K27ac, CTCF), and of "--q 0.05 --broad" for broad marks (i.e. H3K4me1, H3K27me3, H3K36me3, and H3K9me3). Then, we filtered out samples with the clean reads < 10 million, the number of peaks < 10,000 and Jensen-Shannon Divergence (JSD) between two replicates < 0.05 for broad marks and < 0.1 for narrow marks. This filtering step finally left 607 datasets that remained for subsequent analyses. For ATAC-seq, we applied the same processing steps, including adaptor trimming, read alignments, the removal of PCR duplicated reads, and peak calling, while the peaking calling was applied with the specific options of "-f BEDPE --nomodel". The filtering criteria for ATAC-seq included a FRiP score >0.3 and checked TSS enrichment manually. For RNA-seq, a similar pipeline was applied although the read aligner employed STAR (v2.7.0)[66]. Gene expression levels were quantified as Transcripts Per Million (TPM) using StringTie2[67], and raw read counts were obtained using featureCounts (release 2.0.6)[68] based on the ARS_UCD1.2 gene annotation (Ensembl release v105). For 12 public Hi-C datasets, we uniformly processed them using the pipeline SnakeHiC (https://github.com/guandailu/SnakeHiC). For DNA methylation data generated by WGBS, adaptors were trimmed using Trim Galore (v0.6.6) and we preprocessed using the Bismark pipeline (Version 0.24.2). The resulting alignment files in BAM format were subsequently converted into compressed formats, including "beta and "pat", using wgbs_tools (version 0.2.2)[69,70]. Homogeneously methylated blocks (defined as contiguous CpG sites with similar methylation levels) were therefore identified using the "*segment*" command, and blocks with fewer than 4 CpG sites were excluded from downstream analyses. For each segmented

block, we computed the average methylation level across CpG sites based on the corresponding *beta* formatted files.

Imputation of missing and low-quality epigenomic signals

To obtain complete epigenomes for chromatin state annotations, we employed the ChromImpute pipeline (v1.0.5) developed by the Roadmap Epigenomics Consortium[17,26,27]. Briefly, we generated signal tracks for experimental samples using the *bdgcmp* function of the MACS2 tool[65] , computing $-\log_{10}P$ as recommended by[26,27]. For brevity, we only considered signals for autosomes 1-29. ChIP-seq signal tracks were converted to 25-bp resolution by averaging signal intensity across each 25-bp bin. Imputation of missing or low-quality signals followed four main steps: 1) computing global distance between datasets with the *ComputeGlobalDist* function; 2) generating training features with the *GenerateTrainData* command; 3) training mark-specific and sample-specific predictors with *Train* command; and 4) applying these predictors to generate imputed signal tracks using *Apply*. To assess imputation accuracy, we implemented a cross-validation-like procedure. Two experimental samples were randomly masked, and imputation was performed on these held-out samples. The correlation between imputed and observed signals was then assessed using the *Eval* function. This process was repeated 15 times to ensure that each epigenetic mark was tested at least twice. In addition, imputed signals were manually inspected to ensure quality. Additionally, we validated the robustness of the pipeline using ChIP-seq datasets published by Prowse-Wilkins et al.[25], masking 15 samples using the same procedure. Comparable high imputation accuracy was observed across these held-out datasets.

Predicting chromatin states and super-enhancers

We employed ChromHMM (v1.26)[71,72] to predict chromatin states for 72 bovine epigenomes, each of which included the core set of four histone modifications: H3K4me1, H3K4me3, H3K27ac, and H3K27me3. Briefly, we used the *LearnModel* function in ChromHMM to train models with state numbers ranging from 2 to 15—the maximum number supported by the software. To identify the optimal model, we compared emission probability across models following the approach described by Gorkin et al.[73] The optimal number of states was determined as the point at which the median correlation of emission probability across two or more replicates plateaued. For the 72 bovine epigenomes, a 10-state model was ultimately selected, and state labels were manually assigned based on an extensive literature review, particularly as described in ref[74]. For 65 bovine complete epigenomes with two additional marks (OCRs and CTCF), we ran a similar pipeline and learned 15 chromatin states. In order to compare cattle data with that of humans, we also conducted the chromatin state annotation for 134 human epigenomes with the same pipeline.

We also annotated super-enhancers, a class of regulatory elements that differ from typical enhancers in terms of their size, transcription factor density and content, transcriptional activation potential, and sensitivity to perturbation[31,75]. Super-enhancer identification was performed using the ROSE algorithm (version 0.1)[31] across 33 tissue contexts. For each tissue, H3K27ac ChIP-seq profiles from two biological replicates were merged using SAMtools (v1.19.2)[76]. As input, we used the genomic regions annotated as typical enhancers (i.e. chromatin

states E4, E5, and E6) to define candidate super-enhancer regions. These were then ranked based on H3K27ac signal intensity to identify super-enhancers following the standard ROSE pipeline[31].

Linking enhancers to genes

To predict target genes of enhancer elements, we employed the Activity-by-Contact (ABC) model[30]. We applied the ABC pipeline across 62 samples from 37 tissues that had matched ATAC-seq and H3K27ac ChIP-seq data. The ABC score for a given enhancer-gene pair was calculated as:

$$ABC\ score_{(E,G)} = \frac{A_E \times C_{E,G}}{\sum A_e \times C_{e,G}}$$

where E and G represent the enhancer element and the gene, respectively. A(E) represents the activity of element E, and C(E, G) denotes the Hi-C contact frequency between element E and the promoter of gene G. The denominator sums this product over all candidate elements within ±5 Mb of the gene. Activity was quantified using read counts from ATAC-seq and H3K27ac ChIP-seq data. Candidate enhancer elements were defined as ~500 bp regions centered on ATAC-seq peaks.

Context specificity analyses

For gene expression: We employed the Tspex[77] as described previously[55]. For a given tissue, we computed the average gene expression across biological replicates. Tissue-specific expression was then defined using a Z-score approach across all tissues, with genes considered tissue-specific if Z-score >= 0.8. Adult and fetal tissues were separate in this analysis. Subsequently, we performed enrichment analysis on the identified tissue-specific gene sets using the Enrichr package (v3.2)[78] and the GO_Biological_Process_2023 database. The significant threshold was set as adjusted *P*-value threshold of <= 0.05 with the False Discovery Rate (FDR) approach[79].

For chromatin states: We applied an intuitive comparison between the target tissue and the test to identify tissue-specific chromatin states with BEDtools (version 2.31.0)[80]. That is, a state observed in one tissue but not in others. Motif enrichment of tissue-specific chromatin states was conducted using the Hypergeometric Optimization of Motif EnRichment (HOMER) tool (v5.1, 7-16-2024)[32].

For open chromatin regions: To identify tissue-specific OCRs, we first gathered all peaks across valid samples to create the consensus set of peaks. Then we quantified raw reads counts of consensus peaks across samples. Tissue-specific OCRs were identified by using DESeq2 tool (v 1.49.1)[81] when comparing a given tissue with the rest. The significant threshold was set as FDR-adjusted $P < 0.05$ and fold change at $\log_2$ scale > 1.

For DNA methylation: Tissue-specific differentially methylated regions (DMRs) were identified by comparing each focal tissue against all others using the *find_markers* function in the wgbs_tools[69] package (v0.2.2). Statistical significance was assessed via two-sided t-tests, and DMRs with $P > 0.05$ were excluded using the --pval 0.05 option.

Comparing adult and fetal samples: We identified differential chromatin states between adult and fetal stages using an BEDtools (version 2.31.0). That is, a state observed in one stage but not in another one.

Allele-specific identification

We identified genome-wide single nucleotide polymorphisms (SNPs) from two biological replicates (M08 and M22) using the Genome Analysis Toolkit (GATK, v4.5.0.0) following the "best practices" pipeline[82]. To minimize reference mapping bias and enable accurate detection of allele-specific (AS) events, we applied the WASP (v0.16.3) pipeline[83] to all epigenomic datasets. Read counts for each SNP were quantified using the *ASEReadCounter* module in GATK. SNPs were retained for AS analysis if they (i) were heterozygous, (ii) overlapped epigenetic peaks previously identified by MACS2[65], (iii) were located on autosomes, and (iv) had >6 total reads, with each allele supported by at least 3 reads. For each retained SNP within a sample, a two-sided binomial test was performed to assess allelic imbalance. *P*-values were then adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure[79], and SNPs with a false discovery rate (FDR) < 5% were considered significant AS events.

Training ChromBPNet to learn sequence motifs

We employed ChromBPNet (v1.0)[40], a CNN-based deep learning framework, to learn sequence regulatory syntax of chromatin accessibility in the bovine genome. Briefly, for each of the 55 tissue contexts, we merged BAM files from biological replicates to generate input data. The analysis was restricted to ATAC-seq data, as ChromBPNet is specifically designed to model chromatin accessibility and also for simplicity. OCR peaks, previously called by MACS2[65], were merged by tissue using the Irreproducible Discovery Rate (IDR) framework (version 2.0.3)[84] and filtered to exclude regions overlapping with highly repetitive sequences. Genome mappability for the ARS-UCD1.2 reference genome was computed using GenMap (v1.3.0) [85]. ChromBPNet models were trained using a chromosome-based data split: training set (chr2, chr4-5, chr7, chr9-19, chr21-29), validation set (chr8, chr20), and testing set (chr1, chr3, chr6). As negative sets, we generated GC-matched background regions using the "prep nonpeaks" function of the ChromBPNet pipeline. Model training proceeded in two steps: (1) training a bias model to account for Tn5 insertion bias, and (2) training a bias-factorized ChromBPNet model. Chromatin accessibility in OCRs from the test set was then predicted using the *pred_bw* function, which outputs both read counts and signal shape of profiles. Prediction accuracy was evaluated using Pearson's correlation, Spearman's correlation, and JSD. In accordance with the original publication[40], Pearson's correlation between observed and predicted read counts demonstrated desirable evaluations and was therefore selected for subsequent analyses. To interpret model predictions and identify sequence motifs underlying OCRs, nucleotide-level contribution scores were computed and stored as BigWig files. These scores were used for *de novo* motif discovery with TF-MoDISCo (v1.0.0)[43]. To construct a non-redundant motif catalog across tissue contexts, we clustered predicted motif instances and merged clusters with pairwise similarity scores greater than 0.1, yielding a final set of 301 non-redundant sequence motifs.

Training gkm-SVM classifier

To assess the regulatory impact of DNA sequence variants in the cattle genome, we adopted the deltaSVM approach[50], which offers both high predictive accuracy and computational efficiency, making it suitable for genome-wide variant scoring across diverse epigenetic marks and tissue contexts. Specifically, we trained gapped k-mer Support Vector Machine (gkm-SVM) classifiers[86,87] for each tissue-epigenetic mark combination. For each epigenetic mark, we first merged MACS2[65] identified peaks from two biological replicates using the IDR framework

(v2.0.3) [84] and filtered out regions overlapping with highly repetitive sequences. The remaining peaks were then aggregated across all tissues to define a positive set. To generate the negative set, we sampled GC-matched genomic regions that did not overlap any peak regions. We used LS-GKM (v0.1.1) [88], an optimized implementation of gkm-SVM with improved performance[87,88], to train the classifiers. Model performance was evaluated using 10-fold cross-validation, with parallelization enabled via the "-i" option.

To compute deltaSVM scores, we first generated all non-redundant 11-mers using the *nrkmers.py* script. Then, for ~28 million SNPs identified from 3,530 whole-genome sequences (generated by collaborators, unpublished), we synthesized 40-bp allelic oligonucleotides for both reference and alternative alleles (centered on the SNP with ±19 bp flanking sequence). Variant impact scores were computed using the deltasvm.pl script with the following command: perl deltasvm.pl chr${chr}.ref.fa chr${chr}.alt.fa weights.txt deltaSVM.tsv. For each SNP, we defined the maximum deltaSVM score across all tissue-mark combinatory contexts as its regulatory magnitude, representing the predicted change in regulatory activity between alleles. To assess whether CattleGTEx eQTLs[18,45] and traits QTLs[5] are enriched for high regulatory magnitude scores, we conducted an enrichment analysis by fitting a linear model. For eQTLs, we compared fine-mapped causal variants (posterior inclusion probability [PIP] > 0.8) to non-causal eQTL SNPs (PIP < 0.01). For trait QTLs, we contrasted them with randomly selected SNPs matched for minor allele frequency (MAF). In both analyses, we regressed out genomic confounders, including distance to the nearest TSS, context specificity, and GERP conservation scores, as these features were significantly correlated with regulatory magnitude.

GWAs fine-mapping analyses

Genome-wide significant association peaks for 22 dairy bull traits were identified from GWAS summary statistics (Wang et al., in press) using a significance threshold of $P < 5 \times 10^{-8}$. For fine-mapping analysis, candidate regions were defined as 2-Mb windows centered on the lead SNP of each associated peak. Fine-mapping was performed using SuSiE-adj[89], an adaptation of SuSiE[90] that improves fine-mapping accuracy in samples with related individuals by incorporating linear mixed model (LMM)-derived summary statistics. We applied SuSiE-adj both with and without scaled regulatory magnitude as prior weights across 22 traits. Conditional analysis was conducted by including the lead SNP as a covariate and performing association analysis using the SLEMM-GWA method[91].

Linkage disequilibrium score regression analysis

To investigate whether the conservation of regulatory elements defined in this study could inform the genetic architecture of human complex traits, we collected GWAS summary statistics for 25 traits (details in **Table S18**). We performed tissue-specific (equivalent to cell-type-specific[92]) heritability enrichment analyses using the LDSC package (v1.0.0)[93]. GWAS summary statistics were standardized using the *munge_sumstats.py* script to remove SNPs with MAF ≤ 0.01, INFO score ≤ 0.9, genotype call rate ≤ 0.75, duplicate rsIDs, out-of-bounds *P*-values, extreme chi-squared statistics, strand ambiguity, or discordance with SNPs used in LD score calculations. For each category of regulatory elements, we identified tissue-specific segments by

contrasting the target tissue with others using BEDtools (v2.31.0)[80]. Annotations were generated for LDSC analysis using the *make_annot.py* script, and stratified LD scores were then calculated within a 500 kb window using SNPs from the 1000 Genomes Project Phase 3. We used default SNP weights and reference panels obtained from https://zenodo.org/records/7768714. Tissue-specific heritability enrichment analyses were then conducted using the "--h2-cts" option in LDSC[92].

Figure 1. Overview of the study design and datasets. (a) Schematic overview of the analytical framework used in this study, including data processing, imputation of missing or low-quality histone marks, chromatin state annotation, enhancer-gene linking, tissue-specificity analysis, sequence grammar modelling using deep learning, fine-mapping GWAS loci of complex traits, and cross-species epigenomic comparison. (b) Summary of multi-omic datasets integrated in this study. (c) Illustration of tissue samples used for chromatin state annotation in this study. Icons and the cattle image were adapted from BioRender.com.

Figure 2. Chromatin state annotation and context-specific regulatory dynamics. (a) Definition and characterization of chromatin states, including emission probabilities, functional descriptions (annotations), abbreviations, enrichment near transcription start sites (TSS), overlap with genomic features, associated gene expression levels, and evolutionary constraint based on GERP scores[94] (Ensembl release v105). CDS: coding sequence. UTR: untranslated regions. (b) Genome coverage of strong enhancers (EnhA) compared to annotations from that of Kern et al.[11] (c)-(e) Enhancer-gene linking based on the Activity-by-Contact (ABC) model[30]: number of genes linked per enhancer (c), number of enhancers linked per gene (d), and genomic distance between enhancers and their predicted target genes (e). (f) Number of tissue-specific strong enhancers (EnhA) identified across tissues. (g) Motif enrichment analysis of tissue-specific EnhA elements. Representative enriched motifs and associated transcription factors (TFs) with known tissue relevance are shown at right. (h) Venn diagram depicting the overlap of strong enhancers (EnhA) identified in myoblasts (cell line), adult muscle and fetal muscle (bulk tissue). (i) Genome browser tracks displaying gene expression and epigenomic signals around the *MYH5* and *MYH6* loci (chr5:10262831-10302829 bp).

Figure 3. Sequence motif discovery across tissue contexts. (a) Number of seqlets (i.e., short informative subsequences) supporting each of the 301 unique motifs (y-axis) across 42 tissue contexts (x-axis). (b) Comparison between the CTCF motif identified by ChromBPNet[40] in this study (top) and the canonical CTCF motif from the JASPAR database (bottom). (c) Representative motifs identified in this study showing high similarity to known TF motifs curated in the JASPAR database. (d) Five motifs discovered in the rumen, a ruminant-specific organ, which show low similarity to any known motifs in JASPAR motifs (q-value > 0.05). (e) Proportion of motif instances identified across tissue contexts. The KLF15 motif was the most frequently detected. The unit of y-axis for motifs patterns is "bits", while the ChromBPNet derived motifs are represented using contribution weight matrix (CWM) score. The known non-redundant motifs from the JASPAR database were last accessed on February 11, 2022. TF: transcription factor.

Figure 4. Identification and characterization of allele-specific (AS) epigenomic variation. (a) Genome-wide distribution of AS SNPs across bovine autosomes. (b) Proportion of SNPs (x-axis) by variant categories (y-axis). The variant category was obtained using SnpEff[95]. (c) Genomic distribution of significant AS SNPs relative to gene transcription start sites (TSS). (d) Correlation of AS effect sizes ($\log_2$ scaled) in RNA-seq (x-axis) and matched epigenomic assays (y-axis). (e) Comparison of AS effect sizes ($\log_2$ scale) in RNA-seq between this study (x-axis) and Prowse-Wilkins et al.[46] (y-axis). (f) Comparison of AS effect sizes ($\log_2$ scale) in epigenomic assays between this study (x-axis) and Prowse-Wilkins et al.[46] (y-axis). P-values in (d–f) were computed using the asymptotic t-distribution approximation. FC: fold change. (g) Heatmap showing fold changes ($\log_2$ scale) for 439 AS SNPs overlapping known trait-associated QTLs curated in the CattleQTLdb[5]. (h) Gene expression profile of *PKLR* across tissues measured in

transcripts per million (TPM). (i) Genomic tracks showing an AS SNP in the promoter region of *PKLR*, detected in jejunum H3K27ac data. This SNP overlaps a predicted ETV5:FIGLA binding motif (identified using FIMO[96]).

Figure 5. Predicting regulatory effects of sequence variants. (a) Correlation between deltaSVM scores (y-axis) and allelic fold change (log□-scaled, x-axis) across four assays: ATAC-seq, CTCF, H3K27ac, and H3K4me3 (left to right, respectively). (b) Relationship between regulatory magnitude (left y-axis: defined as the maximum deltaSVM score across all assay-tissue contexts) and distance to the nearest transcription start site (TSS; right y-axis) as a function of context specificity (i.e., number of assay–tissue contexts affected). (c) Mean distance to TSS for genetic variants stratified by deciles of regulatory magnitude. (d) Density plot showing the distribution of variant distances to TSS categorized by context specificity, including: "null" (impacting 0 assay-tissue contexts), "specific" (impacting only 1 assay-tissue context), "multiple" (impacting > 1 but <80% of assay-tissue contexts); and "shared" (impacting >80% of assay-tissue contexts). (e) Proportion of SNPs residing within enhancer-like chromatin states (EnhPos, EnhWk, EnhA) and promoter-like chromatin states (TssWk, TssFlnk, TssA), stratified by "Specific" and "Shared" variant categories. (f) Regulatory magnitude weighted fine mapping produces smaller median credible set sizes across 22 complex traits (P=4.43 × 10^{-8}, Wilcoxon test). (g) Fine-mapping a cattle GWAS locus for protein percentage in milk using regulatory magnitude as prior weights in SuSiE-adj[89,90]. The panels from top to down are: GWAS significance plot ($-\log_{10}P$); posterior inclusion probability (PIP) for fine-mapped variants without prior weights; posterior inclusion probability (PIP) for fine-mapped variants with regulatory magnitude prior weights; chromatin state annotations across 65 bovine epigenomes. The vertical red line highlights the fine-mapped variant overlapping both chromatin states and transcription factor binding sites (TFBS).

Figure 6. Cross-species comparison and conservation of regulatory elements. (a) Chromatin state annotations comprising 15 states were independently predicted in cattle and human. The top panel summarizes chromatin state labels, functional descriptions, abbreviations, and corresponding emission probabilities for both species. The bottom panel categorizes cis-regulatory elements (CREs) based on conservation: 1) sfCREs, sequence- and functionally conserved elements; 2) sdCREs, sequence conserved but functionally divergent; 3) soCREs, sequence-only conserved; and 4) ssCREs, human-specific elements. (b) Proportions of regulatory elements stratified by conservation category. (c) phyloP conservation scores for strong enhancers (EnhA), grouped by conservation category. Significance was assessed using a two-sided Student's t-test with FDR adjustment. phyloP scores were obtained from Cactus 241-way mammalian alignments provided by the Zoonomia Project[97]. (d) Genomic distance between strong enhancers (EnhA) and the nearest transcription start site (TSS), stratified by conservation category. Color legend is same as panel (b). (e) Tissue-specific heritability enrichment of EnhA elements stratified by conservation category, represented as $-\log_{10}P$ values. (f) Number of significantly enriched tissue–trait pairs (x-axis) contributed by EnhA elements from each conservation category (y-axis). Trait name abbreviations are provided at **Table S28.**

Reference:

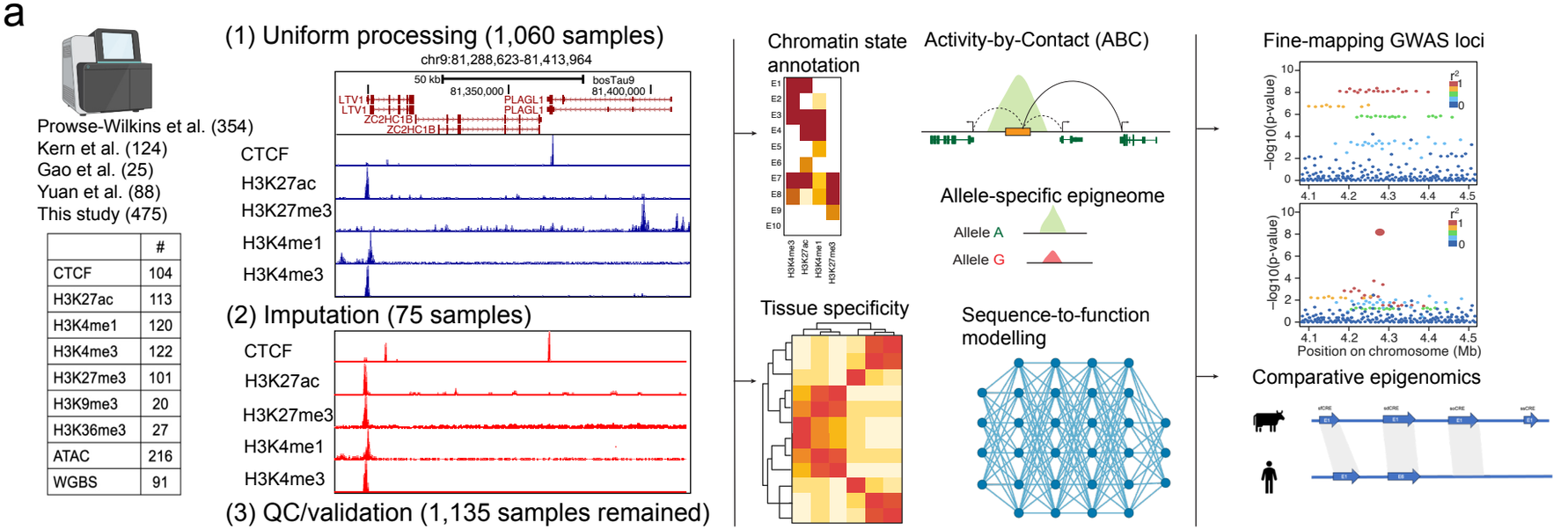
1. Hubert, J.-N., Perret, M., Riquet, J. & Demars, J. Livestock species as emerging models for genomic imprinting. *Front. Cell Dev. Biol.* **12**, (2024).
2. Kemper, K. E. & Goddard, M. E. Understanding and predicting complex traits: knowledge from cattle. *Human Molecular Genetics* **21**, R45–R51 (2012).
3. Georges, M., Charlier, C. & Hayes, B. Harnessing genomic information for livestock improvement. *Nat Rev Genet* **20**, 135–156 (2019).
4. Wiggans, G. R., Cole, J. B., Hubbard, S. M. & Sonstegard, T. S. Genomic Selection in Dairy Cattle: The USDA Experience. *Annual Review of Animal Biosciences* **5**, 309–327 (2017).
5. Hu, Z.-L., Park, C. A. & Reecy, J. M. Bringing the Animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Research* **50**, D956–D961 (2022).
6. Koufariotis, L., Chen, Y.-P. P., Bolormaa, S. & Hayes, B. J. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. *BMC Genomics* **15**, 436 (2014).
7. Zhou, H. *et al.* Comparative Genomics and Epigenomics of Transcriptional Regulation. *Annual Review of Animal Biosciences* **13**, 73–98 (2025).
8. Hayes, B. J. *et al.* Potential approaches to create ultimate genotypes in crops and livestock. *Nat Genet* **56**, 2310–2317 (2024).
9. Tait-Burkard, C. *et al.* Livestock 2.0 – genome editing for fitter, healthier, and more productive farmed animals. *Genome Biology* **19**, 204 (2018).
10. Fang, L. *et al.* The Farm Animal Genotype–Tissue Expression (FarmGTEx) Project. *Nat Genet* 1–11 (2025) doi:10.1038/s41588-025-02121-5.
11. Kern, C. *et al.* Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun* **12**, 1821 (2021).
12. Fang, L. *et al.* Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biology* **17**, 68 (2019).
13. Gao, Y. *et al.* Functional annotation of regulatory elements in cattle genome reveals the roles of extracellular interaction and dynamic change of chromatin states in rumen development during weaning. *Genomics* **114**, 110296 (2022).
14. Salavati, M. *et al.* Improving the annotation of the cattle genome by annotating transcription start sites in a diverse set of tissues and populations using CAGE sequencing. *G3 Genes|Genomes|Genetics* jkad108 (2023) doi:10.1093/g3journal/jkad108.
15. Beiki, H. *et al.* Enhanced bovine genome annotation through integration of transcriptomics and epi-transcriptomics datasets facilitates genomic biology. *GigaScience* **13**, giae019 (2024).
16. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
17. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
18. Liu, S. *et al.* A multi-tissue atlas of regulatory variants in cattle. *Nat Genet* **54**, 1438–1447 (2022).
19. Liu, S. & Fang, L. The CattleGTEx atlas reveals regulatory mechanisms underlying complex traits. *Nat Genet* **54**, 1273–1274 (2022).
20. Li, J. *et al.* A multi-tissue atlas of allelic-specific expression reveals the characteristics, mechanisms, and relationship with dominant effects in cattle. *BMC Biology* **23**, 148 (2025).

21. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535-548.e24 (2019).
22. Zeng, T. & Li, Y. I. Predicting RNA splicing from DNA sequence using Pangolin. *Genome Biology* **23**, 103 (2022).
23. Charles, M. *et al.* Functional impact of splicing variants in the elaboration of complex traits in cattle. *Nat Commun* **16**, 3893 (2025).
24. Yuan, C. *et al.* An organism-wide ATAC-seq peak catalog for the bovine and its use to identify regulatory variants. *Genome Res.* (2023) doi:10.1101/gr.277947.123.
25. Prowse-Wilkins, C. P. *et al.* Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. *Front. Genet.* **12**, (2021).
26. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
27. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**, 364–376 (2015).
28. Pan, Z. *et al.* Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat Commun* **12**, 5848 (2021).
29. Pan, Z. *et al.* An atlas of regulatory elements in chicken: A resource for chicken genetics and genomics. *Science Advances* **9**, eade1204 (2023).
30. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).
31. Whyte, W. A. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307–319 (2013).
32. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010).
33. Li, L., Chambard, J. C., Karin, M. & Olson, E. N. Fos and Jun repress transcriptional activation by myogenin and MyoD: the amino terminus of Jun can mediate repression. *Genes Dev.* **6**, 676–689 (1992).
34. Pataskar, A. *et al.* NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program. *The EMBO Journal* **35**, 24–45 (2016).
35. Ott, M.-O., Bober, E., Lyons, G., Arnold, H. & Buckingham, M. Early expression of the myogenic regulatory gene, myf-5, in precursor cells of skeletal muscle in the mouse embryo. *Development* **111**, 1097–1107 (1991).
36. Hadchouel, J. *et al.* Analysis of a key regulatory region upstream of the Myf5 gene reveals multiple phases of myogenesis, orchestrated at each site by a combination of elements dispersed throughout the locus. *Development* **130**, 3415–3426 (2003).
37. Francetic, T. & and Li, Q. Skeletal myogenesis and Myf5 activation. *Transcription* **2**, 109–114 (2011).
38. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
39. Sokolova, K., Chen, K. M., Hao, Y., Zhou, J. & Troyanskaya, O. G. Deep Learning Sequence Models for Transcriptional Regulation. (2024) doi:10.1146/annurev-genom-021623-024727.
40. Pampari, A. *et al.* ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. 2024.12.25.630221 Preprint at https://doi.org/10.1101/2024.12.25.630221 (2024).

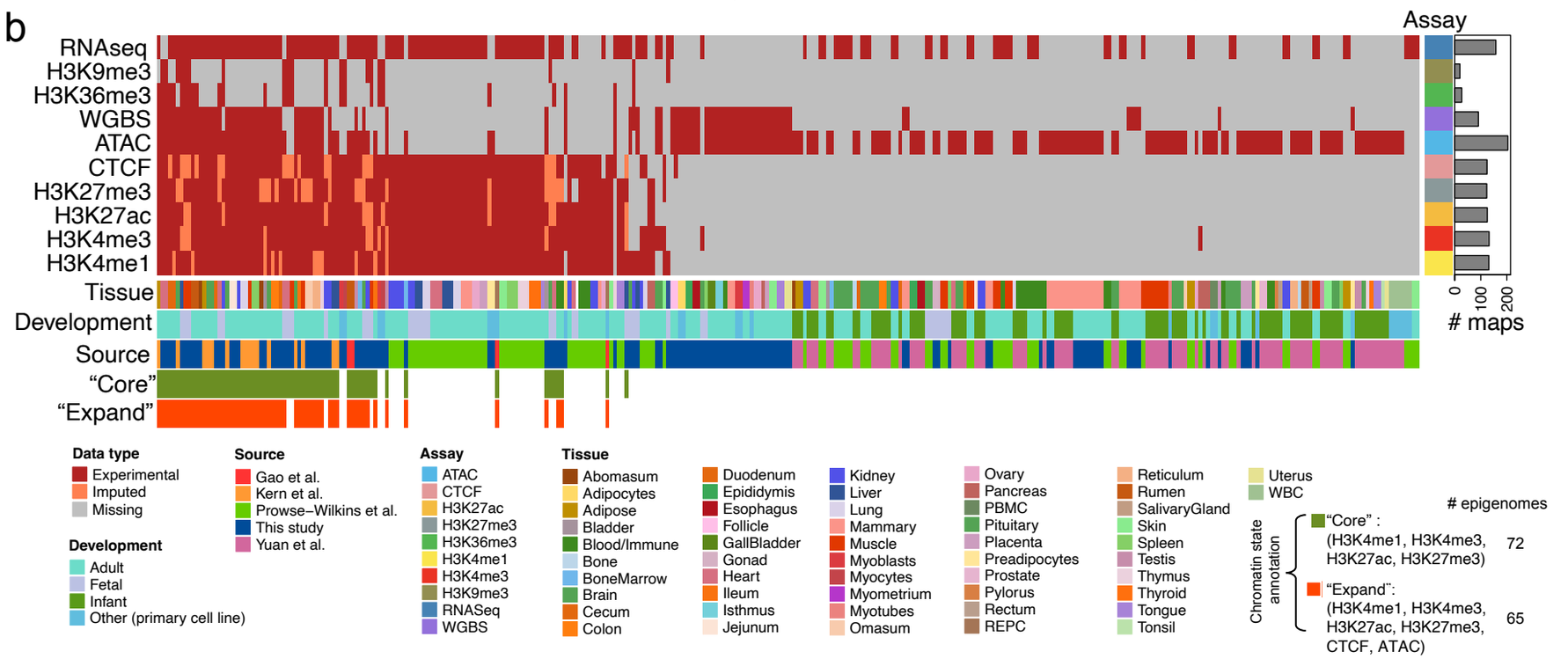
41. Liu, B. B. *et al.* Dissecting regulatory syntax in human development with scalable multiomics and deep learning. 2025.04.30.651381 Preprint at https://doi.org/10.1101/2025.04.30.651381 (2025).
42. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. Preprint at https://doi.org/10.48550/arXiv.1704.02685 (2019).
43. Shrikumar, A. *et al.* Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. Preprint at https://doi.org/10.48550/arXiv.1811.00416 (2020).
44. Fan, L. *et al.* Muscle Krüppel-like factor 15 regulates lipid flux and systemic metabolic homeostasis. *J Clin Invest* **131**, (2021).
45. Li, H. *et al.* An atlas of cell type specific regulatory effects in cattle. 2025.06.23.661035 Preprint at https://doi.org/10.1101/2025.06.23.661035 (2025).
46. Prowse-Wilkins, C. P., Wang, J., Garner, J. B., Goddard, M. E. & Chamberlain, A. J. Allele specific binding of histone modifications and a transcription factor does not predict allele specific expression in correlated ChIP-seq peak-exon pairs. *Sci Rep* **13**, 15596 (2023).
47. Du, A. *et al.* Genetic polymorphisms of PKLR gene and their associations with milk production traits in Chinese Holstein cows. *Front. Genet.* **13**, (2022).
48. Fonseca, P. A. de S., Caldwell, T., Mandell, I., Wood, K. & Cánovas, A. Genome-wide association study for meat tenderness in beef cattle identifies patterns of the genetic contribution in different post-mortem stages. *Meat Science* **186**, 108733 (2022).
49. Guo, D. *et al.* Lipid overload-induced RTN3 activation leads to cardiac dysfunction by promoting lipid droplet biogenesis. *Cell Death Differ* **31**, 292–308 (2024).
50. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**, 955–961 (2015).
51. Marderstein, A. R. *et al.* Mapping the regulatory effects of common and rare non-coding variants across cellular and developmental contexts in the brain and heart. 2025.02.18.638922 Preprint at https://doi.org/10.1101/2025.02.18.638922 (2025).
52. Xiang, R. *et al.* Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. *Proceedings of the National Academy of Sciences* **116**, 19398–19408 (2019).
53. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
54. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat Rev Genet* 1–23 (2022) doi:10.1038/s41576-022-00509-1.
55. Guan, D. *et al.* Genetic regulation of gene expression across multiple tissues in chickens. *Nat Genet* 1–11 (2025) doi:10.1038/s41588-025-02155-9.
56. Teng, J. *et al.* A compendium of genetic regulatory effects across pig tissues. *Nat Genet* **56**, 112–123 (2024).
57. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res.* gr.275488.121 (2021) doi:10.1101/gr.275488.121.
58. Sasse, A., Chikina, M. & Mostafavi, S. Unlocking gene regulation with sequence-to-function models. *Nat Methods* **21**, 1374–1377 (2024).

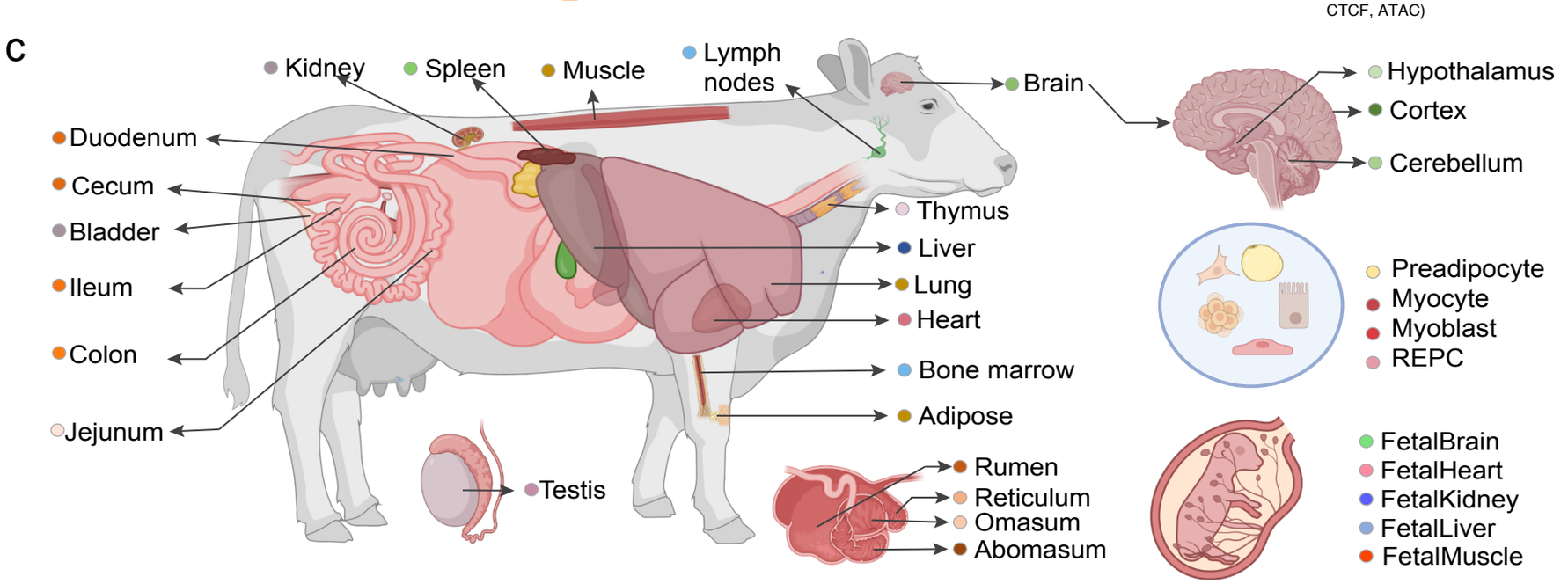
59. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. 2023.03.05.531189 Preprint at https://doi.org/10.1101/2023.03.05.531189 (2023).
60. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
61. Tixier-Boichard, M. *et al.* Tissue Resources for the Functional Annotation of Animal Genomes. *Frontiers in Genetics* **12**, 847 (2021).
62. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14**, 959–962 (2017).
63. Halstead, M. M. *et al.* Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Sci Rep* **10**, 5230 (2020).
64. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://doi.org/10.48550/arXiv.1303.3997 (2013).
65. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
66. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
67. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278 (2019).
68. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
69. Loyfer, N., Rosenski, J. & Kaplan, T. wgbstools: A computational suite for DNA methylation sequencing data representation, visualization, and analysis. 2024.05.08.593132 Preprint at https://doi.org/10.1101/2024.05.08.593132 (2024).
70. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).
71. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**, 2478–2492 (2017).
72. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216 (2012).
73. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).
74. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28**, 817–825 (2010).
75. Lovén, J. *et al.* Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* **153**, 320–334 (2013).
76. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
77. Camargo, A. P., Vasconcelos, A. A., Fiamenghi, M. B., Pereira, G. A. G. & Carazzolle, M. F. tspex: a tissue-specificity calculator for gene expression data. *Research Square* Preprint at https://doi.org/10.21203/rs.3.rs-51998/v1 (2020).
78. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90–W97 (2016).
79. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

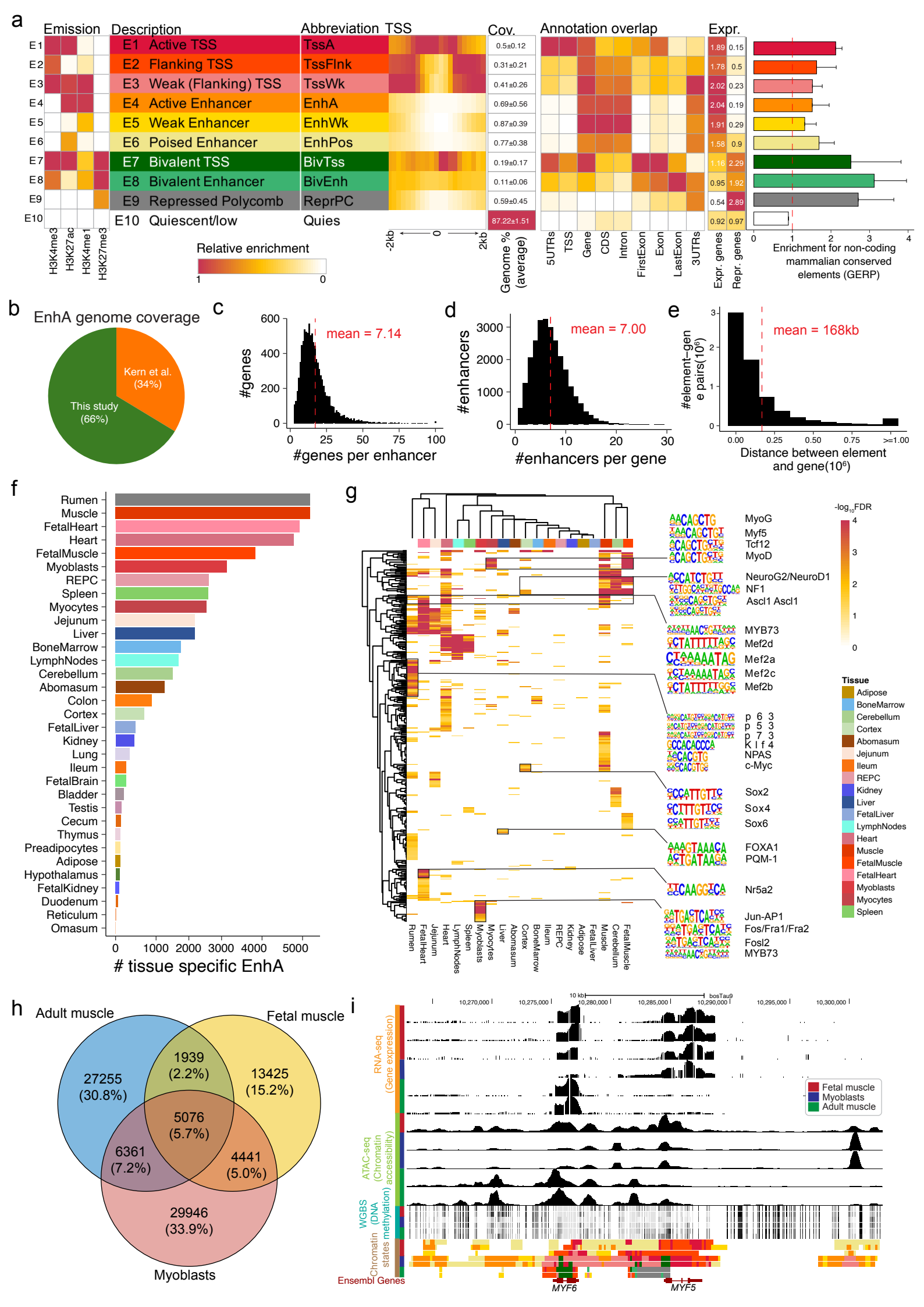
80. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
81. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).
82. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
83. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061–1063 (2015).
84. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, (2011).
85. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).
86. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLOS Computational Biology* **10**, e1003711 (2014).
87. Ghandi, M. *et al.* gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* **32**, 2205–2207 (2016).
88. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets.
89. Wang, J. *et al.* Fine-mapping methods for complex traits: essential adaptations for samples of related individuals. 2021.10.10.463846 Preprint at https://doi.org/10.1101/2021.10.10.463846 (2025).
90. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLOS Genetics* **18**, e1010299 (2022).
91. Cheng, J. *et al.* SLEMM: million-scale genomic predictions with window-based SNP weighting. *Bioinformatics* **39**, btad127 (2023).
92. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621–629 (2018).
93. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).
94. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLOS Computational Biology* **6**, e1001025 (2010).
95. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
96. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
97. Christmas, M. J. *et al.* Evolutionary constraint and innovation across hundreds of placental mammals. 2023.03.09.531574 Preprint at https://doi.org/10.1101/2023.03.09.531574 (2023).

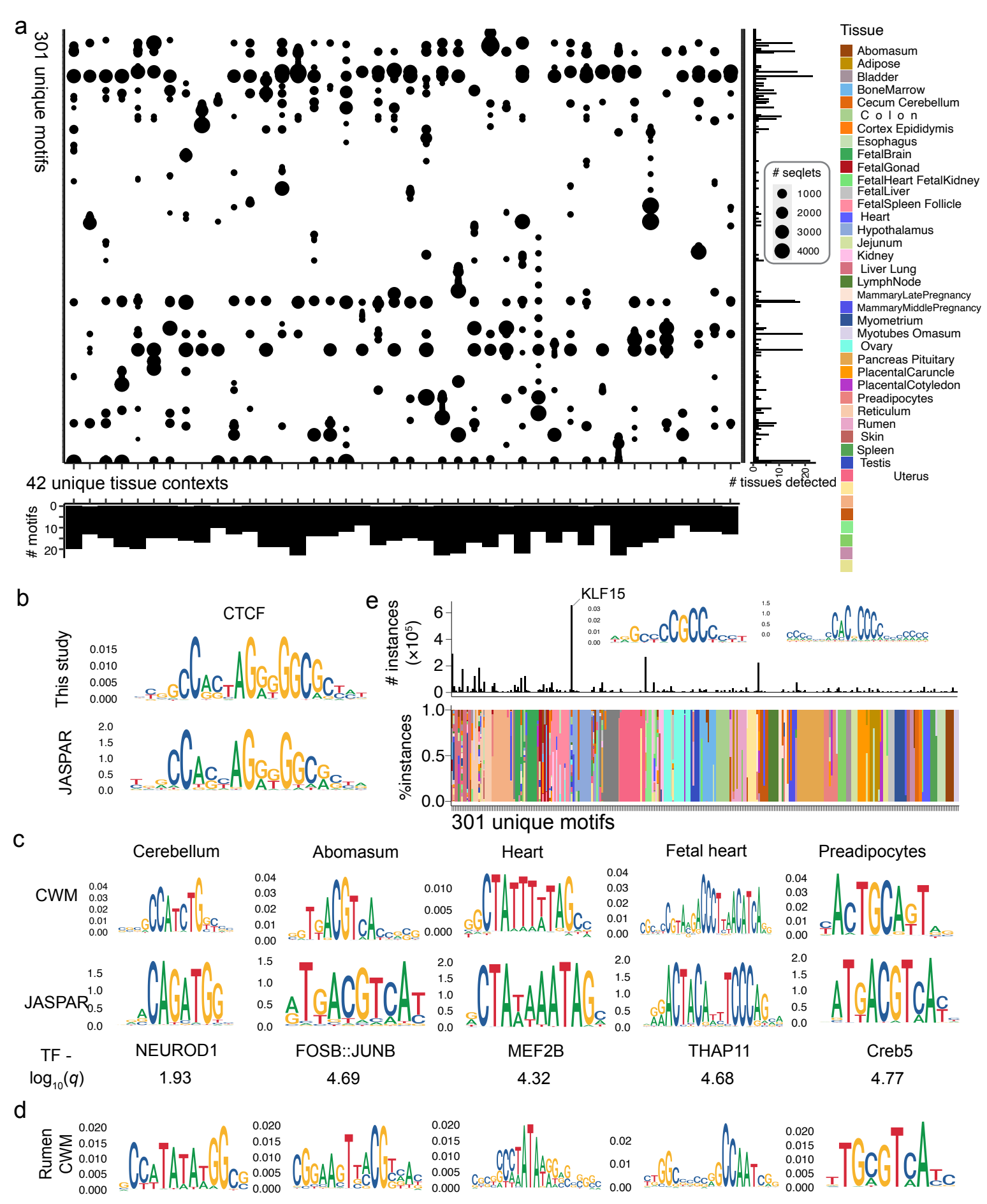


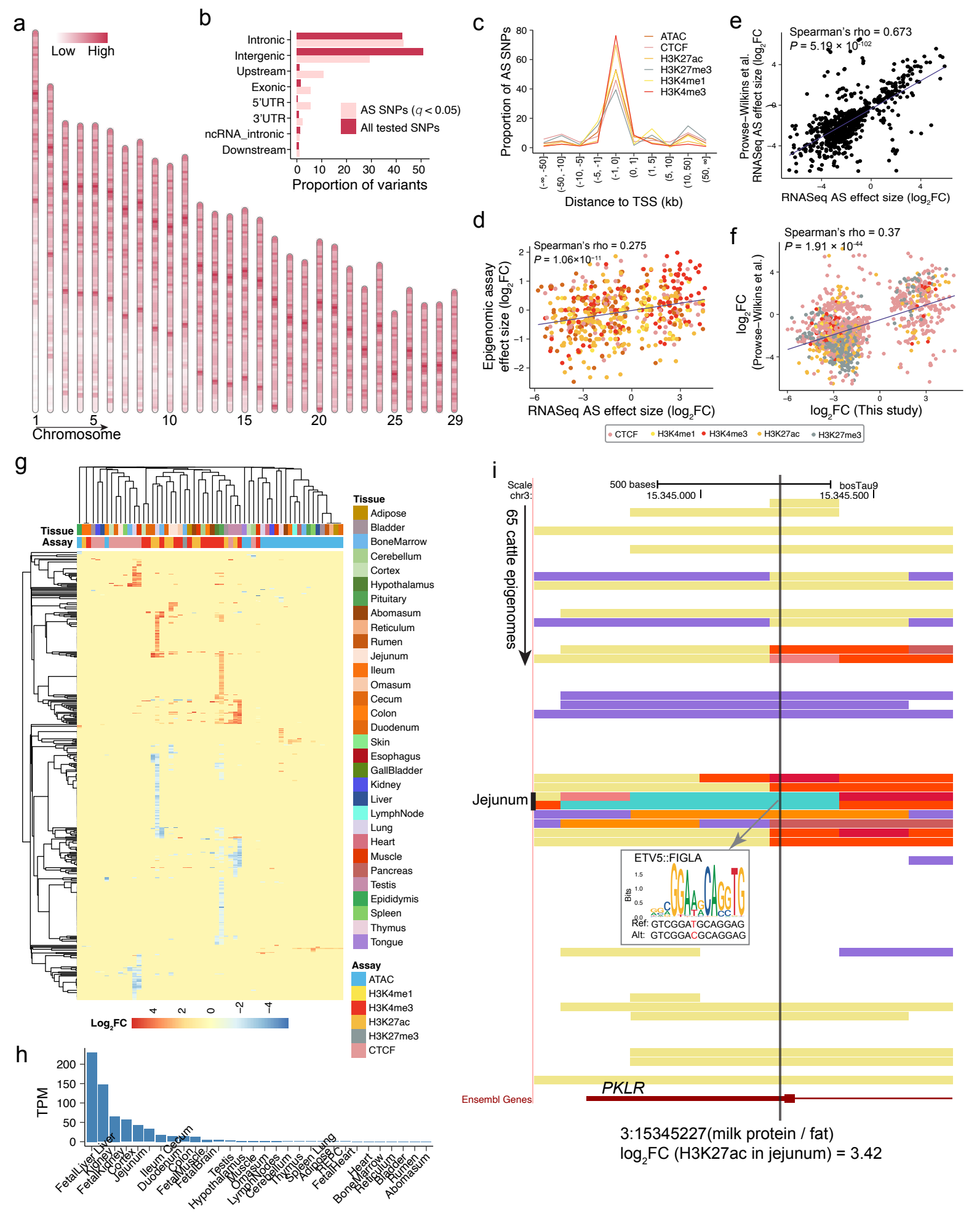
	#
CTCF	104
H3K27ac	113
H3K4me1	120
H3K4me3	122
H3K27me3	101
H3K9me3	20
H3K36me3	27
ATAC	216
WGBS	91

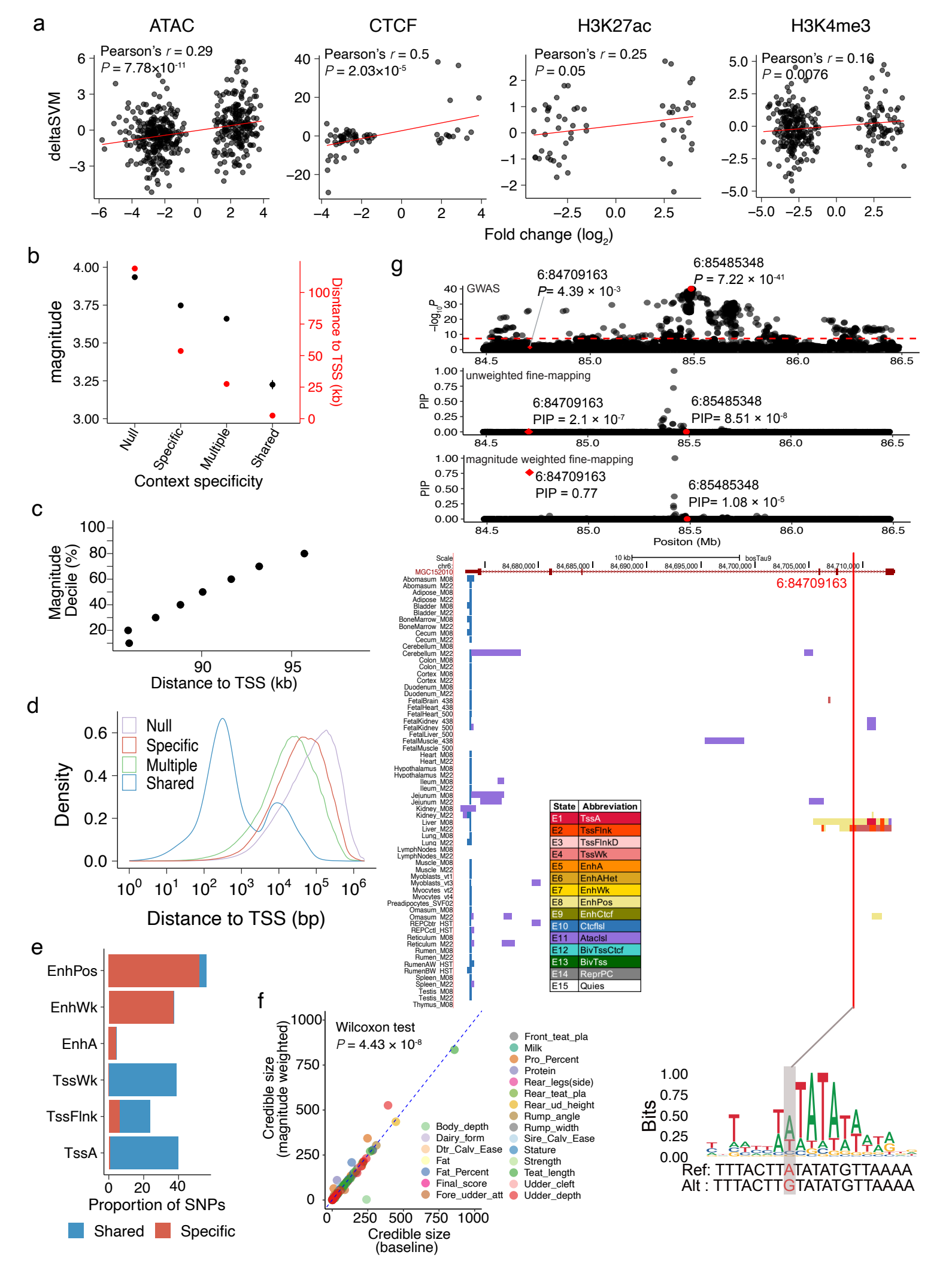


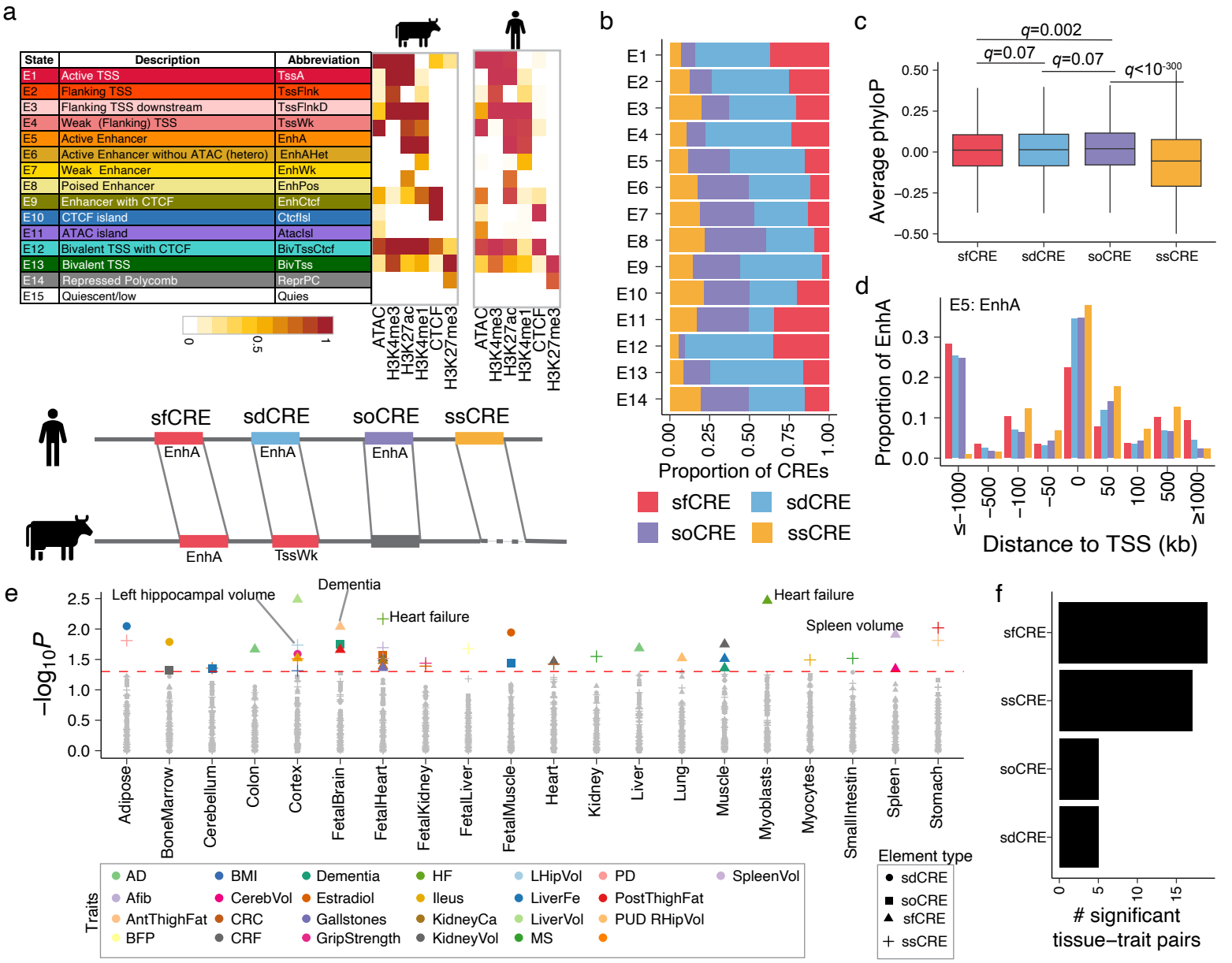



a
Emission
Description
Abbreviation
TSS
Cov.
Annotation overlap
Expr.
E1 Active TSS TssA 0.5±0.12
E2 Flanking TSS TssFlnk 0.31±0.21
E3 Weak (Flanking) TSS TssWk 0.41±0.26
E4 Active Enhancer EnhA 0.69±0.56
E5 Weak Enhancer EnhWk 0.87±0.39
E6 Poised Enhancer EnhPos 0.77±0.38
E7 Bivalent TSS BivTss 0.19±0.17
E8 Bivalent Enhancer BivEnh 0.11±0.06
E9 Repressed Polycomb ReprPC 0.59±0.45
E10 Quiescent/low Quies 87.22±1.51
H3K4me3
H3K27ac
H3K4me1
H3K27me3
Relative enrichment
-2kb
0
2kb
Genome % (average)
5UTRs
TSS
Gene
CDS
Intron
FirstExon
Exon
LastExon
3UTRs
Expr. genes
Repr. genes
Enrichment for non-coding mammalian conserved elements (GERP)
b
EnhA genome coverage
Kern et al. (34%)
This study (66%)
c
mean = 7.14
#genes
#genes per enhancer
d
mean = 7.00
#enhancers
#enhancers per gene
e
mean = 168kb
#element–gene pairs(10^6)
Distance between element and gene(10^6)
f
Rumen
Muscle
FetalHeart
Heart
FetalMuscle
Myoblasts
REPC
Spleen
Myocytes
Jejunum
Liver
BoneMarrow
LymphNodes
Cerebellum
Abomasum
Colon
Cortex
FetalLiver
Kidney
Lung
Ileum
FetalBrain
Bladder
Testis
Cecum
Thymus
Preadipocytes
Adipose
Hypothalamus
FetalKidney
Duodenum
Reticulum
Omasum
tissue specific EnhA
g
-log10FDR
MyoG
Myf5
Tcf12
MyoD
NeuroG2/NeuroD1
NF1
Ascl1 Ascl1
MYB73
Mef2d
Mef2a
Mef2c
Mef2b
NPAS
c-Myc
Sox2
Sox4
Sox6
FOXA1
PQM-1
Nr5a2
Jun-AP1
Fos/Fra1/Fra2
Fosl2
MYB73
Tissue
Adipose
BoneMarrow
Cerebellum
Cortex
Abomasum
Jejunum
Ileum
REPC
Kidney
Liver
FetalLiver
LymphNodes
Heart
Muscle
FetalMuscle
FetalHeart
Myoblasts
Myocytes
Spleen
h
Adult muscle
Fetal muscle
Myoblasts
27255 (30.8%)
1939 (2.2%)
13425 (15.2%)
5076 (5.7%)
6361 (7.2%)
4441 (5.0%)
29946 (33.9%)
i
RNA-seq (Gene expression)
ATAC-seq (Chromatin accessibility)
WGBS (DNA methylation)
Chromatin states
Ensembl Genes
Fetal muscle
Myoblasts
Adult muscle
MYF6
MYF5


a
301 unique motifs
42 unique tissue contexts
motifs
tissues detected
seqlets
1000
2000
3000
4000
Tissue
Abomasum
Adipose
Bladder
BoneMarrow
Cecum Cerebellum
Colon
Cortex Epididymis
Esophagus
FetalBrain
FetalGonad
FetalHeart FetalKidney
FetalLiver
FetalSpleen Follicle
Heart
Hypothalamus
Jejunum
Kidney
Liver Lung
LymphNode
MammaryLatePregnancy
MammaryMiddlePregnancy
Myometrium
Myotubes Omasum
Ovary
Pancreas Pituitary
PlacentalCaruncle
PlacentalCotyledon
Preadipocytes
Reticulum
Rumen
Skin
Spleen
Testis
Uterus
b
CTCF
This study
JASPAR
e
KLF15
instances (×10^5)
%instances
301 unique motifs
c
Cerebellum
Abomasum
Heart
Fetal heart
Preadipocytes
CWM
JASPAR
TF - log10(q)
NEUROD1
1.93
FOSB::JUNB
4.69
MEF2B
4.32
THAP11
4.68
Creb5
4.77
d
Rumen CWM


a
Low
High
Chromosome
b
Intronic
Intergenic
Upstream
Exonic
5'UTR
3'UTR
ncRNA_intronic
Downstream
AS SNPs (q < 0.05)
All tested SNPs
Proportion of variants
c
Proportion of AS SNPs
ATAC
CTCF
H3K27ac
H3K27me3
H3K4me1
H3K4me3
Distance to TSS (kb)
d
Spearman's rho = 0.275
P = 1.06×10-11
Epigenomic assay effect size (log2FC)
RNASeq AS effect size (log2FC)
e
Spearman's rho = 0.673
P = 5.19 × 10-102
Prowse–Wilkins et al. RNASeq AS effect size (log2FC)
RNASeq AS effect size (log2FC)
f
Spearman's rho = 0.37
P = 1.91 × 10-44
log2FC (Prowse–Wilkins et al.)
log2FC (This study)
CTCF
H3K4me1
H3K4me3
H3K27ac
H3K27me3
g
Tissue
Assay
Log2FC
Adipose
Bladder
BoneMarrow
Cerebellum
Cortex
Hypothalamus
Pituitary
Abomasum
Reticulum
Rumen
Jejunum
Ileum
Omasum
Cecum
Colon
Duodenum
Skin
Esophagus
GallBladder
Kidney
Liver
LymphNode
Lung
Heart
Muscle
Pancreas
Testis
Epididymis
Spleen
Thymus
Tongue
ATAC
H3K4me1
H3K4me3
H3K27ac
H3K27me3
CTCF
h
TPM
FetalLiver
Liver
Kidney
FetalKidney
Cortex
Jejunum
Ileum
Cecum
Duodenum
Colon
FetalMuscle
FetalBrain
Testis
Hypothalamus
Muscle
Omasum
LymphNodes
Cerebellum
Thymus
Spleen
Lung
Adipose
RPC
FetalHeart
Heart
BoneMarrow
Reticulum
Bladder
Rumen
Abomasum
i
Scale
chr3:
500 bases
15.345.000
15.345.500
bosTau9
65 cattle epigenomes
Jejunum
ETV5::FIGLA
Bits
Ref: GTCGGATGCAGGAG
Alt: GTCGGACGCAGGAG
Ensembl Genes
PKLR
3:15345227(milk protein / fat)
log2FC (H3K27ac in jejunum) = 3.42


a
ATAC
CTCF
H3K27ac
H3K4me3
Pearson's r = 0.29
P = 7.78×10-11
Pearson's r = 0.5
P = 2.03×10-5
Pearson's r = 0.25
P = 0.05
Pearson's r = 0.16
P = 0.0076
deltaSVM
Fold change (log2)
b
magnitude
Distance to TSS (kb)
Null
Specific
Multiple
Shared
Context specificity
c
Magnitude Decile (%)
Distance to TSS (kb)
d
Null
Specific
Multiple
Shared
Density
Distance to TSS (bp)
e
EnhPos
EnhWk
EnhA
TssWk
TssFlnk
TssA
Proportion of SNPs
Shared
Specific
f
Wilcoxon test
P = 4.43 × 10-8
Credible size (magnitude weighted)
Credible size (baseline)
Body_depth
Dairy_form
Dtr_Calv_Ease
Fat
Fat_Percent
Final_score
Fore_udder_att
Front_teat_pla
Milk
Pro_Percent
Protein
Rear_legs(side)
Rear_teat_pla
Rear_ud_height
Rump_angle
Rump_width
Sire_Calv_Ease
Stature
Strength
Teat_length
Udder_cleft
Udder_depth
g
GWAS
6:84709163
P= 4.39 × 10-3
6:85485348
P = 7.22 × 10-41
unweighted fine-mapping
6:84709163
PIP = 2.1 × 10-7
6:85485348
PIP= 8.51 × 10-8
magnitude weighted fine-mapping
6:84709163
PIP = 0.77
6:85485348
PIP= 1.08 × 10-5
Positon (Mb)
6:84709163
State
Abbreviation
E1 TssA
E2 TssFlnk
E3 TssFlnkD
E4 TssWk
E5 EnhA
E6 EnhAHet
E7 EnhWk
E8 EnhPos
E9 EnhCtcf
E10 CtcfIsl
E11 AtacIsl
E12 BivTssCtcf
E13 BivTss
E14 ReprPC
E15 Quies
Bits
Ref: TTTACTTATATATGTTAAAA
Alt : TTTACTTGTATATGTTAAAA


a
State
Description
Abbreviation
E1 Active TSS TssA
E2 Flanking TSS TssFlnk
E3 Flanking TSS downstream TssFlnkD
E4 Weak (Flanking) TSS TssWk
E5 Active Enhancer EnhA
E6 Active Enhancer withou ATAC (hetero) EnhAHet
E7 Weak Enhancer EnhWk
E8 Poised Enhancer EnhPos
E9 Enhancer with CTCF EnhCtcf
E10 CTCF island CtcfIsl
E11 ATAC island AtacIsl
E12 Bivalent TSS with CTCF BivTssCtcf
E13 Bivalent TSS BivTss
E14 Repressed Polycomb ReprPC
E15 Quiescent/low Quies
ATAC
H3K4me3
H3K27ac
H3K4me1
CTCF
H3K27me3
0
0.5
1
sfCRE
sdCRE
soCRE
ssCRE
EnhA
TssWk
b
E1
E2
E3
E4
E5
E6
E7
E8
E9
E10
E11
E12
E13
E14
0.00
0.25
0.50
0.75
1.00
Proportion of CREs
sfCRE
sdCRE
soCRE
ssCRE
c
q=0.002
q=0.07
q=0.07
q<10-300
Average phyloP
0.50
0.25
0.00
−0.25
−0.50
sfCRE
sdCRE
soCRE
ssCRE
d
E5: EnhA
Proportion of EnhA
0.3
0.2
0.1
0.0
≤−1000
−500
−100
−50
0
50
100
500
≥1000
Distance to TSS (kb)
e
Left hippocampal volume
Dementia
Heart failure
Heart failure
Spleen volume
−log10P
2.5
2.0
1.5
1.0
0.5
0.0
Adipose
BoneMarrow
Cerebellum
Colon
Cortex
FetalBrain
FetalHeart
FetalKidney
FetalLiver
FetalMuscle
Heart
Kidney
Liver
Lung
Muscle
Myoblasts
Myocytes
SmallIntestin
Spleen
Stomach
Traits
AD
Afib
AntThighFat
BFP
BMI
CerebVol
CRC
CRF
Dementia
Estradiol
Gallstones
GripStrength
HF
Ileus
KidneyCa
KidneyVol
LHipVol
LiverFe
LiverVol
MS
PD
PostThighFat
PUD RHipVol
SpleenVol
Element type
sdCRE
soCRE
sfCRE
ssCRE
f
sfCRE
ssCRE
soCRE
sdCRE
0
5
10
15
significant
tissue−trait pairs