Title

Functional annotation of the animal genomes: An integrated annotation resource for the horse

Journal

PLOS Genetics, 19(3)

ISSN

1553-7390

Authors

Peng, Sichong

Dahlgren, Anna R

Donnelly, Callum G

et al.

Publication Date

2023

DOI

10.1371/journal.pgen.1010468

PLOS GENETICS

RESEARCH ARTICLE

Functional annotation of the animal genomes: An integrated annotation resource for the horse

Sichong Peng[1], Anna R. Dahlgren[1], Callum G. Donnelly[1], Erin N. Hales[1], Jessica L. Petersen[2], Rebecca R. Bellone[1,3], Ted Kalbfleisch[4], Carrie J. Finno[1]*

1 Department of Population Health and Reproduction, School of Veterinary Medicine, University of California-Davis, Davis, California, United States of America, **2** Department of Animal Science, University of Nebraska—Lincoln, Lincoln, Nebraska, United States of America, **3** Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California-Davis, Davis, California, United States of America, **4** Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, United States of America

* cjfinno@ucdavis.edu

Citation: Peng S, Dahlgren AR, Donnelly CG, Hales EN, Petersen JL, Bellone RR, et al. (2023) Functional annotation of the animal genomes: An integrated annotation resource for the horse. PLoS Genet 19(3): e1010468. https://doi.org/10.1371/journal.pgen.1010468

Editor: Tosso Leeb, University of Bern, SWITZERLAND

Received: October 11, 2022

Accepted: January 28, 2023

Published: March 2, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1010468

Data Availability Statement: RNA-seq data can be accessed from ENA and SRA under the accession number PRJEB26787 (female tissues) and PRJEB53382 (male tissues). Iso-seq data can be

Abstract

The genomic sequence of the horse has been available since 2009, providing critical resources for discovering important genomic variants regarding both animal health and population structures. However, to fully understand the functional implications of these variants, detailed annotation of the horse genome is required. Due to the limited availability of functional data for the equine genome, as well as the technical limitations of short-read RNA-seq, existing annotation of the equine genome contains limited information about important aspects of gene regulation, such as alternate isoforms and regulatory elements, which are either not transcribed or transcribed at a very low level. To solve above problems, the Functional Annotation of the Animal Genomes (FAANG) project proposed a systemic approach to tissue collection, phenotyping, and data generation, adopting the blueprint laid out by the Encyclopedia of DNA Elements (ENCODE) project. Here we detail the first comprehensive overview of gene expression and regulation in the horse, presenting 39,625 novel transcripts, 84,613 candidate cis-regulatory elements (CRE) and their target genes, 332,115 open chromatin regions genome wide across a diverse set of tissues. We showed substantial concordance between chromatin accessibility, chromatin states in different genic features and gene expression. This comprehensive and expanded set of genomics resources will provide the equine research community ample opportunities for studies of complex traits in the horse.

Author summary

Functional annotation of a reference genome provides critical information that pertains the tissue-specific gene expression and regulation. Non-model organisms often rely on existing annotations of human and mouse genomes and the conservation between species for their genome annotation. This approach has limited power in annotating transcripts

accessed from ENA and SRA under the accession number PRJEB53020. ATAC-seq data can be accessed from ENA and SRA under the accession number PRJEB53037 (https://data.faang.org/dataset/PRJEB53037). Histone modification peaks can be accessed from FAANG data portal (https://data.faang.org/) under the accession number PRJEB35307 (https://data.faang.org/dataset/PRJEB35307). CTCF ChIP-seq data can be accessed from SRA/ENA under project accession PRJEB41079. Histone ChIP-seq data were published by Kingsley et al. (https://doi.org/10.3390/genes11010003) and Barber et al (thesis; https://digitalcommons.unl.edu/animalscidiss/233/). The GO ontology database link used for the pathway analyses can be found at 10.5281/zenodo.6399963 and was released on March 3, 2022. The integrated track hub hosted on UCSC genome browser can be accessed at https://genome.ucsc.edu/s/cjfinno/equCab3. All data underlying this track hub can be accessed via UCSC table browser or at https://github.com/FinnoLab/FAANGtracks/tree/main/data.

Funding: This project was supported by the Grayson-Jockey Club Research Foundation (C.J.F., J.L.P, R.R.B.), USDA National Institute of Food and Agriculture Animal Breeding and Functional Annotation of Genomes (A1201) Grant 2019-67015-29340 (J.L.P, C.J.F., R.B.) as well as NRSP-8 Species Coordinator Funds from the USDA National Institute of Food and Agriculture (C.J.F., J.L.P, R.R.B.), and the UC Davis Center for Equine Health with funds provided by the State of California pari-mutuel fund and contributions by private donors (C.J.F., J.L.P, R.R.B.). Additional support for C.J.F. was provided by NIH NCATS L40 TR001136. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

and regulatory elements that are less evolutionarily conserved. Such are the cases of alternatively spliced isoforms and enhancer elements. In a large-scale collaborated effort, Functional Annotation of Animal Genome (FAANG) aims to generate species-specific and tissue-aware functional annotation for farm animals. In this study, we present the overall annotation efforts and findings from the equine FAANG group. This integrated annotation for the horse genome provides, for the first time, a comprehensive overview of gene expression and regulation landscape in nine major equine tissues, as well as an analytical framework for further inclusion of other important tissues.

Introduction

A reference genome for the horse has been available since 2009 [1], with an improved assembly EquCab3.0 available since 2018 [2]. EquCab3.0 contains 3,771 gaps comprising 9 Mb (0.34% of the genome) and has a scaffold N50 of 86 Mb, with 99.7% mammalian Benchmark Universal Single-Copy Orthologs (BUSCO). This high-quality assembly has enabled development of critical tools and many important discoveries in the horse, which were reviewed by Raudsepp *et al.* [3].

Accompanying the reference genome, annotation was made available via the RefSeq [4] and Ensembl [5] annotation pipelines. The latest RefSeq annotation for EquCab3.0 contains 33,146 genes, of which 21,129 are protein coding, with an average isoform-to-gene ratio of 2.3 [6]. The Ensembl annotation contains 29,969 genes, of which 20,955 are protein coding, with an average isoform-to-gene ratio of 2.0 [7]. With limited public mRNA-seq data for the horse, both RefSeq and Ensembl annotation relied heavily on computational prediction and comparative genomics by translating human and mouse annotation to the horse genome. While this approach produced high-quality annotation for most highly conserved protein-coding genes, it was not able to accurately identify many alternate splicing (AS) in multi-exonic genes. This is evident when comparing the isoform-gene ratios annotated in the horse genome (2.3 and 2.0 in RefSeq and Ensembl, respectively) to that annotated in the human genome (4.0) [8]. While this difference can be attributed to vast quantities of transcriptomic data available in human, recent developments in the long-read sequencing technology provided a unique opportunity for non-model organisms to quickly annotate AS without generating a prohibitively large amount of data. Alternative splicing has been shown to drive cell differentiation and tissue-specific functions [9] and variants leading to aberrant AS have been associated with many diseases [10]. While *in-silico* tools exist to predict variant-induced alterations in AS, accurate annotation of AS isoforms is necessary to establish a reference [11]. Recent advances in long-read sequencing technologies have enabled new approaches to experimentally categorize AS across tissues [12]. In particular, Iso-seq has been successfully applied to various species to characterize AS isoforms [13–15].

In eukaryote genomes, DNA is organized in a three-dimensional structure, where nucleosomes are dynamically unpacked in actively transcribed or regulatory regions [16–19]. This dynamic chromatin remodeling constitutes a crucial aspect of gene regulation: cis-regulatory elements are brought near their target regions by formation of chromatin loops and transcription factors (TF) are recruited to exposed DNA elements. Genetic variants altering this regulatory landscape have been demonstrated to have phenotypic effects [20–22]. Therefore, annotating the genome by specifically defining these cis-regulatory elements (CREs) can provide significant context to understanding genetic variations contributing to many important traits in the horse.

Active CREs are typically characterized by a lack of nucleosome binding and therefore, chromatin accessibilities are often used as a proxy for identifying active regulatory elements [23]. The assay for transposase-accessible chromatin using sequencing (ATAC-seq) is a popular method to assess the genome-wide chromatin accessibilities, owing to its simple protocol and quick turn-around time [24]. Several efforts have been made to adapt the original ATAC-seq protocol to tissue [25] and cryopreserved nuclei [26] samples. We previously demonstrated the feasibility of interrogating genome-wide chromatin accessibility using both flash frozen tissues as well as cryopreserved nuclei in the horse [27].

While the complex molecular mechanism through which this process is regulated remains an active field of research, a growing body of evidence points to histone protein post-translational modifications as an important intermediary of transcription regulation [28–30]. Specifically, histone protein 3 lysine 4 mono- and tri-methylation (H3K4me1 and H3K4me3) have been shown to be enriched around the enhancer and promoter regions, respectively [31,32], with known downstream effectors that further regulate gene expression [33–35]. Additionally, H3K27ac is enriched around active elements and associated with higher levels of gene expression [36]. On the other hand, H3K27me3 is usually found around genes that are not active [37]. Taken together, these protein modifications can be strong indicators of functional activities in specific genomic regions.

To improve AS annotation for the horse transcriptome and to systemically categorize these epigenetic features and identify potential CREs in the equine genome, we collected over 80 tissues from four healthy adult Thoroughbred horses as a part of the Functional Annotation of Animal Genome (FAANG) initiative. Detailed phenotyping and tissue collection protocols have been previously reported [38,39]. Nine prioritized tissues (lamina, liver, left lung, left ventricle of heart, longissimus muscle, skin, parietal cortex, testis, and ovary) were used to generate a diverse set of data that represent different aspects of gene expression and regulation. Additional RNA-seq data from 57 other tissues, generated from these same horses as a result of a community-driven effort, were used to compare our revised FAANG annotation with previous Ensembl and NCBI annotations. Here we present an integrated analysis of the equine FAANG dataset.

Results

Long-read data improved transcriptome annotation

Using Iso-seq data from nine prioritized tissues, we assembled a transcriptome with improved AS and 3' transcription termination site (TTS) annotation for Equcab3 [2]. This transcriptome contained 56,672 transcripts including 39,625 novel transcripts. A majority of these transcripts (51,639) were multi-exonic. Of these novel transcripts, 30,964 (78%) were AS isoforms with either novel combinations of known splice junctions (6,330) or novel splice junctions (24,634). Of the 17,407 known transcripts, 12,470 contain splice junctions fully matched to a reference transcript annotated in the Ensembl gene annotation [5,7] for EquCab3 (full-splice match, FSM). The majority (9,924 or 79.6%) of these transcripts extended reference annotation at the 3' end, with 4,232 transcripts having TTS more than 1 kb downstream of Ensembl annotated TTS (**Fig 1A**). The remaining 4,937 known transcripts lacked known splice junctions at either 5' or 3' end (incomplete-splice match, ISM), of which 2,395 extended the reference annotation at the 3' end (**Fig 1A**). At the transcription start sites (TSS), the majority (98.4%) of transcripts had higher RNA-seq coverage in 100bp windows downstream of TSS than upstream and 89.1% had at least twice coverage in 100bp windows downstream of TSS than upstream (**Fig 1B**).

The tissue-specific expression of these transcripts was quantified using short-read RNA-seq data from 57 tissues of the same animals, nine of which were the same tissues used to generate

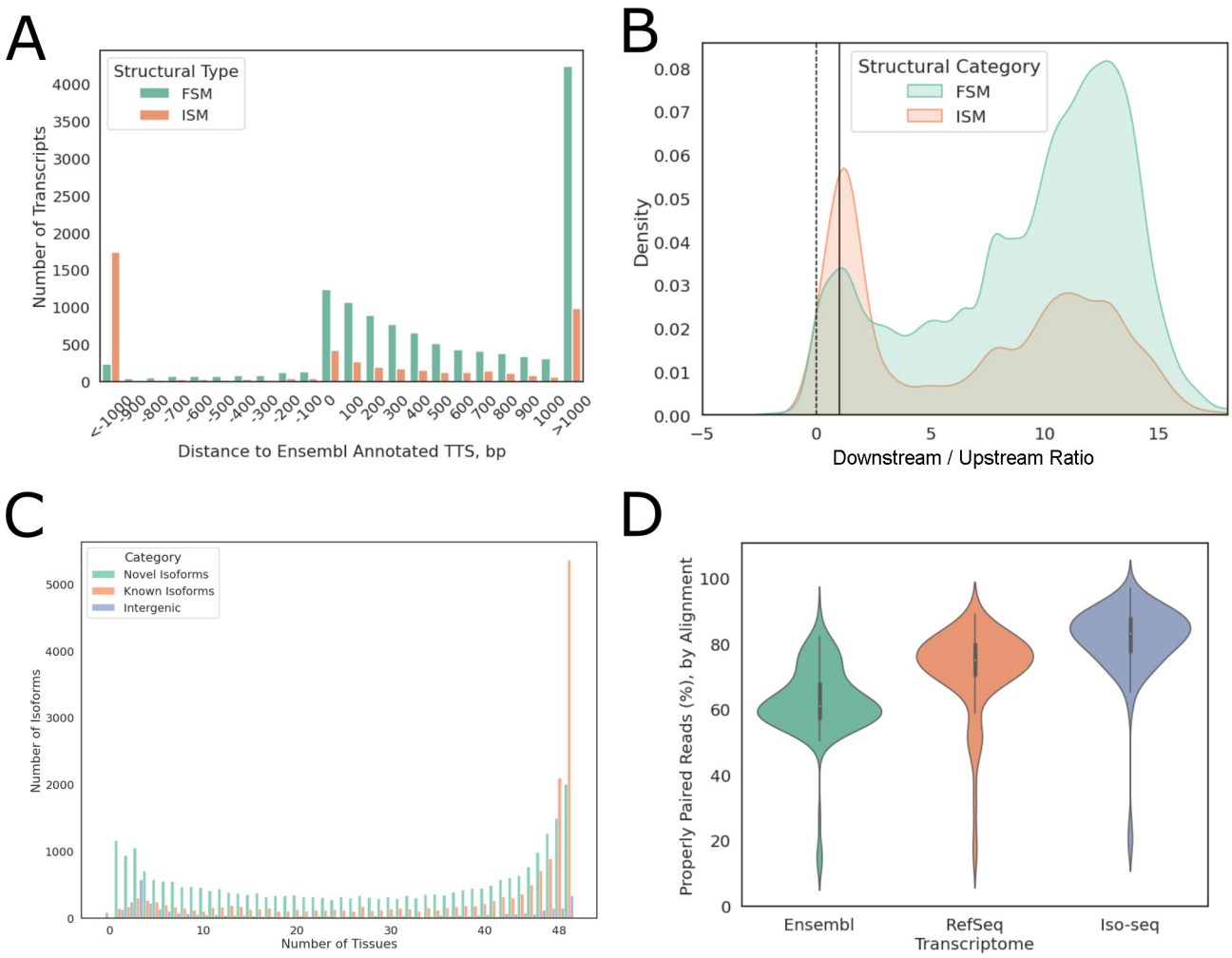

Fig 1. Iso-seq transcriptome improved gene annotation. (A) Distance between Iso-seq Ensembl annotated TTS, negative values indicate shorter 3' ends. Eg. -1000 indicates that Iso-seq annotated TTS is 1000 bp upstream of Ensembl annotated TTS. (B) Log_2 of 100 bp downstream as compared to upstream of TSS RNA-seq coverage. Positive ratios indicate higher coverage downstream of TSS. The dotted line indicates equal coverage up- and down-stream of TSS while the solid line indicates 100% higher coverage downstream of the TSS than upstream. (C) Distribution of known vs. novel transcripts detected in different numbers of tissues. (D) Distribution of mapping rates against Iso-seq transcriptome vs. Ensembl or RefSeq transcriptome across 57 RNA-seq samples. FSM: full-splice match; ISM: incomplete-splice match; TTS: transcription start site; TSS: transcription termination site.

https://doi.org/10.1371/journal.pgen.1010468.g001

Iso-seq data. Approximately 78% of known isoforms were expressed in at least half of the tissues sequenced, while novel isoforms of known genes and novel intergenic transcripts each showed a bimodal distribution, with 44.3% of novel isoforms and 56.8% of intergenic transcripts detected in less than half of the tissues (Fig 1C). We also noted that, on average, 61.4% (33.3%-70.9%) of multi-isoform genes expressed more than one isoform in any given tissue and had different dominant major isoforms (isoform with highest relative expression of a given gene), depending on the tissue type.

The completeness of this transcriptome annotation was assessed by aligning RNA-seq reads directly to transcriptome sequences. Compared to the Ensembl and RefSeq annotations, this Iso-seq transcriptome showed substantial improvement in mapping rates, as measured by percentage of properly paired reads, with a median mapping rate of 83.25% as compared to 61.10% and 75.15% when using Ensembl and RefSeq annotation, respectively (Fig 1D).

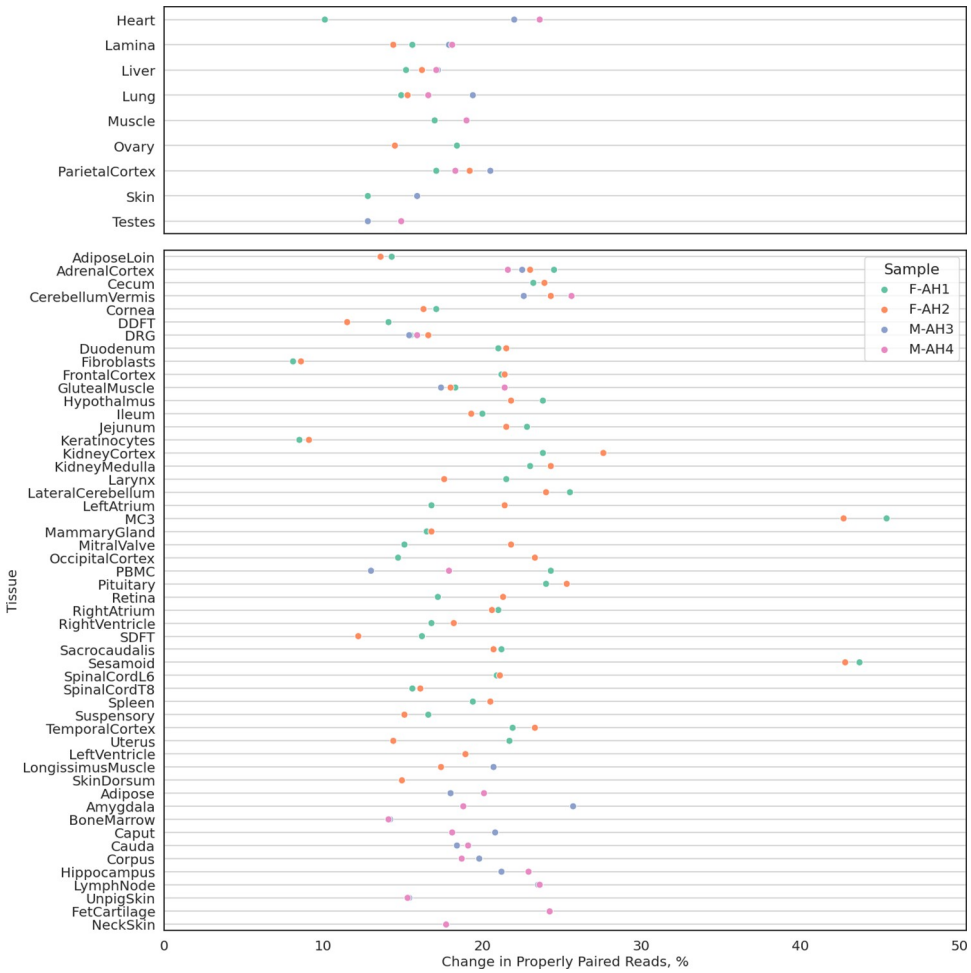

Fig 2. Comparison of FAANG, RefSeq and Ensembl equine transcriptomes. Changes in percentages of properly paired reads aligned to combined FAANG transcriptome when compared to Ensembl or RefSeq transcriptomes, whichever has higher percentage.

https://doi.org/10.1371/journal.pgen.1010468.g002

Since only nine tissues were used to construct this Iso-seq transcriptome, it was expected that many tissue-specific genes and transcripts will be missing. Indeed, several tissues had decreased mapping rates when aligned to the Iso-seq transcriptome as compared to Ensembl or RefSeq (S1 Fig), especially those with large stem cell populations. Therefore, this Iso-seq transcriptome was merged with Ensembl and RefSeq transcriptome to construct a more complete transcriptome annotation, termed the FAANG transcriptome. RNA-seq alignment data indicated that the FAANG transcriptome was substantially more complete than the existing annotations, with an average 19.5% (8–45%) increase in mapping rates across all sequenced tissues (Fig 2). The FAANG transcriptome consisted of 153,492 transcripts (of which 128,723 were multi-exonic) from 36,239 genes, with a gene-isoform ratio of 4.2.

Tissue-specific open chromatin annotation

Chromatin accessibility was profiled from the same nine tissues (adipose, heart, lamina, liver, lung, ovary, testis, muscle, parietal cortex). Most libraries contained 60% to > 90% unique reads, with the exception of liver and cerebral cortex samples. Data from the female liver samples were generated from our previous study, where excessive mitochondria contamination

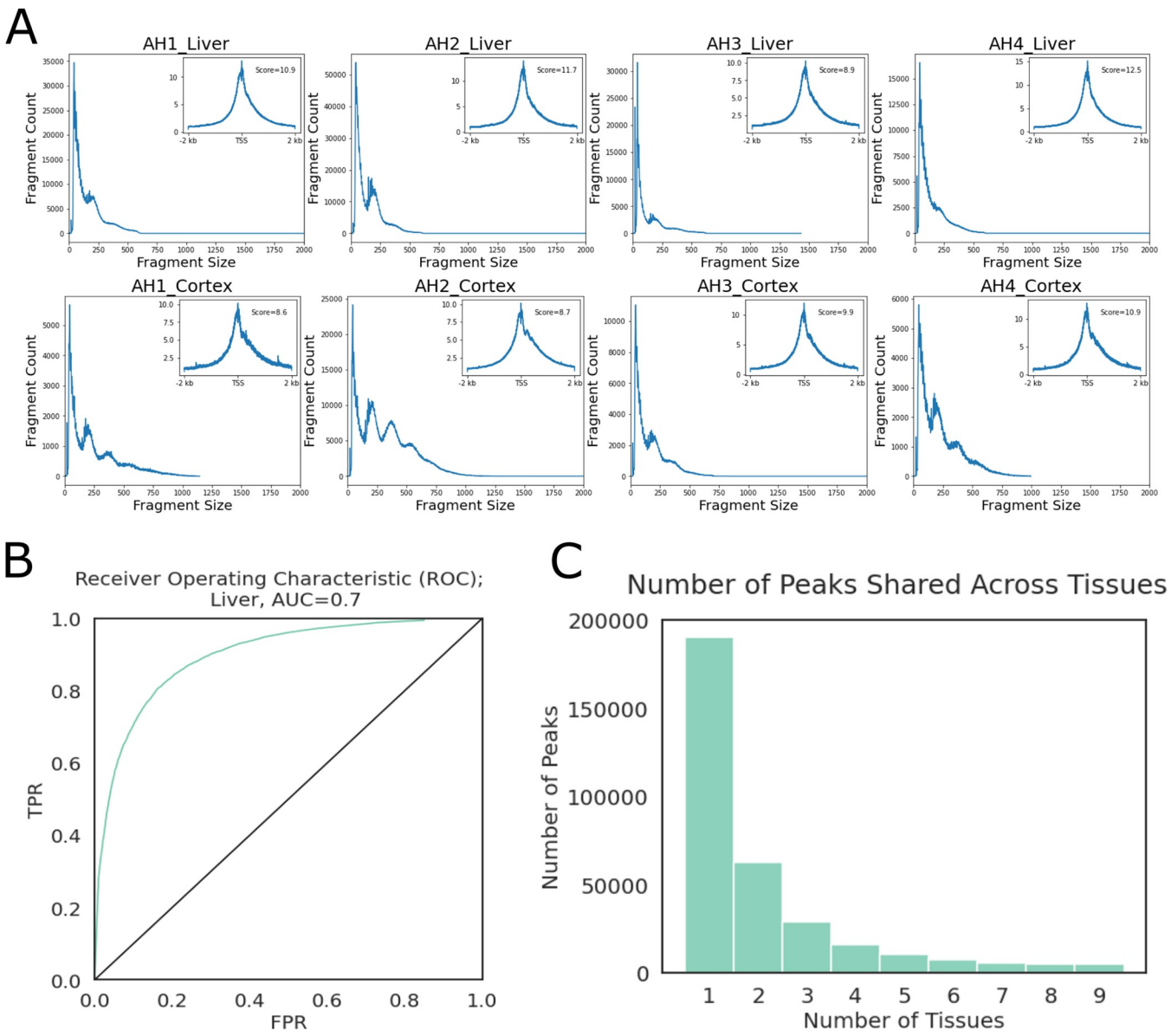

Fig 3. ATAC-seq Quality Control. (A) Fragment size distributions and TSS enrichment plots (upper right inset plots) of liver and cerebral cortex samples as examples. (B) ROC plot of liver peaks, as a representative example across tissues. TPR: true positive rate, FPR: false positive rate, AUC: area under curve. (C) Number of peaks that were observed in a given number of tissues.

https://doi.org/10.1371/journal.pgen.1010468.g003

led to lower library complexities and resequencing was performed to reach desired unique read counts [27]. After removing polymerase chain reaction (PCR) duplicate reads, all libraries contained less than 20% of mitochondria reads. Despite lower library complexities, both liver and cerebral cortex samples showed clear nucleosomal periodicities and high enrichment around TSS (**Fig 3A**).

Peaks were called by MACS3 [40,41] single-end BED mode using both ends of aligned fragments. Accuracies of peak calling were estimated using published histone ChIP-seq data from the same tissues [42]. Briefly, open chromatin peaks were intersected with "true positive" (TP, H3K4me1 or H3K4me3 peaks overlapping H3K27ac) and "true negative" (TN, H3K27me3 peaks) peak sets to calculate true positive rates (TPR) and false positive rates (FPR). Since testes

Table 1. Open Chromatin Peak Metrics. Number of peaks identified in each tissue before and after filtering as well as union and tissue-specific peaks.

	Merged Raw Peak Count	Cutoff	TP	FP	TPR	FPR	Remaining Peak Count	Tissue Specific
Adipose	941,236	67	10,595	2,008	0.59	0.25	77,655	22,884
Cortex	435,514	114	14,109	1,959	0.78	0.25	65,583	18,249
Heart	581,396	85	14,955	2,226	0.83	0.25	86,368	19,730
Lamina	722,387	89	9,423	1,765	0.48	0.25	63,136	21,805
Liver	557,874	76	16,078	2,815	0.87	0.25	95,048	31,460
Lung	522,294	103	13,672	1,957	0.76	0.25	59,024	8,447
Muscle	360,298	98	15,891	2,090	0.86	0.25	74,285	19,772
Ovary	463,426	109	12,583	1,767	0.64	0.25	66,726	16,588
Testis	520,160	N/A	N/A	N/A	N/A	N/A	78,164*	31,880
Union	332,115							
Conserved	5,080							

* Testis peaks were filtered by score at 85th quantile since no histone peaks were available for this tissue

TP: number of true positive peaks, FP: number of false positive peaks; TPR: true positive rate; FPR: false positive rate; cutoff: cutoff score below which peaks were removed from final peak set; union: peaks found in any tissue, after iterative merging; ubiquitous: peaks found in all nine tissues

https://doi.org/10.1371/journal.pgen.1010468.t001

were not included in the histone ChIP-seq dataset from Kingsley et al. [42], testes' libraries were not evaluated at this step. Area under curve (AUC) values of at least 0.6 were achieved for all tissues evaluated (**Figs 3B and S2**). Cutoff scores were set at 25% FPR to filter a final set of peaks for each tissue, except testis. After filtering, the evaluated tissues had 59k-95k peaks remaining (**Table 1**). Testis and liver had the highest amounts of tissue-specific peaks (31,880 and 31,460, respectively), while lung had the lowest number of tissue-specific peaks (8,447). Only a very small number of peaks were conserved across examined tissues (**Fig 3C**).

These open chromatin peaks were annotated by their overlapping genic features as promoter-TSS (2 kb up- or down-stream of a TSS), exon, intron, TTS, and intergenic peaks. Open chromatin peaks across tissues were enriched in TSS, TTS, and exon regions (2.9-, 1.3-, and 1.3-fold enrichment, respectively). This enrichment was more apparent among peaks conserved across tissues (**Fig 4**). Gene ontology (GO) terms overrepresented in genes associated with these conserved peaks were all essential housekeeping biological processes such as TOR signaling and kinase activity (**S1 Table**). For each tissue, 11–22% peaks were located within promoter-TSS regions. However, the same pattern was not observed in tissue-specific peaks. Only 3–5% of tissue specific peaks were in the promoter-TSS regions, while substantially more peaks were located in intronic or intergenic regions (17–22% intronic, 21–34% intergenic

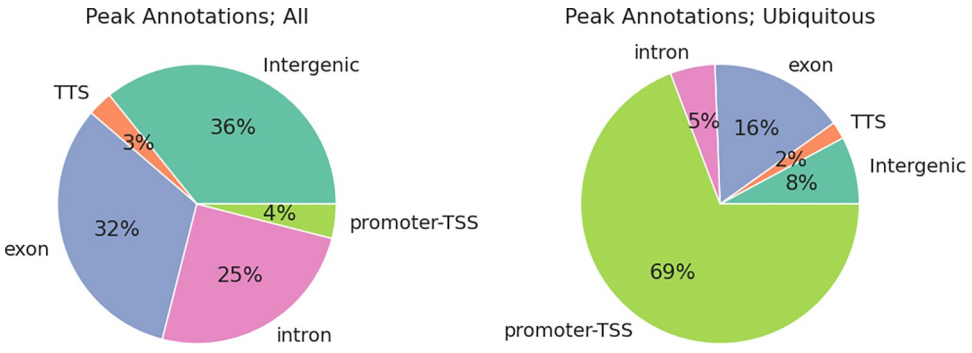

Fig 4. Peak annotations. Left: Composition of peaks by annotation in union peaks; Right: Composition of peaks by annotation in ubiquitous peaks.

https://doi.org/10.1371/journal.pgen.1010468.g004

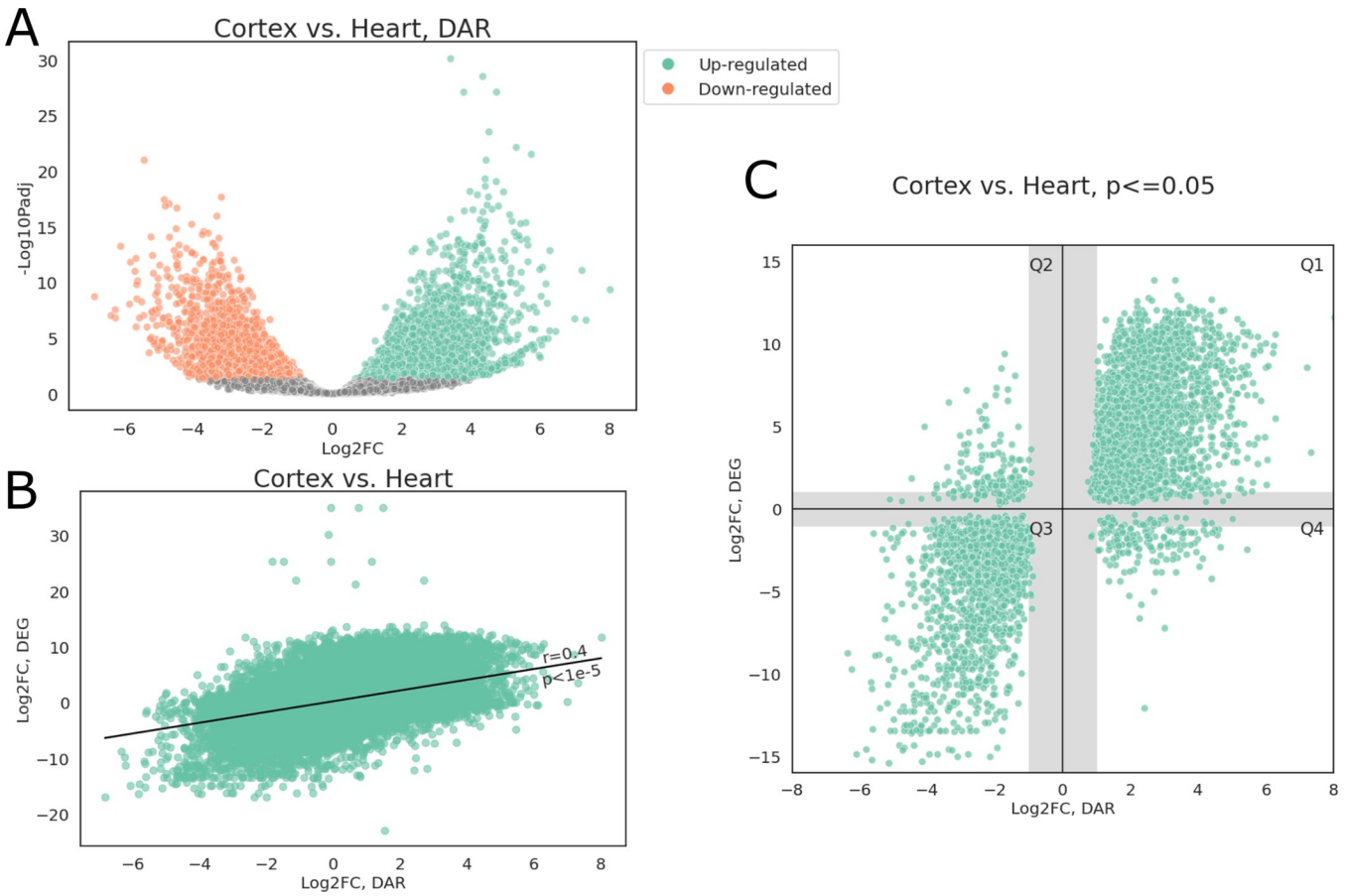

Fig 5. Differential accessibility analysis. (A) Volcano plot of open chromatin peaks; peaks with FDR adjusted p<0.05 and $|\log_2 FC|>1$ were colored by direction of accessibility change, positive $\log_2 FC$ indicate greater accessibility in cerebral cortex. (B) Scatter plot of $\log_2 FC$ from DEG and DAR analyses, Pearson correlation r = 0.4. (C) Scatter plot of $\log_2 FC$ from DEG and DAR analyses, only those with both FDR adjusted p<0.05 were plotted; shaded areas indicate regions where either FC_{DEG} or FC_{DAR} was under 2-fold.

https://doi.org/10.1371/journal.pgen.1010468.g005

peaks across tissues; 23–26% intronic, 23–45% intergenic tissue-specific peaks). Motif analyses of these intergenic regions revealed a diverse range of TF binding sites, such as hepatocyte nuclear factor-4 alpha (HNF-4α) and estrogen-related receptor alpha (ERRα) binding sites in liver-specific intergenic open chromatin regions, myocyte enhancer factor-2 (MEF2) family TF binding sites in heart-specific open chromatin regions, and SRY-related HMG-box (SOX) family TF binding sites in cerebral cortex-specific open chromatin regions. (S2 Table).

Unsurprisingly, TSS accessibility showed significant correlations with their corresponding gene expression. Fig 5 shows a representative differential accessibility and expression analysis of parietal cortex and heart tissues. Overall, approximately 16% of peaks showed differential accessibility (FDR adjusted p<0.05), with 25,144 peaks (7.6%) more accessible in cortex and 29,206 peaks (8.8%) more accessible in heart (Fig 5A). The $\log_2$ fold-change ($\log_2 FC$) of differentially accessible regions (DAR) was significantly correlated with $\log_2 FC$ of differentially expressed genes (DEG) in the same cortex and heart samples (one-sided Wald test, $p<1 \times 10^{-5}$, Pearson correlation coefficient r = 0.4, Fig 5B). After selecting peak-gene pairs whose FDR adjusted p values from both DAR and DEG analyses were below 0.05, we observed that most genes were located in quadrants 1 and 3 (Q1 and Q3 respectively), showing concordant changes in promoter-TSS accessibility and gene expression (Fig 5C). GO enrichment analyses

showed that Q1 genes were primarily associated with neural activities while muscular and cardiac related GO terms were enriched among Q3 genes (**S3 and S4 Tables**). There were also 175 and 177 genes in Q2 and Q4, respectively. GO Terms related to synaptic activity were enriched in Q2 (**S5 Table**) while Q4 genes were overrepresented in actin-filament based processes.

Cis-regulatory element annotation

Chromatin states were first identified using four major histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3) as well as CTCF binding from the same nine tissues. Overall, 14 unique states, corresponding to enhancer, promoter, and insulator states of various degrees of activities, as well as polycomb repressed state were identified (**Fig 6A**). Notably, the CTCF-bound active TSS state (state *4*), co-enriched with CTCF and active promoter marks (H3K4me3 and H3K27ac), was highly enriched around TSS, whereas the CTCF-less active TSS state (state *3*) was more enriched at approximately 500 bp up- and down-stream of TSS. Collectively, states with assayed epigenetic signals (states *1–13*) covered up to 20% of the genome, with the polycomb repressed state (state *13*) covering the largest portion of the genome across tissues, followed by enhancer states (states *6–10*, **Fig 6B**). While promoter states only accounted for 3–5% of the genome, or around 20% of all annotated states, they comprised over 50% of states annotated at TSS regions (**Fig 6B and 6C**).

To correlate gene expression with chromatin state annotation, companion RNA-seq data for each tissue from FAANG was used to quantify the equine FAANG transcriptome via

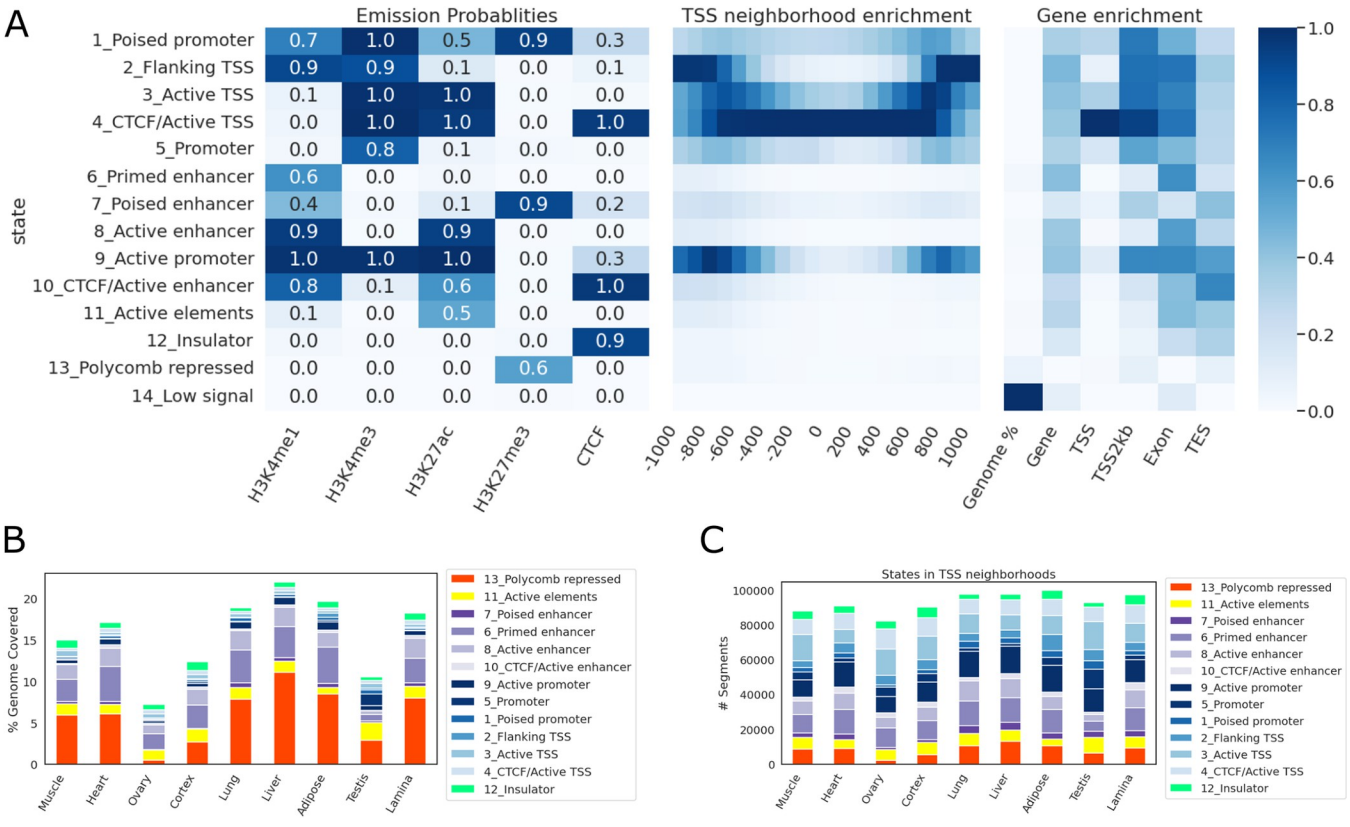

Fig 6. Chromatin states. (A) Emission probabilities, TSS neighborhood enrichment, and different genic features' enrichment at each state. (B) The percentage of genome covered by each state in each tissue. (C) The number of segments from each state in each tissue.

https://doi.org/10.1371/journal.pgen.1010468.g006

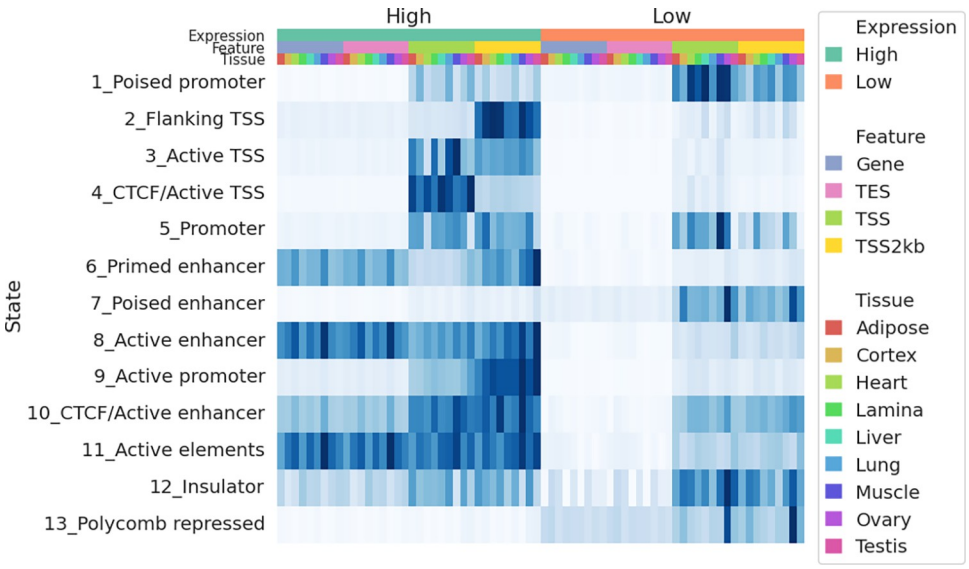

Fig 7. State enrichment in tissue specific genes. Heatmap of enrichment for each state around genes expressed and not expressed in each tissue. The top three color bars denote gene expression status, genic features, and tissues, respectively. Enrichment scores were normalized in each column.

https://doi.org/10.1371/journal.pgen.1010468.g007

quasi-mapping [43]. Transcript level quantification was then summarized to gene level using tximport [44]. For each tissue, genes were classified as high- or low-expression based on their aggregated transcripts per million (TPM) values (high: TPM $\geq$ 1; low: TPM $<$ 1). The enrichment of each state was then estimated across gene bodies, in exonic regions, around TSS and transcription end sites (TES) across all nine tissues (Fig 7). CTCF bound active TSS state (state *4*) showed a 59.4-fold enrichment around TSS of highly expressed (TPM $\geq$ 1) genes, 7.7 times that of lowly expressed genes (TPM $<$ 1). Similarly, active promoter state (state *9*) showed a 14.7-fold enrichment in promoter-TSS neighborhood (TSS2kb), 6.4 times that of lowly expressed genes. On the other hand, poised promoter and enhancer states (states *1* and *7*) were more enriched around TSS of lowly expressed genes (18.7- and 7.5-fold enrichment, respectively). Polycomb repressed states (state *13*) were absent around genes with high expression but enriched in low-expression genes while promoter state marked by a single H3K4me3 mark (state *5*) was observed at a similar level in both categories. Since this promoter state also showed the highest tissue-specificity (S3 Fig), we further examined its distribution among tissues. Most remarkably, testis harbored a substantially greater number of segments of State *5* than any other tissues (46,406 in testis compared to 5,235 in ovary, which was the next highest tissue) (S4 Fig). 61% of promoter state (state *5*) was found in testis and of those, 86% were specific to testis. Similarly, testis also contained the highest numbers of CTCF-less active TSS and poised promoter states (S5 Fig). While less pronounced, it also accounted for 44% of CTCF-less active TSS state (state *3*) and 54% of poised promoter state (state *1*).

To infer potential functions of open chromatin regions, especially those located in the intergenic regions, open chromatin peaks were annotated based on overlap chromatin state segments in each tissue. First, we examined overlap between each chromatin region and different open chromatin states across tissues (Fig 8A). There was an overall agreement in promoter-TSS assignment between open chromatin regions and chromatin state annotations: 92.9% of open chromatin peaks located in TSS-promoter regions overlap a TSS or promoter state (states *1–5*, *9*). Additionally, open chromatin regions located in exonic and intergenic regions showed higher percentages of enhancer states (28.9% and 17.9%, respectively) (Fig 8A). Next, we

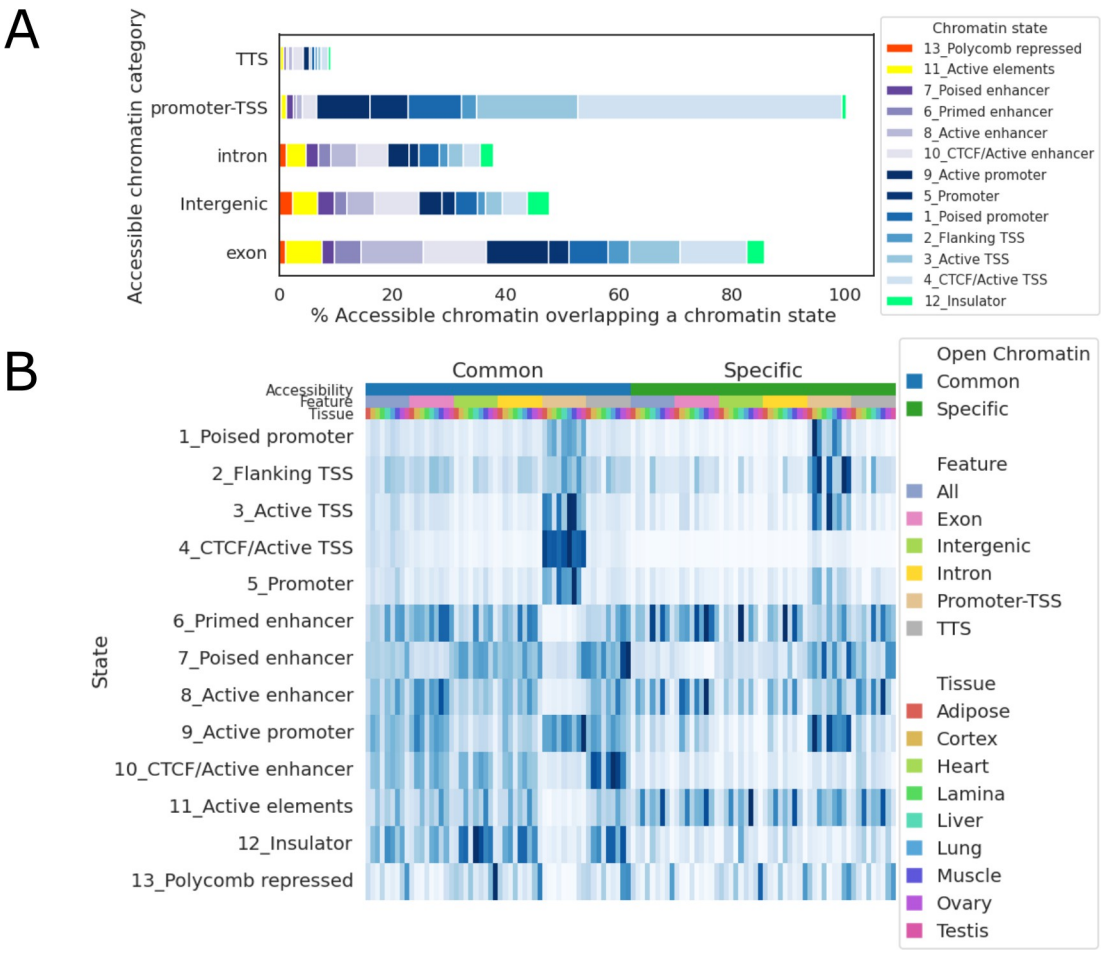

Fig 8. Chromatin accessibility across states. (A) Percentage of open chromatin peaks that overlap each chromatin state. (B) Heatmap of enrichment for each state around open chromatin peaks. The top three color bars denote shared or tissue-specific open chromatin, open chromatin annotation, and tissues, respectively. Enrichment scores were normalized in each column.

https://doi.org/10.1371/journal.pgen.1010468.g008

compared chromatin state enrichment among shared and tissue-specific open chromatin regions (Fig 8B). An open chromatin region was annotated as specific if it was found in only one tissue. CTCF bound active TSS state (state *4*) was highly enriched in common accessible chromatin regions, especially those annotated as promoter-TSS regions (121-fold enrichment), but much less so in tissue-specific accessible chromatin regions (12-fold enrichment). Similarly, CTCF bound enhancer state (state *10*) was also highly enriched in common accessible chromatin regions outside of promoter-TSS neighborhoods (15.3- to 31.5-fold enrichment), and less so in tissue-specific accessible chromatin regions (3.9- to 7.6-fold enrichment).

Since many enhancers interact with genes other than their nearest neighbors [45], in order to predict target genes CREs in absence of chromatin interaction data, we predicted chromatin loops and correlated H3K27ac signal (data originally generated and reported by Kingsley et al. [42]) of CREs with expression of genes resided within the same chromatin loops. First, chromatin loops were predicted using CTCF ChIP-seq data, as described by Oti *et al.* [46]. Overall, we identified 10-14k CTCF-mediated loops per tissue, with testis being the only exception, having only 6,146 CTCF-mediated loops. In all tissues, including testis, these predicted loops covered 80–85% of the genome. Since we only had at most 4 biological replicates per tissue (two for ovary and testis samples), which was not enough samples to reliably estimate

Spearman correlation coefficients [47], we opted for a pan-tissue approach. Tissue-wise chromatin loops were merged across tissues to form a catalog of pan-tissue CTCF-mediated chromatin loops, enabling estimation of correlation across 9 tissues and 4 biological replicates. This catalog contained 4,556 non-overlapping loops, covering 94.0% of the equine genome.

As demonstrated by Kern *et al.* [48], H3K27ac intensity of a CRE is most tightly correlated with its target gene's expression level. Therefore, we correlated H3K27ac ChIP-seq read counts and RNA-seq read counts of each CRE-gene pair that resided within a same predicted chromatin loop. After adjusting for multiple testing using Benjamini-Hochberg procedure to control the false discovery rate at 5%, a total of 84,613 CRE-gene pairs remained as candidates. These CREs were then annotated as genic, intergenic, or TSS-proximal based on their relative proximities. A majority of these candidate pairs had their CREs outside of the gene bodies or TSS-proximal regions (intergenic CREs, 66,051) while only a small portion of them had promoter-like relationship (CREs in the TSS-proximal regions, 8,225). Intergenic CREs were found at varying distances to TSS, with a median distance of 200 Kb and 79% of CREs being within 1 Mb their target TSS. We also observed more CREs (75%) located downstream of their target genes (**S6 Fig**).

To provide the equine community with an integrated, openly access FAANG dataset, we developed a UCSC track hub (https://genome.ucsc.edu/s/cjfinno/equCab3) to host all currently published equine FAANG datasets. All features discussed above can be found in this track hub.

Discussion

In this study, we detailed the efforts to create an integrated annotation for the horse genome that will aid in deeper understanding of gene expression and regulation across tissues in the horse. Utilizing the rich tissue repository from the equine FAANG project, we improved the equine transcriptome with 39,625 novel transcripts and identified 84,613 candidate CRE-gene pairs, 78.1% of which were intergenic CREs. We anticipate these new resources will play a vital role to understanding how genetic variation in the horse contributes to equine biology and health.

First, we outlined an expansive transcriptome annotation for the horse. The previous efforts to annotate the horse genome were limited by the number of tissue types available and sequencing lengths available at that time [49–51]. Specifically, Hestand et al. sequenced 43 different equine tissues in one pool on an Illumina HiSeq 2000, both single- and paired-end at 75 bp on 4 lanes each [50]. While that study included more diverse tissue types than the current FAANG transcriptome, the pooling approach employed in that study limited discovery of rare novel and tissue-specific isoforms. The non-stranded protocol employed in that study also rendered it impossible to identify antisense transcripts. Mansour et al. compared sequences of 8 tissue samples from 59 individuals using short-read RNA-seq libraries from several studies (80–125 bp, single- and paired-end, stranded and unstranded) and identified 36,876 genes with 76,125 isoforms [51]. Due to the limitation of short-read sequencing technology in both the Hestand and Mansour studies, an aggressive filter was necessary to remove mono-exonic transcripts that were not evolutionarily conserved, a common strategy in short-read based transcriptome assemblies [4,5]. This unfortunately would also remove many small noncoding RNAs. Based on recent advances in long isoform sequencing, our approach centered around high-quality full-length reads from Iso-seq and used abundant RNA-seq data to validate splice junction, TSS, and TTS annotation. As a result, novel alternate-spliced isoforms as well as extended 5' and 3' transcribed regions were identified using long-read Iso-seq, validated by abundant short-read mRNA-seq. This combined approach expanded the equine transcriptome

to 153,492 transcripts (of which 128,723 are multi-exonic) from 36,239 genes, with a gene-isoform ratio of 4.2 and an average 19.5% (8–45%) improvement in completeness compared to Ensembl and RefSeq transcriptomes across all sequenced FAANG tissues. The newly discovered genes and isoforms could help identify important coding or regulatory variants in the horse.

Despite these improvements, the present Iso-seq transcriptome was unable to accurately define TSS due to a lack of 5' captured reads. While an aggressive approach was taken to ensure 5' completeness by collapsing transcripts that only differ at 5' ends, a small portion of transcripts were still determined to be potentially 5' incomplete. In addition, this approach may hinder the discovery of alternative TSS. Furthermore, small RNAs with non-polyadenylated tails are missing from the poly-A captured cDNA libraries used for both Iso-seq and RNA-seq. Assays targeting non-polyadenylated RNAs, such as small RNA sequencing and techniques capturing 5' capped transcripts like CAGE-seq are necessary to complement this Iso-seq transcriptome to fully capture the transcriptional landscape in the horse genome. Further, while we demonstrate improved completeness of the FAANG equine transcriptome, only 9 tissues were utilized to construct it, and many rare or tissue-specific transcripts are likely to be missing, especially stem-cell-specific or embryonically specific transcripts. Indeed, short read sequencing data from bone marrow was the only tissue that showed a drastic decrease in mapping rate when compared to RefSeq or Ensembl transcriptomes, suggesting specific isoforms from this tissue are missing in the new transcriptome. In addition, since mare and stallion tissues were prepared at two different laboratories, despite using same protocols, we could not distinguish any sex-specific expression from batch effects during RNA-seq library construction. Lastly, since RNA-seq data was only used to validate and quantify transcripts identified in Iso-seq data, our approach heavily relied on Iso-seq data being complete. This unfortunately was not true as it was evident in **Figs 2** and **S1** that a substantial amount of reads from RNA-seq failed to map to Iso-seq transcriptome. A closer examination of RNA-seq data in conjunction with the Iso-seq data could further refine the FAANG transcriptome.

With an improved transcriptome annotation, we set out to identify other non-transcribed or lowly transcribed regulatory regions in the horse genome. The first step was to identify regions of the horse genome that were accessible to transcription factors, which can then serve as proxies to identifying important regulatory regions. Using ATAC-seq, we identified 332,115 regions with open chromatin genome wide across tissues, with 59,024–95,048 peaks identified in each tissue. We showed that these open regions were enriched with known TF binding sites, further supporting their potential functional roles in gene regulation.

We observed that, while promoter-TSS regions were highly enriched in open chromatin peaks, especially in peaks found in all tissues, they were conspicuously absent from tissue-specific peaks. This echoed the findings from Halstead *et al.* [52], which showed very low numbers of TSS-related peaks in species-specific open chromatin regions. This would corroborate recent findings that enhancers, not promoters, are the main drivers of tissue-specific transcription [53], which would be classified as intergenic in both our study as well as in Halstead *et al.*, as all species interrogated in these studies lacked enhancer annotation.

Combining our ATAC-seq data with previously reported mRNA-seq data from the same tissue and samples, we showed a strong correlation between differential accessibility of a functional element and differential expression of the corresponding gene. When FDR for both DAR and DEG was controlled at 5%, we observed concordant patterns between the gene expression level and accessibility of promoter-TSS region. Interestingly, a small number of genes (352 out of 4,566, 7.8%) also showed discordant patterns between gene expression and promoter-TSS accessibility. It should be noted that, while we took a similar analytical approach for DEG and DAR, ATAC-seq and RNA-seq signals reflect fundamentally different regulatory

features. RNA-seq captures total transcriptional activities in a population of cells, where a small population of cells with extremely high transcription of a given gene can dominate RNA-seq signals for the entire population. On the other hand, ATAC-seq captures the proportion of cells whose DNA is accessible at a given locus. Thus, if a small population of cells have substantially high expression of a given gene, while the remaining cells do not express this gene nor is it accessible, the gene would be upregulated in the RNA-seq dataset but not identified as open in an ATAC-seq dataset, which could explain the discordant patterns we observed in Q2 and Q4 genes. Additionally, the presence of silencers in CREs or the time difference from CRE activation to change in mRNA abundance may also explain the observed discordance. To further dissect the fine regulatory landscape of these genes, single-cell based approaches are advisable for both RNA quantification and chromatin accessibility assessment to best match representing cell types across assays.

In all nine tissues, 21–34% of peaks were located in the intergenic regions while 23–45% of tissue-specific peaks were in intergenic regions. Motif analyses in this study identified common TF binding sites in many of these intergenic open chromatin regions. For example, over 41% (4,112) of the liver-specific intergenic open chromatin regions contained binding sites for ERRα, a TF known as a central regulator of energy metabolism [54]. Similarly, binding sites for various SOX family TFs were detected in 15–32% of cerebral cortex-specific intergenic open chromatin regions. The SOX family TFs have been shown to be important regulators for neural differentiation and adult neurogenesis [55].

It is likely that many of intergenic open chromatin regions have important regulatory functions. To elucidate their functional roles, we identified potential regulatory states genome wide across tissues using signals from biochemical assays (histone modifications and CTCF ChIP-seq) and correlated them with open chromatin peaks. We identified 14 unique chromatin states using data from four major histone marks and CTCF binding assays. These chromatin states were identified in each of the nine tissue types, covering 7–21% of the genome, representing major CREs. Overall, the chromatin state annotation correlated well with chromatin accessibility in the same tissues and provided additional information regarding potential function of REs in these tissues. These annotated REs will be an invaluable addition to the equine reference genome assembly. The similar annotation provided by ENCODE has led to discoveries of many regulatory variants in various diseases [20,22,56]. We anticipate this catalog of REs will prove instrumental in evaluating complex genetic traits and disease in the horse.

In developing these unique chromatin states, we noted a particular difference in chromatin state annotation for shared and tissue-specific open chromatin peaks. CTCF bound promoter and enhancers were highly enriched in shared open chromatin regions but less so in tissue-specific regions. CTCF is a chromatin regulator that facilitates formation of chromatin loops. It has been suggested that a subset of CTCF binding sites is constitutively bound and critical to well-regulated gene expression [57] and that CTCF binding at proximal promoters promotes distal enhancer-promoter interaction, which is essential to the activation of many genes across a diverse range of tissues [58]. Our results suggest that these CTCF-mediated promoter-enhancer interactions play a large role in genes expressed across multiple tissues, rather than tissue-specific genes. This aligns with other findings which suggest that CTCF patterns are established early in embryogenesis [59]. Taking advantage of the extensive research surrounding the relationship between CTCF binding and 3-dimensional chromatin structures (TADs), we used our CTCF ChIP-seq data to predict chromatin loops and were therefore able to predict 84,613 candidate CRE-gene interactions across tissues. This dataset should dramatically improve our ability to both identify important regulatory variants and predict their target genes and gene networks.

Work from this study has opened doors for further exploration. For example, the discordant relationships between differentially accessible regions (DARs) and differentially expressed

genes (DEGs) in brain and heart tissues suggest substantial sub-population differences within each of these tissues. Future studies utilizing single-cell-based technologies could help unravel such differences and identify cell-type-defining genes and CREs. Additionally, we observed substantial differences between testes and all other tissues from both the ATAC-seq and ChIP-seq data. This difference could be a result of significant spermatozoa population in our testis samples, or it could be related to the unique transcriptional landscape of testis. Future research should focus on separating mature spermatozoa with spermatogonium and other cell types in testis to further refine the regulatory landscape of this tissue. Overall, we presented an integrated repository of equine FAANG data, encompassing both transcriptional and regulatory features that are now freely available to the equine community. We anticipate this resource to be integral to future equine research.

Methods

RNA extraction and sequencing

From the outset of the equine FAANG initiative, researchers were invited to "adopt" tissues of interest. This involved sponsorship of the sequencing costs for two biological replicates (2 male or 2 female) of the "adopted" tissue. Under this Adopt-A-Tissue model, along with the eight prioritized tissues funded by both the USDA National Institute of Food and Agriculture and the Grayson Jockey Club Foundation, the equine community collectively generated short-read mRNA-seq data from over forty tissues. All RNA extractions for mRNA-seq were performed at two locations (female samples at UC Davis, male samples at University of Nebraska-Lincoln). Briefly, tissue aliquots were homogenized using Biopulverisor and Genogrinder in TRIzol reagent (ThermoFisher Scientific, Waltham MA). RNA was isolated and purified using RNeasy Plus Mini/Micro columns (Qiagen, Germantown, MD) or Direct-zol RNA Miniprep Plus (Zymo Research, Irvine, CA). A detailed protocol can be found in **S1 and S2 Texts**. For the female tissues, cDNA libraries were prepared with Illumina TruSeq Stranded kit and sequenced at the University of Minnesota sequencing core facility on an Illumina HiSeq 2500 using 125 bp paired-end reads. Male samples went through similar library preparation before 150 bp paired-end sequencing at Admera Health (South Plainfield, NJ) on an Illumina NovaSeq.

Nine tissues (lamina, liver, left lung, left ventricle of heart, longissimus muscle, skin, parietal cortex, testis, and ovary) from the FAANG biobank [38,39] were selected for Iso-seq to represent a wide range of biological functions and therefore, gene expression. RNA for Iso-seq was extracted separately from the same tissues as mRNA-seq using the same protocol. All tissues were processed in one batch for Iso-seq, except for parietal cortex, which was processed in a separate batch as a pilot study. One sample per sex per tissue was selected for sequencing based on sample availability and RNA integrity numbers (RINs selected > 7). cDNA libraries were prepared and sequenced at UC Berkely QB3 Genomics core facility. Two libraries were randomly pooled and sequenced on a single SMRT cell on PacBio Sequel II.

Transcriptome assembly

Pooled subreads were first demultiplexed using Lima (https://lima.how/). Circular consensus reads (ccs) were then constructed from demultiplexed subreads using PacBio Ccs program (https://ccs.how/). PolyA tails were trimmed from ccs reads using Isoseq3 (https://github.com/PacificBiosciences/IsoSeq). This step also removes concatemers and any reads lacking at least 20 bp of polyA tails. Redundant reads were then clustered based on pair-wise alignment using Isoseq3. Clustered transcripts were aligned to the reference genome EquCab3 [2] using minimap2 [60] without reference annotation as guide. Collapsed transcripts were filtered if they

were not supported by at least two full length reads. Filtered transcripts from each sample were then merged into a single transcriptome using Cupcake (https://github.com/Magdoll/cDNA_Cupcake/) and further filtered to retain only those detected in more than one sample. The merged total transcriptome was again aligned to the reference genome and collapsed to remove redundant transcripts. Potential 5' degraded transcripts were also removed by collapsing transcripts that had identical 3' ends and only differed at 5' ends. SQANTI3 [61] was then used to classify and annotate the transcriptome against the RefSeq transcriptome as reference. Finally, the total transcriptome was filtered to remove nonsense-mediated decay transcripts, transcripts with a splice junction not covered by short-read RNA-seq data, and transcripts without short-read coverage support to generate the final FAANG equine transcriptome (5,546, 7,262, and 12,900 transcripts were removed by each filter, respectively). To detect potential intra-primed transcripts, the percentage of adenines in a 20 bp window immediately downstream of the annotated TTS was calculated for every Iso-seq transcript. Transcripts with 80% or more adenines (i.e., allowing for 4 mismatches with poly-T oligonucleotides) in a 20 bp window downstream of annotated TTS were designated as potential intra-priming candidates. Data processing, visualization, and statistical analyses were performed using pandas [62], matplotlib [63], seaborn [64], scipy [65], and scikit-learn [66]. Detailed program versions, commands, parameters, and code can be obtained from https://github.com/FinnoLab/FAANG_IsoSeq.

RNA-seq analysis

Short-read RNA-seq data were trimmed to remove adapters and low-quality reads using trim-galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and Cutadapt [67]. Read qualities were inspected using fastQC [68] and multiQC [69]. Trimmed reads were aligned to equCab3.0 using STAR aligner [70] with standard parameters (with—outSAMstrandField intronMotif—outSAMattrIHstart 0). PCR duplicates were marked using sambamba [71]. Mapping rates, qualities, and fragment lengths were assessed with SAMTools [72] and deepTools [73]. Aligned reads were used to assess completeness of transcriptomes using deepTools. BWA MEM [74] was used to align the RNA-seq reads directly to transcriptomes and SAMTools was used to calculate the percentages of properly paired reads from the transcriptome alignment. Due to the presence of alternatively spliced isoforms in transcriptomes, multiple-alignment reads were not removed. Salmon [43] was used to quantify the transcripts using RNA-seq data. Transcript level TPM values were summarized to gene level using tximport [44]. For the chromatin state enrichment analyses, genes were designated "active" if its aggregated TPM was at least 1 in a tissue. TSS, promoter-TSS neighborhood (TSS±2kb), exon, intron, and TTS coordinates were determined for each gene based on the FAANG transcriptome. Detailed program versions, commands, parameters, and code can be obtained from https://github.com/FinnoLab/FAANG_IsoSeq.

ATAC-seq analysis

ATAC-seq data from the 9 tissues (adipose, lamina, liver, left lung, left ventricle of heart, longissimus muscle, parietal cortex, testis, and ovary) of two sexes collected from the equine FAANG biobank [38,39] were generated according to Peng *et al.* [27] Libraries were sequenced in 50 bp paired-end mode (PE50) on Illumina NovaSeq 6000. Reads were aligned to EquCab3.0 using BWA MEM with default parameters. Alignments were filtered to remove fragments that mapped to mitochondria genome, were discordantly mapped, PCR duplicates, or mapped to multiple loci using SAMTools. Remaining reads were shifted +4/-5 bp on plus/minus strand, respectively, to account for the 9 bp insertion introduced by Tn5 transposase

[24] using deepTools. Both forward and reverse reads of the final fragments were converted to bed format using bedtools [75] and peaks were called and refined using MACS3 [40,41] (-f BED -p 0.01—shift -75—extsize 150—nomodel—call-summits—nolambda—keep-dup all). After peak calling, we extracted summits from called peaks, and extended them on both sides by 250 bp, resulting in a set of 501 bp fixed length peaks. These peaks were then sorted by their score and non-overlapping, most significant peaks were retained, as described in Grandi *et al.* [76]. The same procedure was employed to subsequently merge biological replicates and then all tissue peak sets to generate a union set of peaks. A count matrix was constructed for the union peak set containing number of transposition events per peak per sample. This count matrix was used for differential accessibility analyses using DESeq2 [77]. The union peak set was then intersected with each tissue peak set to determine if a peak was present in each tissue. Peaks only identified in one tissue type were denoted "unique" peaks while those identified in all nine tissues were denoted as "conserved". Detailed program versions, commands, parameters, and code can be obtained from https://github.com/FinnoLab/atac-seq.

Differential accessibility and expression analyses

DAR and DEG were analyzed with similar approaches. First, a matrix of raw counts was constructed containing number of transposition events (DAR) or RNA-seq reads (DEG) for each union open chromatin peaks (DAR) or gene (DEG). The raw counts were then normalized and fitted to a negative binomial generalized linear model using DESeq2 (1.30.1) [77] package with default parameters. Wald tests were applied to obtain p-values for each region (DAR) or gene (DEG) and Benjamini-Hochberg procedure were used to control false discovery rate at 5%.

ROC analyses

For each set of peaks merged by tissues, false positive rates (FPR), true positive rates (TPR), and precision were calculated using published Histone ChIP-seq peaks from Kingsley *et al.* [42]:

First, a set of "real positive" (RP) peaks were collected by merging H3K4me1 and H3K4me3 peaks and intersecting the merged peaks with H3K27ac peaks from each tissue. A set of "real negative" (RN) peaks were collected from H3K27me3 peaks from each tissue. Subsequently, each set of ATAC-seq peaks were intersected with RP and RN peaks, and the number of intersections were recorded as "true positive" (TP) and "false positive" (FP). TPR, FPR, and precision were then calculated as follows:

$$TPR = \frac{n_{TP}}{n_{RP}}$$

$$FPR = \frac{n_{FP}}{n_{RN}}$$

$$Precision = \frac{n_{TP}}{n_{TP} + n_{FP}}$$

Motif and gene ontology analyses

Motifs were analyzed using HOMER [78] (-size 250) with custom genome built from EquCab3.0 assembly and FAANG transcriptome annotation. GO enrichment analyses were performed using PANTHER [79] with default parameters.

Chromatin state discovery

ChIP-seq data for histone modifications were obtained from previously published studies [42,80]. Additionally, ChIP-seq for CTCF was performed for the same nine frozen tissue samples at Diagenode *Inc*. (Belgium). Briefly, CTCF ChIP libraries were sequenced at 50bp single- and paired-end (female and male samples, respectively). Reads were aligned to EquCab3.0 using BWA MEM with default parameters. Aligned reads were subsequently filtered to remove low-quality mapping, PCR duplicates, and mitochondria reads using SAMTools. BAM files for all five marks were binarized using ChromHMM [81] BinarizeBam (-b 100 -n 140 -p 0.00001) and several models with different numbers of states were trained on binarized data using LearnModel function (-b 100). A model with 14 states was selected because it had the minimum number of states with strong correlation to all states identified in other models.

Chromatin loop prediction and CRE-gene interaction analyses

To obtain a pan-tissue set of chromatin loops in absence of chromatin interaction data, CTCF ChIP-seq data was used to predict chromatin loops in each tissue, using the algorithm proposed by Oti *et al*.[46]. Briefly, CTCF ChIP-seq peaks were identified using MACS3 [40] (-q 0.05 -f BAM -g 2365156725—keep-dup auto). CTCF motifs were then extracted from these peaks using FIMO function from MEME [82] (—motif MA1930.1—parse-genomic-coord). Then in each tissue, each chromosome was scanned continuously, and a loop was recorded by detecting a pair of parallel and anti-parallel CTCF motifs. The predicted CTCF loops across tissues were then combined by merging overlapping loops using bedtools [75] merge function.

To obtain a list of CREs, neighboring active chromatin states (states *2–5* and *8–11)* were merged and annotated by their proximity to a known gene (genic, intergenic, and TSS-promoter) as well as their closest genes. These CREs were then quantified by number of overlapping reads from their corresponding H3K27ac ChIP-seq data. Reads counts were normalized by a scaling factor calculated using weighted trimmed mean (TMM) method. Normalized RNA-seq read counts were obtained from previous analysis (see Methods - Differential Accessibility and Expression Analyses). Spearman correlation between H3K27ac read count in each CRE and RNA-seq read count in each gene was calculated using spearmanr function from SciPy [65]. P-values were adjusted using Benjamini-Hochberg procedure and candidate CRE-gene pairs were filtered at 5% false discovery rate.

Enrichment analysis

Enrichment of each state in genes and open chromatin regions was calculated using the following formula:

$$\frac{\frac{N_{Ann \bigcap State}}{N_{Ann}}}{\frac{N_{state}}{N_{genome}}}$$

where N_{Ann} is the number of bases in a particular annotation (gene, exon, TSS, open chromatin peaks, etc) and N_{state} is the number bases in each state. $N_{Ann \bigcap State}$ refers to the number of bases that are in both a particular state and annotation. N_{genome} is the total size of the reference genome.

Supporting information

S1 Fig. Comparison of FAANG, RefSeq and Ensembl equine transcriptomes. Changes in percentages of properly paired reads aligned to combined Iso-seq transcriptome when

compared to Ensembl or RefSeq transcriptomes, whichever has higher percentage.
(TIF)

S2 Fig. ROC plots of ATAC-seq peaks. Receiver Operating Characteristics (ROC) of eight tissues whose ATAC-seq peaks were validated by Histone ChIP-seq data
(TIF)

S3 Fig. Tissue-specificity of states. The proportion of segments from each state that were identified in different numbers of tissues.
(TIF)

S4 Fig. Promoter state shared across tissues. Intersection plot showing number of segments annotated as promoter state (state *5*) unique to each tissue and shared across tissues. Top: bar plot indicates sizes of each intersection; Bottom right: each column denotes a unique set of peaks where filled dots indicat that peaks in this set were found in the corresponding tissue; Bottom left: bar plot indicates number of segments annotated as promoter state (state *5*) in each tissue.
(TIF)

S5 Fig. CTCF-less active TSS and poised promoter states shared across tissues. Intersection plots showing number of segments annotated as (A) CTCF-less active TSS state (state *3*) and (B) poised promoter state (state *1*) unique to each tissue and shared across tissues. Top: bar plot indicates sizes of each intersection; Bottom right: each column denotes a unique set of peaks where filled dots indicate that peaks in this intersection were found in the corresponding tissue; Bottom left: bar plot indicates number of segments annotated as (A) CTCF-less active TSS state (state *3*) or (B) poised promoter state (state *1*) in each tissue.
(TIF)

S6 Fig. Distance from intergenic RE to target genes' TSS. Density plot of distances from intergenic REs to their target genes' TSS. Negative distance denotes RE being upstream of target TSS. Median absolute distance: 200 Kb.
(TIF)

S1 Table. Gene ontology of peaks identified across all 9 tissues. GO enrichment analysis of peaks conserved across all nine tissues.
(XLSX)

S2 Table. De novo motif discoveries in tissue specific intergenic regions. Top known motifs in each tissue, filtered by FDR $q \leq 0.05$.
(XLSX)

S3 Table. Gene ontology of Q1 genes. GO enrichment analysis of Q1 genes upregulated in both DEG and DAR analyses of cerebral cortex vs. heart.
(XLSX)

S4 Table. Gene ontology of Q3 genes. GO enrichment analysis of Q1 genes downregulated in both DEG and DAR analyses of cerebral cortex vs. heart.
(XLSX)

S5 Table. Gene ontology of Q2 genes. GO enrichment analysis of Q1 genes upregulated in DEG but downregulated in DAR analyses of cerebral cortex vs. heart.
(XLSX)

S1 Text. Detailed protocol for RNA Isolation using a column and on-column DNA digestion.
(DOCX)

S2 Text. Detailed protocol for Finno modifications of the RNeasy Lipid Tissue Kit (Qiagen) for sesamoid bone.
(DOCX)

S1 Data. Gtf file for the merged FAANG-refseq-ensembl annotated transcripts in EquCab3.0. This is a tab-delimited text formatted file can be uploaded to the Integrated Genome Viewer (IGV; https://software.broadinstitute.org/software/igv/ or UCSC https://genome.ucsc.edu).
(GZ)

Author Contributions

Conceptualization: Sichong Peng, Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Data curation: Sichong Peng, Anna R. Dahlgren, Callum G. Donnelly, Erin N. Hales, Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Formal analysis: Sichong Peng, Erin N. Hales, Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Funding acquisition: Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Investigation: Sichong Peng, Erin N. Hales, Jessica L. Petersen, Rebecca R. Bellone, Carrie J. Finno.

Methodology: Sichong Peng, Anna R. Dahlgren, Callum G. Donnelly, Erin N. Hales, Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Project administration: Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Resources: Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Software: Ted Kalbfleisch.

Supervision: Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Validation: Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

Visualization: Sichong Peng, Jessica L. Petersen, Rebecca R. Bellone, Carrie J. Finno.

Writing – original draft: Sichong Peng.

Writing – review & editing: Sichong Peng, Anna R. Dahlgren, Callum G. Donnelly, Erin N. Hales, Jessica L. Petersen, Rebecca R. Bellone, Ted Kalbfleisch, Carrie J. Finno.

References

1. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, et al. Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. Science. 2009 Nov 6; 326(5954):865–7. https://doi.org/10.1126/science.1178158 PMID: 19892987
2. Kalbfleisch TS, Rice ES, DePriest MS, Walenz BP, Hestand MS, Vermeesch JR, et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol. 2018; 1:197. https://doi.org/10.1038/s42003-018-0199-z PMID: 30456315

3. Raudsepp T, Finno CJ, Bellone RR, Petersen JL. Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post-genome era. Anim Genet. 2019 Dec; 50(6):569–97. https://doi.org/10.1111/age.12857 PMID: 31568563

4. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4; 44(D1):D733–745. https://doi.org/10.1093/nar/gkv1189 PMID: 26553804

5. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Res. 2021 Jan 8; 49(D1):D884–91. https://doi.org/10.1093/nar/gkaa942 PMID: 33137190

6. Equus caballus RefSeq Annotation Release 103 [Internet]. RefSeq. [cited 2021 Sep 10]. Available from: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Equus_caballus/103/

7. Ensembl Genebuild 106.3, EquCab3.0 [Internet]. 2019. Available from: https://uswest.ensembl.org/Equus_caballus/Info/Annotation

8. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 2019 Jan 8; 47(D1):D766–73. https://doi.org/10.1093/nar/gky955 PMID: 30357393

9. Roundtree IA, He C. RNA epigenetics—chemical messages for posttranscriptional gene regulation. Current Opinion in Chemical Biology. 2016 Feb; 30:46–51. https://doi.org/10.1016/j.cbpa.2015.10.024 PMID: 26625014

10. Lee TI, Young RA. Transcriptional Regulation and Its Misregulation in Disease. Cell. 2013 Mar; 152 (6):1237–51. https://doi.org/10.1016/j.cell.2013.02.014 PMID: 23498934

11. Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frébourg T, et al. Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. Aretz S, editor. PLoS Genet. 2016 Jan 13; 12(1):e1005756. https://doi.org/10.1371/journal.pgen.1005756 PMID: 26761715

12. De Paoli-Iseppi R, Gleeson J, Clark MB. Isoform Age—Splice Isoform Profiling Using Long-Read Technologies. Front Mol Biosci. 2021 Aug 2; 8:711733. https://doi.org/10.3389/fmolb.2021.711733 PMID: 34409069

13. Chen SY, Deng F, Jia X, Li C, Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Sci Rep. 2017 Dec; 7(1):7648. https://doi.org/10.1038/s41598-017-08138-z PMID: 28794490

14. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013 Nov; 31(11):1009–14. https://doi.org/10.1038/nbt.2705 PMID: 24108091

15. Suryamohan K, Krishnankutty SP, Guillory J, Jevit M, Schröder MS, Wu M, et al. The Indian cobra reference genome and transcriptome enables comprehensive identification of venom toxins. Nat Genet. 2020 Jan; 52(1):106–17. https://doi.org/10.1038/s41588-019-0559-8 PMID: 31907489

16. Hansen AS, Iryna P, Claudia C, Tjian R, Xavier D. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. eLife; Cambridge [Internet]. 2017 [cited 2019 Jun 11]; 6. Available from: https://search.proquest.com/docview/1952732110/abstract/B705B22ED1E14523PQ/1 https://doi.org/10.7554/eLife.25776 PMID: 28467304

17. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature. 2017 Apr; 544(7648):59–64. https://doi.org/10.1038/nature21429 PMID: 28289288

18. Sos BC, Fung HL, Gao DR, Osothprarop TF, Kia A, He MM, et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol. 2016 Feb 4; 17:20. https://doi.org/10.1186/s13059-016-0882-7 PMID: 26846207

19. Liu C, Wang M, Wei X, Wu L, Xu J, Dai X, et al. An ATAC-seq atlas of chromatin accessibility in mouse tissues. Sci Data. 2019 Dec; 6(1):65. https://doi.org/10.1038/s41597-019-0071-0 PMID: 31110271

20. Warburton A, Breen G, Rujescu D, Bubb VJ, Quinn JP. Characterization of a REST-Regulated Internal Promoter in the Schizophrenia Genome-Wide Associated Gene MIR137. Schizophr Bull. 2015 May; 41 (3):698–707. https://doi.org/10.1093/schbul/sbu117 PMID: 25154622

21. Giorgio E, Robyr D, Spielmann M, Ferrero E, Di Gregorio E, Imperiale D, et al. A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). Hum Mol Genet. 2015 Jun 1; 24(11):3143–54. https://doi.org/10.1093/hmg/ddv065 PMID: 25701871

22. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. Nature. 2010 Apr; 464(7291):1071–6.

23. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet. 2009 Mar; 10(3):161–72. https://doi.org/10.1038/nrg2522 PMID: 19204718

24. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology [Internet]. 2015 Jan [cited 2020 Oct 28]; 109(1). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb2129s109 PMID: 25559105

25. Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods. 2017 Oct; 14(10):959–62. https://doi.org/10.1038/nmeth.4396 PMID: 28846090

26. Halstead MM, Kern C, Saelao P, Chanthavixay G, Wang Y, Delany ME, et al. Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. Sci Rep. 2020 Dec; 10(1):5230. https://doi.org/10.1038/s41598-020-61678-9 PMID: 32251359

27. Peng S, Bellone R, Petersen JL, Kalbfleisch TS, Finno CJ. Successful ATAC-Seq From Snap-Frozen Equine Tissues. Front Genet. 2021 Jun 16; 12:641788. https://doi.org/10.3389/fgene.2021.641788 PMID: 34220931

28. Zentner GE, Henikoff S. Regulation of nucleosome dynamics by histone modifications. Nature Structural & Molecular Biology. 2013 Mar; 20(3):259–66. https://doi.org/10.1038/nsmb.2470 PMID: 23463310

29. Zhang Y, Sun Z, Jia J, Du T, Zhang N, Tang Y, et al. Overview of Histone Modification. In: Fang D, Han J, editors. Histone Mutations and Cancer [Internet]. Singapore: Springer Singapore; 2021 [cited 2022 Jun 18]. p. 1–16. (Advances in Experimental Medicine and Biology; vol. 1283). Available from: http://link.springer.com/10.1007/978-981-15-8104-5_1

30. Hyun K, Jeon J, Park K, Kim J. Writing, erasing and reading histone lysine methylations. Exp Mol Med. 2017 Apr; 49(4):e324–e324. https://doi.org/10.1038/emm.2017.11 PMID: 28450737

31. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007 Mar; 39 (3):311–8. https://doi.org/10.1038/ng1966 PMID: 17277777

32. Santos-Rosa H, Schneider R, Bannister AJ, Sherriff J, Bernstein BE, Emre NCT, et al. Active genes are tri-methylated at K4 of histone H3. Nature. 2002 Sep; 419(6905):407–11. https://doi.org/10.1038/nature01080 PMID: 12353038

33. Lauberth SM, Nakayama T, Wu X, Ferris AL, Tang Z, Hughes SH, et al. H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. Cell. 2013 Feb; 152 (5):1021–36. https://doi.org/10.1016/j.cell.2013.01.052 PMID: 23452851

34. Bian C, Xu C, Ruan J, Lee KK, Burke TL, Tempel W, et al. Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation: Sgf29 functions as an H3K4me2/3 binder in SAGA. The EMBO Journal. 2011 Jul 20; 30(14):2829–42.

35. Eberl HC, Spruijt CG, Kelstrup CD, Vermeulen M, Mann M. A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics. Molecular Cell. 2013 Jan; 49(2):368–78. https://doi.org/10.1016/j.molcel.2012.10.026 PMID: 23201125

36. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010 Dec 14; 107(50):21931–6. https://doi.org/10.1073/pnas.1016071107 PMID: 21106759

37. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006 May; 441(7091):349–53. https://doi.org/10.1038/nature04733 PMID: 16625203

38. Burns EN, Bordbari MH, Mienaltowski MJ, Affolter VK, Barro MV, Gianino F, et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim Genet. 2018 Dec; 49(6):564–70. https://doi.org/10.1111/age.12717 PMID: 30311254

39. Donnelly CG, Bellone RR, Hales EN, Nguyen A, Katzman SA, Dujovne GA, et al. Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. Front Genet. 2021 Mar 8; 12:650305. https://doi.org/10.3389/fgene.2021.650305 PMID: 33763124

40. Liu T. MACS: Model-based Analysis for ChIP-Seq [Internet]. 2022. Available from: https://github.com/macs3-project/MACS

41. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9(9):R137. https://doi.org/10.1186/gb-2008-9-9-r137 PMID: 18798982

42. Kingsley NB, Kern C, Creppe C, Hales EN, Zhou H, Kalbfleisch TS, et al. Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. Genes. 2019 Dec 18; 11(1):3. https://doi.org/10.3390/genes11010003 PMID: 31861495

43. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017 Apr; 14(4):417–9. https://doi.org/10.1038/nmeth.4197 PMID: 28263959

44. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015 Dec 30; 4:1521. https://doi.org/10.12688/f1000research.7563.2 PMID: 26925227

45. Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature. 2013 Dec; 504(7479):306–10. https://doi.org/10.1038/nature12716 PMID: 24213634

46. Oti M, Falck J, Huynen MA, Zhou H. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. BMC Genomics. 2016 Dec; 17(1):252. https://doi.org/10.1186/s12864-016-2516-6 PMID: 27004515

47. Zwillinger D, Kokoska S. CRC standard probability and statistics tables and formulae. Boca Raton: Chapman & Hall/CRC; 2000. 554 p.

48. Kern C, Wang Y, Xu X, Pan Z, Halstead M, Chanthavixay G, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun. 2021 Dec; 12(1):1821. https://doi.org/10.1038/s41467-021-22100-8 PMID: 33758196

49. Coleman SJ, Zeng Z, Wang K, Luo S, Khrebtukova I, Mienaltowski MJ, et al. Structural annotation of equine protein-coding genes determined by mRNA sequencing: Structural annotation of equine protein-coding genes. Animal Genetics. 2010 Dec; 41:121–30.

50. Hestand MS, Kalbfleisch TS, Coleman SJ, Zeng Z, Liu J, Orlando L, et al. Annotation of the Protein Coding Regions of the Equine Genome. Ouzounis CA, editor. PLoS ONE. 2015 Jun 24; 10(6): e0124375. https://doi.org/10.1371/journal.pone.0124375 PMID: 26107351

51. Mansour TA, Scott EY, Finno CJ, Bellone RR, Mienaltowski MJ, Penedo MC, et al. Tissue resolved, gene structure refined equine transcriptome. BMC Genomics. 2017 Dec; 18(1):103. https://doi.org/10.1186/s12864-016-3451-2 PMID: 28107812

52. Halstead MM, Kern C, Saelao P, Wang Y, Chanthavixay G, Medrano JF, et al. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. BMC Genomics. 2020 Dec; 21(1):698. https://doi.org/10.1186/s12864-020-07078-9 PMID: 33028202

53. Ko JY, Oh S, Yoo KH. Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. Mol Cells. 2017 Mar; 40(3):169–77. https://doi.org/10.14348/molcells.2017.0033 PMID: 28359147

54. Xia H, Dufour CR, Giguère V. ERRα as a Bridge Between Transcription and Function: Role in Liver Metabolism and Disease. Front Endocrinol. 2019 Apr 5; 10:206.

55. Stevanovic M, Drakulic D, Lazic A, Ninkovic DS, Schwirtlich M, Mojsin M. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. Front Mol Neurosci. 2021 Mar 31; 14:654031. https://doi.org/10.3389/fnmol.2021.654031 PMID: 33867936

56. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb; 518(7539):317–30. https://doi.org/10.1038/nature14248 PMID: 25693563

57. Khoury A, Achinger-Kawecka J, Bert SA, Smith GC, French HJ, Luu PL, et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat Commun. 2020 Dec; 11(1):54. https://doi.org/10.1038/s41467-019-13753-7 PMID: 31911579

58. Kubo N, Ishii H, Xiong X, Bianco S, Meitinger F, Hu R, et al. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat Struct Mol Biol. 2021 Feb; 28(2):152–61. https://doi.org/10.1038/s41594-020-00539-5 PMID: 33398174

59. Franco MM, Prickett AR, Oakey RJ. The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction1. Biology of Reproduction [Internet]. 2014 Nov 1 [cited 2022 Aug 5]; 91(5). Available from: https://academic.oup.com/biolreprod/article-lookup/doi/10.1095/biolreprod.114.122945

60. Li H. New strategies to improve minimap2 alignment accuracy. Alkan C, editor. Bioinformatics. 2021 Dec 7; 37(23):4572–4.

61. Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 2018 Feb 9;

62. Reback J, McKinney W, brockmendel J, Bossche JVD, Augspurger T, Cloud P, et al. pandas-dev/pandas: Pandas 1.1.3 [Internet]. Zenodo; 2020 [cited 2020 Oct 28]. Available from: https://zenodo.org/record/3509134

63. Caswell TA, Droettboom M, Lee A, Hunter J, Firing E, Stansby D, et al. matplotlib/matplotlib v3.1.3 [Internet]. Zenodo; 2020 [cited 2020 Oct 28]. Available from: https://zenodo.org/record/3633844

64. Waskom M. seaborn: statistical data visualization. JOSS. 2021 Apr 6; 6(60):3021.

65. SciPy 1.0 Contributors, Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar; 17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543

66. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12(85):2825–30.

67. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011 May 2; 17(1):10.

68. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. 2010 [cited 2018 Aug 12]. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

69. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1; 32(19):3047–8. https://doi.org/10.1093/bioinformatics/btw354 PMID: 27312411

70. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan; 29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635 PMID: 23104886

71. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015 Jun 15; 31(12):2032–4. https://doi.org/10.1093/bioinformatics/btv098 PMID: 25697820

72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15; 25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

73. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016 Jul 8; 44(W1):W160–5. https://doi.org/10.1093/nar/gkw257 PMID: 27079975

74. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

75. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15; 26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

76. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc [Internet]. 2022 Apr 27 [cited 2022 May 30]; Available from: https://www.nature.com/articles/s41596-022-00692-9 https://doi.org/10.1038/s41596-022-00692-9 PMID: 35478247

77. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec; 15(12):550. https://doi.org/10.1186/s13059-014-0550-8 PMID: 25516281

78. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28; 38(4):576–89. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

79. Mi H, Ebert D, Muruganujan A, Mills C, Albou LP, Mushayamaha T, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Research. 2021 Jan 8; 49(D1):D394–403. https://doi.org/10.1093/nar/gkaa1106 PMID: 33290554

80. Barber A. Annotating Gene Expression and Regulatory Elements in Tissues from Healthy Thoroughbred Horses and Identifying Candidate Mutations Associated with Perosomus Elumbis in an Angus Calf. Theses and Dissertations in Animal Science. 2022 Apr; 233:143.

81. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature Methods. 2012 Mar; 9(3):215–6. https://doi.org/10.1038/nmeth.1906 PMID: 22373907

82. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Research. 2015 Jul 1; 43(W1):W39–49. https://doi.org/10.1093/nar/gkv416 PMID: 25953851