A Multi-omics Approach to Annotate the Horse Genome

By

SICHONG PENG

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Carrie J. Finno, Chair

Rebecca R. Bellone

Huaijun Zhou

Committee in Charge

2022

Dedication

To my aunt, in loving memory.

Acknowledgements

At the end of this long journey, I can't help but feel immense gratitude towards my major advisor and chair of my committee, Dr. Finno, whose guidance not only helped me navigate through the academic life but also empowered me to identify and pursue life-long career goals. Additionally, I could not have undertaken this journey without my committee, Dr. Bellone and Dr. Zhou, as well as my advisor Dr. Brown, who generously provided their knowledge and expertise.

This work is the culmination of a long-term collaborative project supported by a large group of colleagues, whose work laid down the foundation for my thesis. I had the pleasure of working with many of them and I would like to extend my sincere thanks to all of them. Especially Dr. Petersen, Dr. Kalbfleisch, Dr. Hales, Dr. Dahlgren, Janel Petersen, Annee Nguyen, and Dr. Donnelly, as well as members of the equine FAANG consortium, whose contributions were integral to the success of this project.

I am also grateful to all my family and friends, especially my cohort, who were incredibly supportive throughout this journey. And of course, I cannot imagine finishing this journey without my amazing partner, Kyle Richard, who always believed in me, even when I didn't.

Special thanks should go to my cat, Milo, whose cuddles and purrs always helped to calm my soul in difficult times.

Abstract

The genomic sequence of the horse has been available since 2007, providing critical resources for discovering important genomic variants regarding both animal health and population structure. However, to fully understand the functional implications of these variants, detailed annotation of the horse genome is required. Currently, the horse genome is annotated using the limited available RNA-seq data, as well as through comparative genomics by translating human and mouse genome annotation. While this approach has served the equine researchers well and led to a number of discoveries improving the care and management of horses, many important questions remain unanswered.

The limitation of the current annotation is two pronged. First, a comparative genomics approach is insufficient to identify many genes that are less evolutionarily conserved, especially those that are noncoding. The sole reliance on short-read RNA-seq data also meant that alternate isoforms could not be accurately resolved. Second, epigenomic regulatory elements are crucial to detailed understanding of gene expression network but are yet to be systemically identified in the horse. Many regulatory elements, including enhancers, promoters, and insulators, are not transcribed or transcribed at a very low level, necessitating alternate approaches to identify them. To solve these 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.

This thesis describes the equine FAANG team's effort to map tissue-specific gene expression and regulation in the horse genome. ***Chapter 1*** provides an overview of the equine FAANG project's approach to functional annotation. ***Chapter 2*** describes an improved transcriptome that includes novel genes and alternate isoforms compared to the current annotation. In ***Chapter 3***, we use ATAC-seq to create a catalog of tissue-specific open chromatin regions, which can serve as proxies to active

regulatory elements. ***Chapter 4*** provides a complete annotation of chromatin states across nine tissues. The ***Addendum*** detailed our effort to validate assays for transposase accessible chromatin using sequencing (ATAC-seq) in both frozen tissues and cryopreserved nuclei from fresh tissues. This thesis presents the first comprehensive overview of gene expression and their regulation in the horse, enabling interrogation of complex gene regulatory network and further studies of complex traits in horses. Future work should focus on both widening the scope of the equine FAANG project by including more tissue types and developmental stages, as well as refining gene network at single-cell resolution.

Table of Contents

List of Figures and Tables

Chapter 1 Decoding the Equine Genome: Lessons from ENCODE

Keywords: FAANG; gene regulation; horse; functional annotation; transcriptome; epigenetics; welfare; health

Published in: Peng, S.; Petersen, J.L.; Bellone, R.R.; Kalbfleisch, T.; Kings-ley , N.B.; Barber, A.M.; Cappelletti, E.; Giulotto, E.; Finno, C.J. Decoding the Equine Genome: Lessons from ENCODE. *Genes* 2021, 12, x. https://doi.org/10.3390/genes12111707.

Abstract

The horse reference genome assemblies, EquCab2.0 and EquCab3.0, have enabled great advancements in the equine genomics field, from tools to novel discoveries. However, significant gaps of knowledge regarding genome function remain, hindering the study of complex traits in horses. In an effort to address these gaps and with inspiration from the Encyclopedia of DNA Elements (ENCODE) project, the equine Functional Annotation of Animal Genome (FAANG) initiative was proposed to bridge the gap between genome and gene expression, providing further insights into functional regulation within the horse genome. Three years after launching the initiative, the equine FAANG group has generated data from more than 400 experiments using over 50 tissues, targeting a variety of regulatory features of the equine genome. In this review, we examine how valuable lessons learned from the ENCODE project informed our decisions in the equine FAANG project. We report the current state of the equine FAANG project and discuss how FAANG can serve as a template for future expansion of functional annotation in the equine genome and be used as a reference for studies of complex traits in horse. A well-annotated reference functional atlas will also help advance equine genetics in the pan-genome and precision medicine era.

1. The Horse Genome

The horse reference genomes (Equcab2.0 [1] and EquCab3.0 [2]) are based on a Thoroughbred mare Twilight and remain the only high-quality genome assemblies for equids. EquCab2.0 has 42,304 gaps comprising 55 Mb (2.2% of the genome) in total, with a scaffold N50 of 46 Mb. In comparison,

EquCab3.0 contains 3771 gaps comprising 9 Mb (0.34% of the genome) with a scaffold N50 of 86 Mb. It has 99.7% mammalian Benchmarking Universal Single-Copy Orthologs (BUSCO) (5 fragmented and 7 missing out of 4104 mammalian universal orthologs), compared to that of 99.0% (4064 complete orthologs) in EquCab2.0 [2]. Owing to the availability of a high-quality reference genome sequence, researchers have been able to utilize a wide variety of high-throughput tools to interrogate genetic etiologies for various equine traits. Recently, Raudsepp et al. provided a comprehensive review of major discoveries using combinations of recent technologies including genome-wide association studies (GWAS), whole-genome sequencing (WGS), and RNA-seq [3].

Using these tools, successful identification of the genetic variants responsible for simple Mendelian traits have been identified, including a novel variant in glutamate metabotropic receptor 6 (GRM6) associated with congenital stationary night blindness [4] and a nonsense variant in rap guanine nucleotide exchange factor 5 (RAPGEF5) as-sociated with equine familial isolated hypoparathyroidism [5]. However, many GWA studies conducted in horses have identified significant regions of association that do not contain any known genes. In humans, it was estimated that 88% of trait/disease as-sociated single nucleotide polymorphisms (SNPs) identified from GWAS were either intergenic or intronic [6]. These SNPs would later be recognized as enriched in various functional elements [7]. Since then, numerous studies have examined different mechanisms by which noncoding variants may affect phenotype. Variants near these significantly associated SNPs have been found to create transcription factor (TF) binding sites [8], disrupt binding motifs [9], or alter TF binding affinities [10,11].

These findings support the notion that many noncoding regions of DNA have important regulatory functions that affect gene expression. With a comprehensive registry of 926,535 human regulatory elements [12], it is now common to include functional annotation in the fine mapping of traits post-GWAS [13]. However, no such resources are available for most animal species, including horses. To

address this critical gap in knowledge, FAANG was proposed as an effort to identify important regulatory elements in the major livestock species [14].

2. Functional Annotation of Animal Genomes

The ENCODE initiative was proposed in 2003 as an ambitious effort to "identify all functional elements in the human genome sequence" [15]. In 2017, ENCODE concluded its third phase, delivering an integrated set of DNA transcription, regulation, and epigenetic modifications from a total of 7495 experiments in more than 500 cell types and tissues [12].

After almost two decades, ENCODE improved our understanding of gene regulation and delivered a wide range of computational tools, as well as a rich deposit of well-documented, publicly available experimental datasets [12]. Inspired by its phenomenal success, an international group of researchers proposed a similar, coordinated effort to systematically annotate animal genomes, providing vital resources to animal genetics research communities, termed Functional Annotation of Animal Genomes (FAANG) [14]. As part of the FAANG initiative, the equine FAANG group has been actively working with the larger FAANG community and ENCODE researchers to lead the annotation efforts for the horse genome.

The first stage of the equine FAANG initiative was to generate a biobank of reference tissues from comprehensively phenotyped animals. Burns et al. [16] and Donnelly et al. [17] detailed the phenotyping of four selected reference animals (UCD_AH1 – UCD_AH4) and a collection of over 80 tissues from each individual. These healthy animals were selected from the same breed (Thoroughbred) as Twilight, the horse used to construct the equine reference genome. When considering selection for the FAANG horses, the priority was placed on representing healthy Thoroughbred horses. Because Twilight was selected for the equine reference sequence based on homozygosity across the equine leukocyte antigen (ELA) region [1], the decision was made to include three unrelated Thoroughbreds and one (AH4) half-

sibling of Twilight to achieve this goal while still aligning well with the reference sequence. A unique aspect of this biobank is that horses were extensively phenotyped, both antemortem by experienced veterinarians and postmortem by veterinary pathologists. This not only ensured that there was no evidence of clinical or subclinical disease in these animals, but it also provided in-sight into the cellular composition of the tissues selected for assays. These tissues are stored at −80^{o}C in a biobank at UC Davis and are available to all equine FAANG re-searchers.

Here, we briefly discuss some of the most relevant findings from ENCODE and their implications for functionally annotating the equine genome.

3. Transcriptome

The transcriptome is the collection of all transcripts in an organism. It includes protein-coding mRNAs as well as noncoding RNAs. During the second phase of EN-CODE, 62% of the human genome was found to be transcribed with 31% of transcribed bases located in intergenic regions [18]. Many of these transcripts have been recognized as noncoding RNAs with important regulatory roles [19–23]. Additionally, in any cell line, 39% of the genome was transcribed on average. Up to 56.7% of transcriptome was detected in at least one of fifteen studied cell lines. Interestingly, only 7% of protein-coding genes were cell-line specific, while 53% were constitutive. In comparison, long-noncoding RNAs (lncRNAs) appeared to contribute more to cell-line specificity, with 29% of lncRNAs detected in only one of the fifteen studied cell lines and 10% ex-pressed in all cell lines [18]. These results highlighted the necessity of characterizing transcriptome in a cell-specific manner.

As part of ENCODE, GENCODE was initially founded to provide high-quality reference gene annotation for the human genome and subsequently expanded into a long-running partnership between several groups and institutes. In its most recent re-lease based on GRCh38, a total of 60,649 genes have been identified in the human genome, of which 19,955 are protein coding, with an average isoform-to-gene

ratio of 3.9 [24]. It was also demonstrated that genes tend to express many isoforms simultaneous-ly, with a dominant isoform comprising 30% or more of its corresponding gene expression. Isoforms also appeared to contribute to cell type specificity, with over 75% of protein-coding genes having different dominant isoforms in different cell lines [18].

In addition to protein-coding transcripts, the transcriptome also consists of many noncoding RNA species, including both small and long noncoding RNAs. The functions of these RNAs have been extensively examined and implicated in important bio-logical pathways [25–28]. The small noncoding RNAs present a unique opportunity to new therapeutic approaches [29]. Extensive efforts have been put into cataloguing noncoding RNAs in the human and mouse genome [30,31]. These efforts have further detailed the extent of noncoding RNA regulatory network and the diversity of noncoding RNA species and their functions.

Taken together, these findings from ENCODE demonstrated the importance of noncoding RNAs and of alternative splicing in cell-specific expression and regulation. Both Ensembl [32] and RefSeq [33] provide noncoding RNA and isoform annotation for EquCab3.0 by utilizing the high-quality annotation of the human genome as well as publicly available horse RNA-seq data. RefSeq annotation for EquCab3.0 consists of 30,022 genes, of which 21,129 are protein coding, with an average isoform-to-gene ratio of 2.6 [34]. The Ensembl annotation of the equine genome contains 30,371 genes (20,955 protein coding) with an average isoform-to-gene ratio of 1.9 [35]. Assuming the human and equine genomes have a similar number of genes and consistent isoform-to-gene ratio, the current horse gene annotation likely lacks many noncoding RNAs and alternate isoforms.

The FAANG initiative proposed RNA-seq assays for both mRNA and smRNA to identify and quantify these transcripts in a tissue-specific manner [14]. These assays have been performed for eight prioritized tissues (liver, lamina, heart, parietal cortex, adipose, skeletal muscle, ovary/testis, and lung) (Table 1.1).

Table 1.1 Overview of Available Data and Assay Details						
Project Accession	**Assay**	**Samples**	**Tissues**	**Instrument**	**Library Layout**	**Number of Experiments**
PRJEB26698	WGS	Two females	1	HiSeq 2500 (San Diego, CA, USA)	2 × 250 bp	2
PRJEB42407	WGS	Two males	1	NovaSeq 6000 (San Diego, CA, USA)	2 × 150 bp	2
PRJEB26787	RNA-seq	Two females	30	HiSeq 2500 (San Diego, CA, USA)	2 × 250 bp	60
PRJEB32645	RRBS	Two females	10	HiScanSQ (San Diego, CA, USA)	1 × 50 bp	20
PRJEB35307	Histone ChIP-seq	Two females	8	HiSeq 4000 (San Diego, CA, USA)	1 × 50 bp	80
PRJEB42315	Histone ChIP-seq	Two females	4	HiSeq 4000 (San Diego, CA, USA)	1 × 50 bp	38
PRJEB41079	CTCF ChIP-seq	Two females	8	HiSeq 4000 (San Diego, CA, USA)	1 × 50 bp	28
PRJEB41317	ATAC-seq pilot	Two females	2	HiSeq 4000/NextSeq 500 (San Diego, CA, USA)	2 × 75 bp/2 × 42 bp	16
WGS: whole-genome sequencing; RNA-seq: mRNA sequencing; RRBS: reduced-representation bisulfite sequencing; Histone ChIP-seq: chromatin immunoprecipitation using sequencing for the four major histone marks; CTCF ChIP-seq: chromatin immunoprecipitation using sequencing for CTCF protein; ATAC-seq pilot: assay for transposase accessibility using sequencing.						

To facilitate data generation for the remaining biobanked tissues, we proposed a unique “Adopt-A-Tissue” model for mRNA-seq. Researchers were invited to “adopt” a tissue or tissues fitting their research interests, which meant they would cover the as-say and sequencing costs. All library preparations and sequencing were performed at the same two locations (female samples at UC Davis, male samples at University of Nebraska-Lincoln) to minimize variability. This approach allowed the

community to contribute to the initiative together while still being able to limit technical variations across laboratories during library constructions [36]. Owing to this unique strategy, the equine community has sequenced over 40 tissues, and the data have been made publicly available (Table 1.1).

More recently, long-read sequencing assays such as PacBio Isoform sequencing (Iso-seq) have emerged as powerful tools to determine the splicing patterns of transcripts. To address the poor isoform annotations currently available for the horse genome, Iso-seq assays are being performed in 8 tissues (liver, lung, lamina, heart, ovary, testis, muscle, skin, and parietal cortex) across eight PacBio Sequel 8M SMRT cells. By combining a wide variety of assays, the equine FAANG initiative aims to deliver a comprehensively annotated transcriptome for the horse genome.

4. Chromatin Accessibility

In mammalian cells, DNA molecules are packed by histone proteins to form nucleosomes and are subsequently compacted into chromatin [37,38]. Compact chromatin restricts access to DNA molecules by transcription factors and serves as a way to regulate gene expression [39]. For example, nucleosomes are densely arranged in facultative and constitutive heterochromatin while depleted in active regions such as active enhancers, insulators, and transcribed gene bodies [40,41]. Using DNase-seq, a DNase I assay quantifying susceptibility of chromatin to DNase I, Boyle et al. identified 94,925 DNase I hypersensitive sites (DHS) covering 2.1% of the human genome [42]. It was also found that only 13% of DHS were located within promoters, while up to 78% were in intergenic or intronic regions. Remarkably, DHS were found in or near the transcription start sites (TSS) of nearly all highly expressed genes. However, while DNase I hypersensitivity appeared to be necessary for gene expression, it was not sufficient as DHS were also observed in unexpressed genes [42]. The association between accessible chromatin and active elements present a unique opportunity to study tissue- and cell-specific gene regulation [43–47].

Echoing their strong functional implications, accessible chromatin was also shown to be associated with noncoding variants identified in GWAS studies of common traits. Maurano et al. examined 5654 noncoding variants identified in the GWAS studies of 207 diseases and 447 quantitative traits and found 76.6% of these variants lie either within a DHS or in complete linkage disequilibrium (LD) with another SNP in DHS [48]. The data further demonstrated that many of these DHS were strongly correlated with the promoter of a distal gene target [48]. Gusev et al. analyzed the heritability of 11 common diseases and found that SNPs contained within DHS explained up to 79% of heritability [49]. The strong association between accessible chromatin and functional elements warranted efforts to establish a catalog of tissue-specific DHS to facilitate discoveries of functionally relevant variants [47].

Although DNase-seq has proven successful in identifying accessible chromatin, its laborious protocol, slow turn-around time, and large sample size requirements severely limit large-scale applications [50,51]. Buenrostro et al. developed Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq), which greatly reduced both time and labor costs while requiring lower nuclei input [51]. Owing to its simple protocol and comparable output [52], ATAC-seq has been widely adopted as a state-of-the-art method for interrogating genome-wide chromatin accessibility; further, several variations in methodology have been developed to apply ATAC-seq to frozen tissues [53], cryopreserved nuclei [54], or to improve sensitivity in low-input materials [55].

Using ATAC-seq on cryopreserved nuclei from eight tissues across pig, cattle, and mouse, Halstead et al. showed a lack of conservation of sequence and accessibility in accessible sites across evolutionary distance, with 20% shared sites between pig and cattle and only 10% between mouse and ungulates [56]. Therefore, it is necessary to establish a tissue-specific catalog of accessible sites specifically for the horse genome. A pilot study was recently carried out to evaluate the suitability of frozen equine tissue derived nuclei for ATAC-seq [57]. Following protocols established by this study, additional ATAC-seq

experiments are underway to expand this assay to eight prioritized tissues for the equine FAANG project.

5. Histone Modifications

Histone proteins form the basic building blocks of hierarchical chromatin structures and have been recognized to play an important role in modulating gene expression through post-transcriptional modifications [58–61]. A nucleosome core is formed by two copies of each of the four major types of histone proteins: H2A, H2B, H3, and H4 [62]. Since Allfrey first suggested the potential role of histone acetylation in regulating gene expression in 1964 [58], extensive research has been carried out to under-stand the roles, mechanisms, and implications of different histone modifications. His-tone 3 lysine 4 monomethylation (H3K4me1), H3K4me3, H3K27me3, and H3K27ac are among some of the most studied and best understood modifications. Hyun et al. pro-vided a detailed review of molecular mechanisms associated with histone lysine modifications and their regulatory functions [63]. Here, we briefly discuss ENCODE findings regarding histone marks and how they can be integrated to provide a more com-prehensive view of regulatory activities.

Barski et al. first comprehensively assayed histone modifications across the hu-man genome using high-throughput sequencing [64]. Consistent with previous studies, H3K4 methylation marks were enriched in promoter regions. A significant drop in signal between −200 bp and +50 bp of TSS was observed for H3K4me3 with major peaks at −300 bp and +100 bp [64]. This was consistent with observations that H3K4me3 was primarily associated with promoter regions [65] and that nucleosomes were depleted near active TSS [40]. On the other hand, H3K4me1 showed a distinct bimodal signal with peaks around –900 bp and +1000 bp of TSS [64], in agreement with previous observations that H3K4me1 was enriched in enhancer regions [66]. Similarly, H3K27me3 was observed at a higher level around the TSS of silent genes than those around active genes, supporting correlation between H3K27me3 and gene repression

[67]. Conversely, H3K27ac was observed around active elements and associated with higher expression level [68].

Taken together, the four histone modifications discussed in this manuscript rep-resent major regulatory elements and can provide valuable information regarding tis-sue-specific regulatory activities in the horse genome. Using genome-wide chromatin immunoprecipitation sequencing (ChIP-seq) for these four marks in eight prioritized tissues in the two female FAANG horses, Kingsley et al. reported over one million putative regulatory sites [69]. The utility of these data was demonstrated when a 16 kB intergenic deletion associated with an ocular condition in horses, namely distichiasis, was discovered and FAANG ChIP-seq data showed that this region harbors a tissue specific active enhancer [70]. Undoubtedly, these data will continue to aid in the understanding of other structural variants causing or associated with disease in the horse as additional tissues are evaluated. Following the success of the mRNA Adopt-A-Tissue initiative, similar efforts have facilitated characterization of histone marks in four tissues important to equine health and traits of economic impact (spleen, meta-carpal 3, sesamoid, and skin) [71]. Furthermore, additional Adopt-A-Tissue efforts are currently ongoing to facilitate histone ChIP-seq assays for the remaining FAANG tis-sues.

6. CTCF Binding

CCCTC-binding factor (CTCF) is a well-studied zinc finger protein that serves a central role in the formation of chromatin topology and remodeling. It was first dis-covered as a repressive transcription factor in chicken for c-MYC [72] as well as LYZ [73]. It was later shown that CTCF may also serve as an activator for the Amyloid β-Protein Precursor gene (APP) [74]. In 1999, Bell et al. reported a CTCF binding site at the core of an insulator element at the 5’ end of the chicken β-globin gene HBB [75]. Insulators are genomic regions that separate genes from cis-regulatory elements [76]. This site also sits

at a boundary between active and inactive chromatin [77], a typical feature of an insulator element [78,79].

Many seemingly contradictory functions of CTCF have attracted extensive efforts to understand the mechanisms of its multivalent roles. CTCF is highly conserved across species [80,81] and embryonically lethal when knocked out in mice [82]. The binding motif of CTCF consists of a ~20 bp core consensus sequence and less conserved peripheral sequences, comprising ~50 bp [83,84]. ChIP assays targeting CTCF revealed several unique patterns. First, CTCF binding sites were observed across the genome, with over 40% within intergenic regions [64,83,85]. Consistent with the insulator activity of CTCF, two distinct types of loci with opposing CTCF binding patterns were observed. Loci depleted of CTCF binding sites tend to include clusters of related gene families and transcriptionally coregulated genes, while loci enriched in CTCF binding sites tend to have genes with alternative promoters [83]. Furthermore, CTCF was shown to be crucial for chromatin loop formation at the mouse β-globin locus [86]. Similarly, Hou et al. described an alternative loop formation by inserting a CTCF binding insulator HS5 between the β-globin locus and its upstream locus control region [87]. Additionally, cohesin has been functionally associated with CTCF in mediating chromatin loops [88,89]. These results suggested a potential mechanism via which CTCF mediates regulation of chromatin conformation and gene expression.

The introduction of Hi-C technology that enabled genome-wide interrogation of long-range interactions [90] quickly brought about new insights into the mechanisms of CTCF function. Refining the resolution of the Hi-C interaction maps to kilobases, Rao et al. observed that the majority of chromatin loops were associated with convergent pairs of CTCF motifs, as well as colocalizing with cohesin proteins [91]. The orientation of CTCF motifs was also shown to determine the directionality of the CTCF mediated interactions [92]. Finally, the significance of such directionality was functionally demonstrated by inverting CTCF sites with CRISPR to alter genome topology as well as promoter function [93].

These findings led to a proposed extrusion model [94,95], where a chromatin loop is pulled through an extrusion complex consisting of cohesin and CTCF and is stabilized by a CTCF dimer. This model explains the convergence of a CTCF pair surrounding a chromatin loop, as well as the many regulatory functions of CTCF observed in early studies. More evidence is emerging in support of this model. Based on this model, Fudenberg et al. used simulation to reproduce topologically associated domains (TADs) and contact frequencies observed in Hi-C studies as well as to recapitulate experimental results where TADs were observed to spread upon depletion of CTCF binding sites [96]. Haarhuis et al. showed that cohesin release factor WAPL could restrict chromatin loop extrusion by releasing cohesin from DNA and that knocking out WAPL results in enlarged chromatin loops between incorrectly orientated CTCF motifs [97]. Allahyar et al., employing a multi-contact 4C technology, showed that such enlarged loops in WAPL knockout cells are a result of aggregated CTCF loop anchors, or a “cohesin traffic jam” [98].

Given its central role in chromatin loop formation, CTCF binding sites can be considered an intermediate between the 1D genomic sequence and 3D chromatin topology. Although there is no simple rule to determine the functional outcome of a disrupt-ed CTCF binding site, as it largely depends on its interaction with surrounding regulatory elements, there is no doubt that a catalog of CTCF binding sites in a given cellular context can provide valuable information when decoding the functional implications of DNA variants.

Following the practices established by the FAANG community, characterization of CTCF binding sites using ChIP-seq is being performed on eight prioritized tissues for both sexes. Analyses to identify both tissue and sex-specific CTCF binding and integrate all of the FAANG ChIP-seq data into chromatin state annotations are currently underway.

7. Chromatin States

While the associations between individual histone marks and regulatory activities are noteworthy, combinations of histone marks have proven to be more reliable in the fine-scale predictions of regulatory elements. For example, Creyghton et al. observed that the H3K27ac mark could distinguish active enhancers from inactive/poised enhancers, which are both marked by H3K4me1 [68]. Bernstein et al. similarly identified a bivalent signal with both H3K4 methylation and H3K27 methylation, suggesting a poised regulatory element [99]. These findings prompted hypotheses that various regulatory functions of noncoding DNA could be explained by either additive properties [100] or unique combinations of histone modifications [101]. New unsupervised computational approaches were subsequently developed to classify histone modification patterns and partition them into different chromatin states [102,103]. Ernst et al. identified 11 promoter states, all marked by H3K4me3 and varying presence and levels of several other marks, as well as 4 enhancer-associated states, all marked by H3K4me1 and varying frequencies of acetylation marks [103]. These findings suggest that some histone modifications (H3K4me1, H3K4me3) designate unique regulatory elements while other modifications (acetylation marks including H3K27ac) enhance regulatory activity in an additive fashion.

The recognition of chromatin states and introduction of computation tools such as ChromHMM [104] provided a way to systematically profile the regulatory landscape in any given cellular context. Taking advantage of this development and the availability of ChIP-seq data from the four major histone marks and CTCF, efforts to compose an integrated tissue-specific chromatin state map are currently underway for the equine genome.

8. Unique Aspects of the Horse Genome

Centromeres are enigmatic structures because, contrary to other genetic loci, their function is not determined by the underlying DNA sequence but depends on epigenetic factors. The Centromere

Protein A (CENP-A) is a centromere-specific variant of his-tone H3 that epigenetically identifies, maintains, and propagates centromere function [105]. The characteristics of its binding domain have been elusive to investigators due to its typical association with tandemly repeated DNA (satellite DNA). In this context, a turning point was the discovery that the centromere of horse chromosome 11 (ECA11) was completely devoid of satellite DNA, demonstrating for the first time that a natural mammalian centromere, fixed in a species, can exist without satellite sequences [1]. Owing to the lack of satellite repeats at the centromere of ECA11 and the availability of the horse reference genome, the genomic position of the corresponding CENP-A binding domain could be precisely identified by ChIP-on-chip with an anti-CENP-A antibody [1]. Later, several satellite-less centromeres were identified by ChIP-seq in the donkey genome [106]. These peculiar centromeres found in equid species represent an immature stage of “centromerization”, being the result of centromere re-positioning, which is the movement of the centromeric function without detectable chromosomal rearrangements. This event was exceptionally frequent during the rapid evolution of the genus Equus [107–109]. Such centromeres, being uncoupled from satellite DNA, provide a unique model for dissecting the molecular structure of the centromere [110].

The position of the ECA11 satellite-less centromere, identified as the CENP-A binding domain, is not fixed in the horse population but slides within an about the 500 kb region, giving rise to different positional alleles or “epialleles” [106,111,112]. The analysis of these epialleles carried out on families composed by horses, donkeys, and their hybrid offspring (mule/hinny) revealed that they are inherited as Mendelian traits, but their position can slide in one generation [106]. Conversely, the position of the centromere is stable during mitotic propagation of cultured cells grown for several population doublings, suggesting that the sliding may presumably take place during meiosis or early embryogenesis [106].

The absence of satellite DNA at these centromeres also provides a unique opportunity to understand whether some typical features of mammalian centromeres de-pend on the presence of satellite DNA. In particular, it was possible to demonstrate that satellite DNA was not necessary for segregation fidelity of the centromere [113] and was not implicated in the suppression of meiotic recombination, which is typically exerted by the centromere [112].

The rich repository of tissues from different developmental origins available through the FAANG project will allow us to answer other important questions on centromere biology using the ECA11 centromere as model system. We will test whether the centromere position is conserved during development or if it can slide during tis-sue differentiation. In addition, thanks to the large amount of data regarding the functional annotation of the horse genome, generated within the FAANG effort, we will be able to map the epigenetic marks available through the consortium in the ECA11 centromeric region. The results will indicate whether chromatin markers and transcriptional activity at ECA11 centromere vary across tissues and individuals, and with respect to centromere position. Furthermore, CENP-A has been shown to bind at TF binding sites and promoters, suggesting potential regulatory activities [114]. Therefore, utilizing FAANG data, we will be able to identify the regulatory activities of CENP-A and any roles centromeres may play during tissue differentiation

9. Summary and Future Perspectives

Just three years after starting the tissue and data collection for the equine FAANG initiative, the community has completed over 400 experiments from more than 50 tis-sues using a variety of assays targeting different features of the horse regulatory landscape (Table 1.2). Data are being made available to the public as they are generated and evaluated for passing quality control measures; these data have been and continue to be utilized in unrelated research projects [5,70,115]. Integrated analysis is

currently on-going to provide a systematic annotation of major functional elements in the horse genome available, as a central hub hosted on UCSC genome browser to the research community.

Table 1.2 Overview of Completed Assays

Assay	Animals	Tissue Types	Total Experiments
WGS	AH1-AH4	Blood	4
mRNA-seq	AH1	47	140
	AH2	46	
	AH3	23	
	AH4	24	
Iso-seq	AH1–AH4	12	48
ChIP-seq–H3K4me1	AH1–AH2	12	40
	AH3–AH4	8	
ChIP-seq–H3K4me3	AH1–AH2	12	40
	AH3–AH4	8	
ChIP-seq–H3K27ac	AH1–AH2	12	40
	AH3–AH4	8	
ChIP-seq–H3K27me3	AH1–AH2	12	40
	AH3–AH4	8	
ChIP-seq–CTCF	AH1–AH2	8	32
	AH3–AH4	8	
ATAC-seq	AH1–AH4	10	40
RRBS	AH1–AH2	10	20
smRNA-seq	AH1–AH2	48	96
Total		**48**	**444**

With over 80 tissues collected from four healthy and comprehensively phenotyped animals, we will be able to generate a map of gene expression and regulation throughout the horse body, providing unique opportunities to investigate tissue-specific gene expression and gene networks. However, this tissue

collection presents a serious challenge for data analyses. Heterogeneity both within tissues as a result of cell-type differentiation and across tissues as a result of tissue infiltration or contamination during collection, can confound analysis of tissue-specific expression and regulation. The prevalence of this issue was recently reported by Sturm et al. [116]. To mitigate this issue, careful histological assessment was performed during the tissue collection phase to minimize the possibility of tissue infiltration or contamination. However, caution should be taken to assess the extent of tissue heterogeneity during data analysis. Additionally, single-cell based technologies have proven useful to profile cell types from complex tissues [117–120], and the adoption of these technologies to equine FAANG data are being discussed within the community and will likely be integrated in the next steps of the multi-phased approach of this project.

While the equine FAANG biobank represents a wide variety of tissue types, the four horses these tissues were collected from represent only a narrow subset of the horse population, as well as developmental stages. These horses were intentionally selected to be of the same breed as the reference genome assembly in order to better an-notate the reference genome assembly. However, caution should be taken with interpretation and extrapolation of these data to other breeds or developmental stages. Regardless, this initiative will serve as a template and reference point for the future expansion of the transcriptome and epigenome of equids.

FAANG represents a notable international collaborative effort in the equine community that has brought together equine researchers and practitioners from around the globe. Most importantly, FAANG collaborators have been vocal proponents of open science and broad data accessibility within the equine community. The growing number of publicly available datasets is accelerating discoveries and powering large-scale analyses. Well-annotated and carefully documented FAANG data with ac-companying comprehensive metadata will serve as a reference point for many future discoveries in horse.

Acknowledgments: We would like to acknowledge UC Davis FAANG group for their in-put in experimental designs and data analyses as well as three anonymous reviewers for their suggestions and comments.

Disclosure: Portions of this work were supported by Animal Breeding and Functional Annotation of Genomes (A1201) Grant 2019-67015-29340/Project Accession 1018854 from the USDA National Institute of Food and Agriculture, the Grayson Jockey Club Foundation, USDA NRSP-8 and the UC Davis Center for Equine Health, Italian Ministry of Education, University and Research (MIUR) [Dipartimenti di Eccellenza Program (2018–2022)—Dept. of Biology and Biotechnology "L. Spallanzani", University of Pavia]. Support for C.J.F was provided by the National Institutes of Health (NIH) (L40 TR001136). None of the funding agencies had any role in the design of the study, analysis, interpretation of the data, or writing of the manuscript.

References

1. Wade, C. M. et al. Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse. Science 326, 865–867 (2009).
2. Kalbfleisch, T. S. et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun. Biol. 1, 197 (2018).
3. Raudsepp, T., Finno, C. J., Bellone, R. R. & Petersen, J. L. Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post‐genome era. Anim. Genet. 50, 569–597 (2019).
4. Hack, Y. L. et al. Whole‐genome sequencing identifies missense mutation in GRM6 as the likely cause of congenital stationary night blindness in a Tennessee Walking Horse. Equine Vet. J. 53, 316–323 (2021).
5. Rivas, V. N. et al. A nonsense variant in Rap Guanine Nucleotide Exchange Factor 5 (RAPGEF5) is associated with equine familial isolated hypoparathyroidism in Thoroughbred foals. PLOS Genet. 16, e1009028 (2020).
6. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. 106, 9362–9367 (2009).
7. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
8. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466, 714–719 (2010).

9. Bauer, D. E. et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science 342, 253–257 (2013).
10. Tuupanen, S. et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat. Genet. 41, 885–890 (2009).
11. Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c- MYC in cis through a Large Chromatin Loop Linked to a Cancer Risk-Associated Single-Nucleotide Polymorphism in Colorectal Cancer Cells. Mol. Cell. Biol. 30, 1411–1420 (2010).
12. The ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710 (2020).
13. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: Illuminating the Dark Road from Association to Function. Am. J. Hum. Genet. 93, 779–797 (2013).
14. Andersson, L. et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 16, 57 (2015).
15. Consortium, T. E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636–640 (2004).
16. Burns, E. N. et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim. Genet. 49, 564–570 (2018).
17. Donnelly, C. G. et al. Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. Front. Genet. 12, 650305 (2021).
18. Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).
19. Schmitz, S. U., Grote, P. & Herrmann, B. G. Mechanisms of long noncoding RNA function in development and disease. Cell. Mol. Life Sci. 73, 2491–2509 (2016).
20. Turner, M., Galloway, A. & Vigorito, E. Noncoding RNA and its associated proteins as regulatory elements of the immune system. Nat. Immunol. 15, 484–491 (2014).
21. Lin, N. et al. An Evolutionarily Conserved Long Noncoding RNA TUNA Controls Pluripotency and Neural Lineage Commitment. Mol. Cell 53, 1005–1019 (2014).
22. Long, J. et al. Long noncoding RNA Tug1 regulates mitochondrial bioenergetics in diabetic nephropathy. J. Clin. Invest. 126, 4205–4218 (2016).
23. St. Laurent, G., Wahlestedt, C. & Kapranov, P. The Landscape of long noncoding RNA classification. Trends Genet. 31, 239–251 (2015).
24. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2019).
25. Meller, V. H., Joshi, S. S. & Deshpande, N. Modulation of Chromatin by Noncoding RNA. Annu. Rev. Genet. 49, 673–695 (2015).
26. Shen, Y. et al. Insights into Enhancer RNAs: Biogenesis and Emerging Role in Brain Diseases. The Neuroscientist 107385842110468 (2021) doi:10.1177/10738584211046889.
27. Moazzendizaji, S. et al. microRNAs: Small molecules with a large impact on colorectal cancer. Biotechnol. Appl. Biochem. bab.2255 (2021) doi:10.1002/bab.2255.
28. Wen, Z.-J. et al. Emerging roles of circRNAs in the pathological process of myocardial infarction. Mol. Ther. - Nucleic Acids S2162253121002456 (2021) doi:10.1016/j.omtn.2021.10.002.
29. Winkle, M., El-Daly, S. M., Fabbri, M. & Calin, G. A. Noncoding RNA therapeutics — challenges and potential solutions. Nat. Rev. Drug Discov. 20, 629–651 (2021).
30. He, P. et al. The changing mouse embryo transcriptome at whole tissue and single-cell resolution. Nature 583, 760–767 (2020).

31. Lorenzi, L. et al. The RNA Atlas expands the catalog of human non-coding RNAs. Nat. Biotechnol. (2021) doi:10.1038/s41587-021-00936-1.
32. Howe, K. L. et al. Ensembl 2021. Nucleic Acids Res. 49, D884–D891 (2021).
33. Pruitt, K. D. et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42, D756–D763 (2014).
34. Equus caballus RefSeq Annotation Release 103. RefSeq https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Equus_caballus/103/.
35. Equus caballus Ensembl Annotation Release 105.3. Ensembl https://uswest.ensembl.org/Equus_caballus/Info/Annotation.
36. McIntyre, L. M. et al. RNA-seq: technical variability and sampling. BMC Genomics 12, 293 (2011).
37. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA. Science 184, 868–871 (1974).
38. Olins, D. E. & Olins, A. L. Chromatin history: our view from the bridge. Nat. Rev. Mol. Cell Biol. 4, 809–814 (2003).
39. Lorch, Y., LaPointe, J. W. & Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. Cell 49, 203–210 (1987).
40. Lee, C.-K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat. Genet. 36, 900–905 (2004).
41. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat. Rev. Genet. 7, 703–713 (2006).
42. Boyle, A. P. et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132, 311–322 (2008).
43. Stergachis, A. B. et al. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell 154, 888–903 (2013).
44. Song, L. et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 21, 1757–1767 (2011).
45. Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res. 22, 1711–1722 (2012).
46. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).
47. Meuleman, W. et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature 584, 244–251 (2020).
48. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–1195 (2012).
49. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. 95, 535–552 (2014).
50. Crawford, G. E. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 16, 123–131 (2005).
51. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC‐seq: A Method for Assaying Chromatin Accessibility Genome‐Wide. Curr. Protoc. Mol. Biol. 109, (2015).
52. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962 (2017).
53. 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).

54. Sos, B. C. et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol. 17, 20 (2016).
55. Halstead, M. M. et al. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. BMC Genomics 21, 698 (2020).
56. Peng, S., Bellone, R., Petersen, J. L., Kalbfleisch, T. S. & Finno, C. J. Successful ATAC-Seq From Snap-Frozen Equine Tissues. Front. Genet. 12, 641788 (2021).
57. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. Proc. Natl. Acad. Sci. U. S. A. 51, 786–794 (1964).
58. Yang, X.-J. & Seto, E. HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. Oncogene 26, 5310–5318 (2007).
59. Oki, M., Aihara, H. & Ito, T. Role of histone phosphorylation in chromatin dynamics and its implications in diseases. Subcell. Biochem. 41, 319–336 (2007).
60. Bedford, M. T. & Clarke, S. G. Protein arginine methylation in mammals: who, what, and why. Mol. Cell 33, 1–13 (2009).
61. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature 389, 251–260 (1997).
62. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. Exp. Mol. Med. 49, e324–e324 (2017).
63. Barski, A. et al. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129, 823–837 (2007).
64. Liang, G. et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc. Natl. Acad. Sci. U. S. A. 101, 7357–7362 (2004).
65. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. 39, 311–318 (2007).
66. Boyer, L. A. et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441, 349–353 (2006).
67. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. 107, 21931–21936 (2010).
68. Kingsley, N. B. et al. Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. Genes 11, 3 (2019).
69. Hisey, E. A. et al. Whole genome sequencing identified a 16 kilobase deletion on ECA13 associated with distichiasis in Friesian horses. BMC Genomics 21, 848 (2020).
70. Kingsley, N. B. et al. “Adopt-a-Tissue” Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare. Front. Genet. 12, 649959 (2021).
71. Lobanenkov, V. V. et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5’-flanking sequence of the chicken c-myc gene. Oncogene 5, 1743–1753 (1990).
72. Baniahmad, A., Steiner, C., Köhne, A. C. & Renkawitz, R. Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. Cell 61, 505–514 (1990).
73. Vostrov, A. A. & Quitschke, W. W. The Zinc Finger Protein CTCF Binds to the APBβ Domain of the Amyloid β-Protein Precursor Promoter. J. Biol. Chem. 272, 33353–33359 (1997).
74. Bell, A. C., West, A. G. & Felsenfeld, G. The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. Cell 98, 387–396 (1999).

75. Dorsett, D. Distance-independent inactivation of an enhancer by the suppressor of Hairy-wing DNA-binding protein of Drosophila. Genetics 134, 1135–1144 (1993).
76. Hebbes, T. R., Clayton, A. L., Thorne, A. W. & Crane-Robinson, C. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. EMBO J. 13, 1823–1830 (1994).
77. Kellum, R. & Schedl, P. A group of scs elements function as domain boundaries in an enhancer-blocking assay. Mol. Cell. Biol. 12, 2424–2431 (1992).
78. Udvardy, A., Maine, E. & Schedl, P. The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. J. Mol. Biol. 185, 341–358 (1985).
79. Filippova, G. N. et al. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol. Cell. Biol. 16, 2802–2813 (1996).
80. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet. 17, 520–527 (2001).
81. Heath, H. et al. CTCF regulates cell cycle progression of αβ T cells in the thymus. EMBO J. 27, 2839–2850 (2008).
82. Kim, T. H. et al. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell 128, 1231–1245 (2007).
83. Nakahashi, H. et al. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. Cell Rep. 3, 1678–1689 (2013).
84. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res. 36, 5221–5231 (2008).
85. Splinter, E. et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. Genes Dev. 20, 2349–2354 (2006).
86. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. Proc. Natl. Acad. Sci. 105, 20398–20403 (2008).
87. Wendt, K. S. et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 451, 796–801 (2008).
88. Parelho, V. et al. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. Cell 132, 422–433 (2008).
89. Lieberman-Aiden, E. et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009).
90. Rao, S. S. P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).
91. Vietri Rudan, M. et al. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Rep. 10, 1297–1309 (2015).
92. Guo, Y. et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell 162, 900–910 (2015).
93. Nichols, M. H. & Corces, V. G. A CTCF Code for 3D Genome Architecture. Cell 162, 703–705 (2015).
94. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. 112, E6456–E6465 (2015).
95. Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. 15, 2038–2049 (2016).

96. Haarhuis, J. H. I. et al. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell 169, 693-707.e14 (2017).
97. Allahyar, A. et al. Enhancer hubs and loop collisions identified from single-allele topologies. Nat. Genet. 50, 1151–1160 (2018).
98. Bernstein, B. E. et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125, 315–326 (2006).
99. Schreiber, S. L. & Bernstein, B. E. Signaling Network Model of Chromatin. Cell 111, 771–778 (2002).
100. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. Nature 403, 41–45 (2000).
101. Hon, G., Ren, B. & Wang, W. ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome. PLoS Comput. Biol. 4, e1000201 (2008).
102. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. 28, 817–825 (2010).
103. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods 9, 215–216 (2012).
104. Fachinetti, D. et al. A two-step mechanism for epigenetic specification of centromere identity and function. Nat. Cell Biol. 15, 1056–1066 (2013).
105. Nergadze, S. G. et al. Birth, evolution, and transmission of satellite-free mammalian centromeric domains. Genome Res. 28, 789–799 (2018).
106. Carbone, L. et al. Evolutionary movement of centromeres in horse, donkey, and zebra. Genomics 87, 777–782 (2006).
107. Piras, F. M. et al. Phylogeny of Horse Chromosome 5q in the Genus Equus and Centromere Repositioning. Cytogenet. Genome Res. 126, 165–172 (2009).
108. Giulotto, E., Raimondi, E. & Sullivan, K. F. The Unique DNA Sequences Underlying Equine Centromeres. in Centromeres and Kinetochores (ed. Black, B. E.) vol. 56 337–354 (Springer International Publishing, 2017).
109. Purgato, S. et al. Centromere sliding on a mammalian chromosome. Chromosoma 124, 277–287 (2015).
110. Cappelletti, E. et al. CENP-A binding domains and recombination patterns in horse spermatocytes. Sci. Rep. 9, 15800 (2019).
111. Roberti, A. et al. Satellite DNA at the Centromere is Dispensable for Segregation Fidelity. Genes 10, E469 (2019).
112. Athwal, R. K. et al. CENP-A nucleosomes localize to transcription factor hotspots and subtelomeric sites in human cancer cells. Epigenetics Chromatin 8, 2 (2015).
113. Gao, S. et al. Comparative Transcriptome Profiling Analysis Uncovers Novel Heterosis-Related Candidate Genes Associated with Muscular Endurance in Mules. Anim. Open Access J. MDPI 10, E980 (2020).
114. Sturm, G., List, M. & Zhang, J. D. Tissue heterogeneity is prevalent in gene expression studies. NAR Genomics Bioinforma. 3, lqab077 (2021).
115. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64 (2013).
116. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401 (2014).
117. Pott, S. & Lieb, J. D. Single-cell ATAC-seq: strength in numbers. Genome Biol. 16, 172 (2015).

118. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat. Biotechnol. 33, 1165–1172 (2015).

Chapter 2 Long-read RNA sequencing improves the annotation of the equine transcriptome

Keywords: FAANG; horse; transcriptome; annotation

Manuscript in preparation: Sichong Peng, Anna R. Dahlgren, Erin N. Hales, Alexa M. Barber, Callum G. Donnelly, Ted Kalbfleisch, Jessica L. Petersen, Rebecca R. Bellone, Mariusz Mackowski, Katia Cappelli, Stefano Capomaccio, Stephen J. Coleman, Ottmar Distl, Elena Giulotto, Bianca Waud, Natasha A. Hamilton, Tosso Leeb, Gabriella Lindgren, Leslie A. Lyons, Molly McCue, James N. MacLeod, Julia Metzger, James R. Mickelson, Barbara A. Murphy, Ludovic Orlando, Cecilia Penedo, Terje Raudsepp, Eric Strand, Teruaki Tozaki, Dagmar S. Trachsel, Brandon D. Velie, Claire M. Wade, Jakub Cieslak, Carrie J. Finno. Long-read RNA sequencing improves the annotation of the equine transcriptome.

Abstract

A high-quality reference genome assembly, a biobank of diverse equine tissues from the Functional Annotation of the Animal Genome (FAANG) initiative, and incorporation of long-read sequencing technologies have enabled efforts to build a comprehensive and tissue-specific equine transcriptome. The equine FAANG transcriptome reported here provides up to 45% improvement in transcriptome completeness across tissue types when compared to either RefSeq and/or Ensembl transcriptomes. This updated transcriptome also provides major improvements in the identification of alternatively spliced isoforms, novel noncoding genes, and annotation of 3' transcription termination site (TTS). The equine FAANG transcriptome will aid functional studies investigating important equine traits, while providing opportunities to identify allele-specific expression and differentially expressed genes across tissues.

Introduction

Equine genome assemblies [1,2] have provided vital resources for equine genetics research. However, it is evident that detailed annotation of the genome is necessary for further investigation of both simple and complex traits in horses. Current equine genome assemblies have annotations provided by both Ensembl [3,4] and NCBI [5,6] gene annotation pipelines. These annotations relied primarily on limited

available RNA-seq data, cross-species alignments, and computational predictions. The equine RefSeq annotation release 103 from NCBI contains 33,146 genes and pseudogenes, of which 21,129 are protein coding, and 77,102 transcripts, including 60,887 mRNAs and 16,215 non-coding RNAs [6]. This presents a total isoform-to-gene ratio of 2.3, or 2.8 if only coding genes are considered. Similarly, the equine Ensembl annotation release 105.3 contains 20,955 protein coding genes and 9,014 non-coding genes, with 59,087 transcripts, resulting in a transcript-to-gene ratio of 2.04. For comparison, the most recent GENCODE human gene annotation release (release 39, GRCh38.p13) includes 61,533 genes, of which 19,982 are protein coding, with an average isoform-to-gene ratio of 3.9, or 4.3 when only considering protein coding genes [7]. Furthermore, the human ENCODE projects determined that genes tend to express many isoforms simultaneously, but different dominant isoforms exist in different cell lines [8]. The data from humans, in comparison to the horse, supports there are potentially missing alternatively spliced isoforms not represented in the current equine gene annotation. In addition, the tissue-specific nature of isoform expression represented in the human literature underscores the need for a transcriptome with more complete isoform annotation.

Noncoding RNAs play an important role in many biological pathways [9–12]. With the rising use of noncoding RNA therapeutics [13], a more comprehensive noncoding RNA annotation for the horse genome will certainly be an asset to the equine research community. The Ensembl annotation for EquCab3 includes 9,014 noncoding genes, while the RefSeq annotation contains 8,893 noncoding genes. In comparison, the GENCODE human gene annotation release 39 includes 26,378 noncoding genes. A particular challenge with annotating noncoding RNAs comes from the fact that noncoding RNAs are usually less evolutionarily conserved [14] and present at very low levels [15,16]. Therefore, without deep sequencing of diverse tissue types, noncoding RNAs typically remain unannotated. Since the current equine gene annotation relies heavily on cross-species conservation, and a limited number of RNA-seq data that are publicly available, it is expected that a large number of noncoding RNAs are

currently unannotated. Indeed, a previous study identified 20,800 candidate long noncoding RNAs (lncRNAs) with low expression, low exon diversity, and low levels of sequence conservation [17].

To address these challenges, the equine Functional Annotation of Animal Genome (FAANG) project has collected over 80 tissue types and body fluids from 4 adult Thoroughbred horses (two females and two males) [18,19]. These horses underwent thorough clinical examinations and were selected as healthy references. The FAANG biobank has produced a diverse dataset describing various aspects of equine gene regulation [20–22]. Here, we report our efforts to build a comprehensive transcriptome for the horse genome using long-read sequence technologies across nine diverse tissues.

Results

Transcript Annotation

Full-length non-redundant transcripts were categorized based on their annotated splice junctions as compared to reference Ensembl transcripts [4,23] and the genomic overlap between the two, following the schematics introduced by Tardaguila et al [24]. Overall, isoforms with novel splice sites (categorized as novel not in catalog or NNC) account for over 40% of all Iso-seq transcripts identified (**Table 2.1**). The majority of novel genes (96.6%) identified in the Iso-seq transcriptome have only one isoform per gene, 59.5% of which are mono-exonic (**Fig 2.1A-D**). Compared to the Ensembl annotation, the Iso-seq transcriptome contains fewer short transcripts (<0.5 Kb, 172 or 0.3% in Iso-seq transcriptome; 2,392 or 4% in Ensembl transcriptome) but proportionally more long transcripts (>1.5 Kb, 49,523 or 97.0% of Iso-seq transcriptome; 52,896 or 89.5% of Ensembl transcriptome) (**Fig 2.1E**).

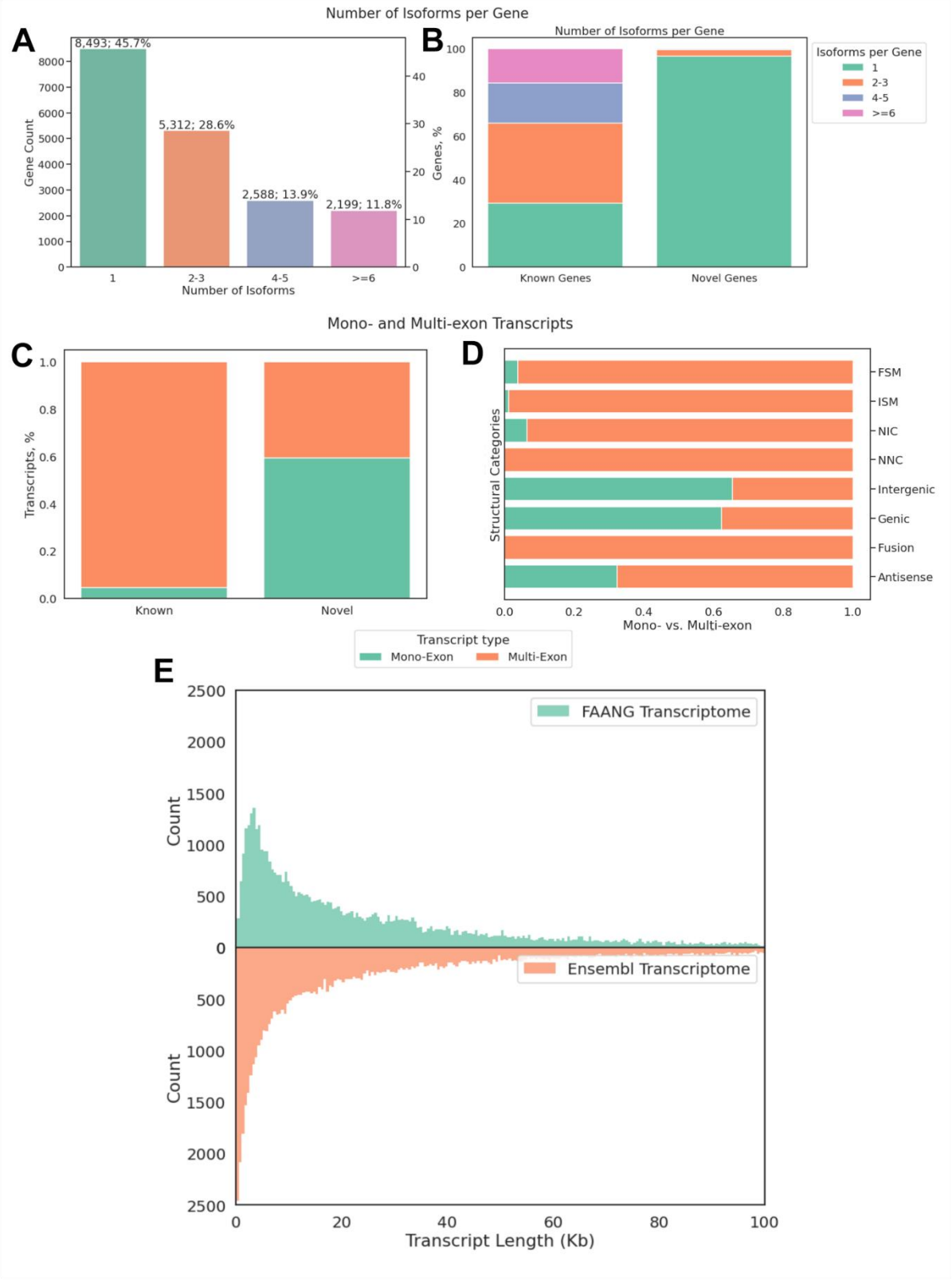
Number of Isoforms per Gene
A
8,493; 45.7%
5,312; 28.6%
2,588; 13.9%
2,199; 11.8%
Gene Count
Number of Isoforms
1
2-3
4-5
>=6
B
Number of Isoforms per Gene
Genes, %
Known Genes
Novel Genes
Isoforms per Gene
1
2-3
4-5
>=6
Mono- and Multi-exon Transcripts
C
Transcripts, %
Known
Novel
D
Structural Categories
FSM
ISM
NIC
NNC
Intergenic
Genic
Fusion
Antisense
Mono- vs. Multi-exon
Transcript type
Mono-Exon
Multi-Exon
E
FAANG Transcriptome
Ensembl Transcriptome
Count
Transcript Length (Kb)

Figure 2.1 Summary of the Iso-seq equine transcriptome.

The overall number of isoforms (A), known vs. novel genes (B), percentages of known and novel transcripts (C) and portion of mono- and multi-exon transcripts by structural categories that were annotated in the Iso-seq equine transcriptome. (D) Percentages of mono- and multi-exonic transcripts in each structural category; FSM: full-splice match, all exons and splice junctions match a known reference; ISM: incomplete-splice match, like FSM but missing 3' and/or 5' ends; NIC: novel-in-catalog, novel transcripts with known exons and splicing sites; NNC: novel-not-in-catalog, novel transcripts with novel splicing sites; Intergenic: novel transcripts with no overlapping known genes; Genic: novel transcripts overlapping known introns; Fusion: fusion transcripts; Antisense: novel transcripts on the opposite strand of known transcripts (E) Transcripts by length of Iso-seq transcriptome as compared to the Ensembl transcriptome.

Table 2.1 FAANG transcript breakdown by structural category

Structural Category	Count	Percentage
Antisense	928	1.64
FSM	12,470	22.01
Fusion	684	1.21
Genic	2,137	3.77
ISM	4,937	8.71
Intergenic	4,543	8.02
NIC	6,330	11.17
NNIC	24,634	43.47

FSM: full-splice match; ISM: incomplete-splice match; NIC: novel-in-catalog; NNC: novel-not-in-catalog

5' Completeness

Since standard Iso-seq libraries do not capture 5' caps of transcripts [25], an aggressive collapsing approach was utilized to remove potentially 5' degraded transcripts. To assess the completeness of 5' ends of annotated transcripts, short-read RNA-seq and ATAC-seq data from the same tissues were used to compare coverage near annotated TSS. Overall, 98.4% transcripts have higher short read RNA-seq coverage in the 100bp window downstream of the Iso-seq annotated TSS than upstream (**Fig 2.2A**).

Transcripts with a $\log_2$ ratio of greater than 1 were designated as 5' complete and 89.1% of transcripts overall were identified to be complete. A majority of transcripts (66.6%-93.7%) across all structural categories were determined to have complete 5' ends, with novel genes (genic (73.2%), intergenic (66.6%), and antisense (71.3%) having a lower percentage of 5' complete transcripts (**Fig 2.2B**). Additionally, ATAC-seq of the same tissues also show substantial enrichment at annotated TSS with $\log_2$ ratio of ATAC-seq coverage in 100bp immediately down- and up-stream of annotated TSS exceeding 2 in all nine tissues (**Fig 2.2C**).

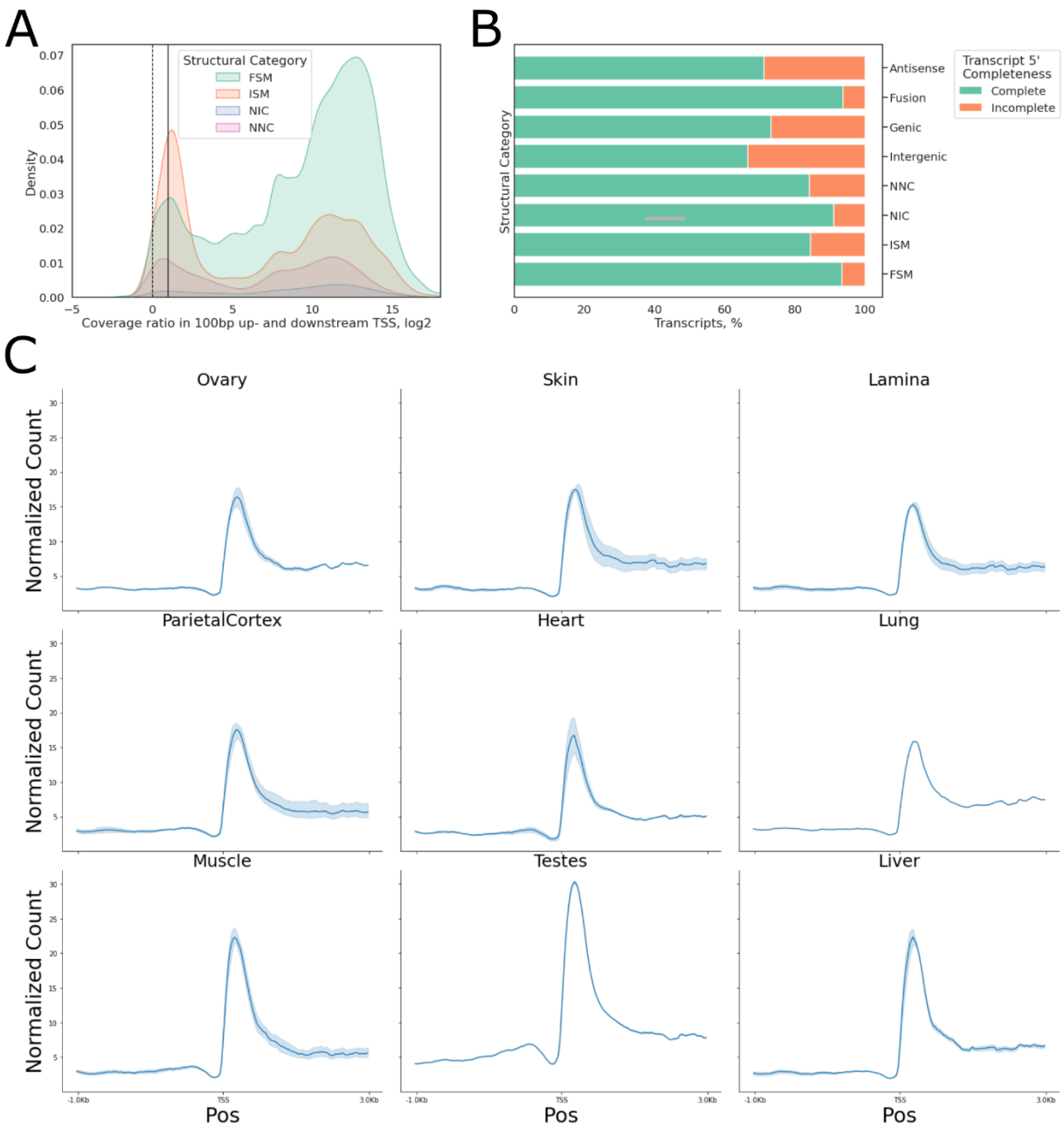

Figure 2.2 5' Completeness of the FAANG equine transcriptome.

(A) 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 (B) Percentages of transcripts whose log_2 ratios are greater than 1, denoted as 5' complete (green) (C) ATAC-seq read coverage in 1 kb upstream and 3 kb downstream of annotated TSS. FSM: full-splice match; ISM: incomplete-splice match; NIC: novel-in-catalog; NNC: novel-not-in-catalog; Schematics defined by Tardaguila et al [24].

3’ Completeness

To capture polyadenylated transcripts with complete 3’ ends, poly-T oligonucleotides were used during library construction of Iso-seq. However, internal stretches of adenines could also bind to poly-T oligonucleotides, a phenomenon known as intra-priming, which results in truncated transcripts [24]. Multi-exonic transcripts across all structural categories had approximately 25% adenines on average, with fewer than 5% transcripts having over 80% adenines, suggesting a high level of 3’ completeness (**Fig 2.3A**). Over 30% of the novel mono-exonic isoforms (NIC) are flagged as potentially intra-primed (**Fig 2.3B**). Many of these transcripts retain a partial intron and may be intron-retaining isoforms undergoing nonsense-mediated decay (NMD). A comparison between Iso-seq and Ensembl annotation showed significant improvement of TTS full-splice matched (FSM) transcripts, with over 4,232 transcripts having TTS more than 1 kb downstream of Ensembl annotated TTS, as well as minor improvements of TSS annotation with 5,103 transcripts having TSS at least 100 bp upstream of Ensembl annotated TSS (**Fig 2.3C**).

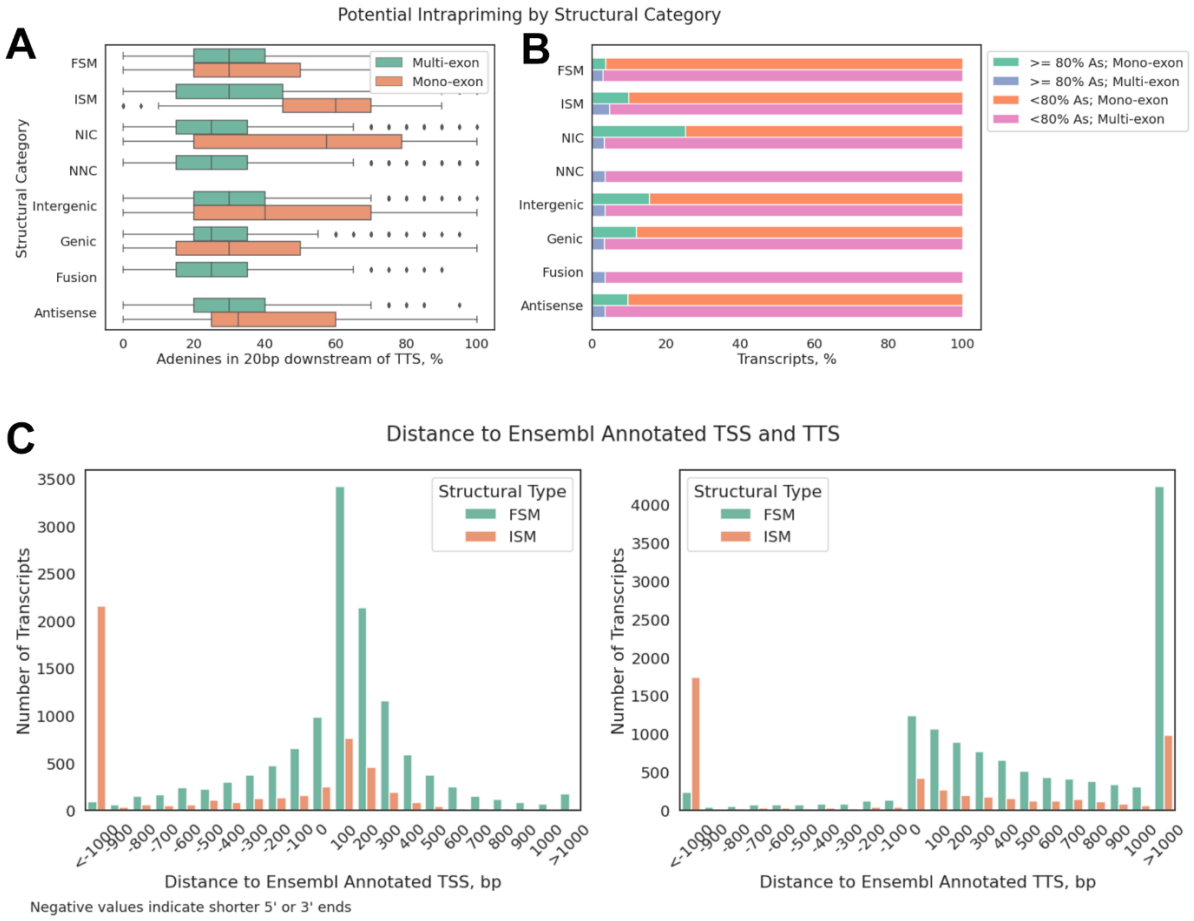

Figure 2.3 3' Completeness of the FAANG equine transcriptome.

(A) Percentages of adenines in 20 bp genomic regions immediately downstream of annotated TTS. Boxes indicate interquartile range (IQR) and whiskers indicate 1.5*IQR (B) Portions of transcripts with more than 80% adenines in 20bp genomic regions immediately downstream of annotated TTS, by structural categories and exon counts. (C) Distance between Iso-seq annotated TTS/TSS and Ensembl annotated TTS/TSS, negative values indicate shorter 5' or 3' ends. Eg. -1000 indicates that Iso-seq annotated TSS is 1000 bp downstream of Ensembl annotated TSS (left) or that Iso-seq annotated TTS is 1000 bp upstream of Ensembl annotated TTS (right). IQR: interquartile range, the range of second and third quartile of data.

Protein Coding and Noncoding Transcripts

Open reading frames (ORFs) were predicted using GeneMarkS-T (GMST) algorithm [26] by SQANTI3 [24] to identify protein-coding transcripts in the FAANG transcriptome. The vast majority of transcripts belonging to known genes had ORFs (97.6% of FSM, 96.8% of ISM, 92.5% of NIC, and 95.4% of NNC),

while a significant proportion of novel genes had transcripts without ORFs (28% of genic, 60% of intergenic, and 67.6% antisense transcripts) (**Fig 2.4A**). A substantial difference in exon counts among coding and noncoding transcripts was identified, with 44.6% of noncoding transcripts being mono-exonic as compared to 4.6% of coding transcripts. Specifically, coding transcripts with novel junctions (NIC) were 96.8% multi-exonic, while noncoding NIC transcripts are 56% multi-exonic. Similarly, 53% of coding transcripts that overlap or fall within annotated introns were identified as multi-exonic, while only 7.5% of those without an ORF were identified as multi-exonic (**Fig 2.4B**).

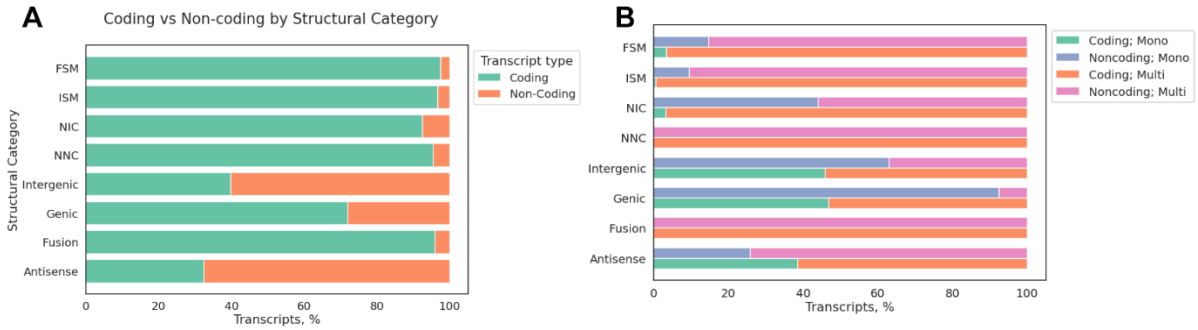

Figure 2.4 Protein coding and non-coding transcripts in the FAANG equine transcriptome.

(A) Portions of coding vs. noncoding transcripts by structural categories and (B) portions of coding vs. noncoding transcripts by structural categories and exon counts.

Splice Junctions

Any junctions not covered by at least one uniquely mapped read from RNA-seq data were removed, along with their associated transcripts. A total of 8,476 transcripts containing 14,738 such junctions were removed at this step. On average, known junctions had 4.8x RNA-seq coverage as compared to novel junctions. This difference primarily came from canonical junctions (GT-AG, GC-AG and AT-AC) (**Fig 2.5A**). Novel isoforms and transcripts of novel genes also had lower minimum junction coverage as compared to known isoforms (FSM and ISM, Kruskal-Wallis H-test, $p<0.0001$; post-hoc Dunn's test $p < 3.5 \times 10^{-68}$, Bonferroni corrected $\alpha=0.003$; **Fig 2.5B**).

Approximately 10% of the splice junctions annotated in Iso-seq transcriptome were novel as compared to the Ensembl transcriptome (56,503 out of 581,782). These novel splice junctions contributed to the discovery of 36,795 novel isoforms (**Fig 2.5C**). The GT-AG splice site was observed in 99.2% of splice junctions, with GC-AG and AT-AC sites observed in 0.68% and 0.05% of transcripts, respectively. Non-canonical splice sites were primarily observed at very low frequencies (<3%) (**Fig 2.5C**; **Table 2.2**).

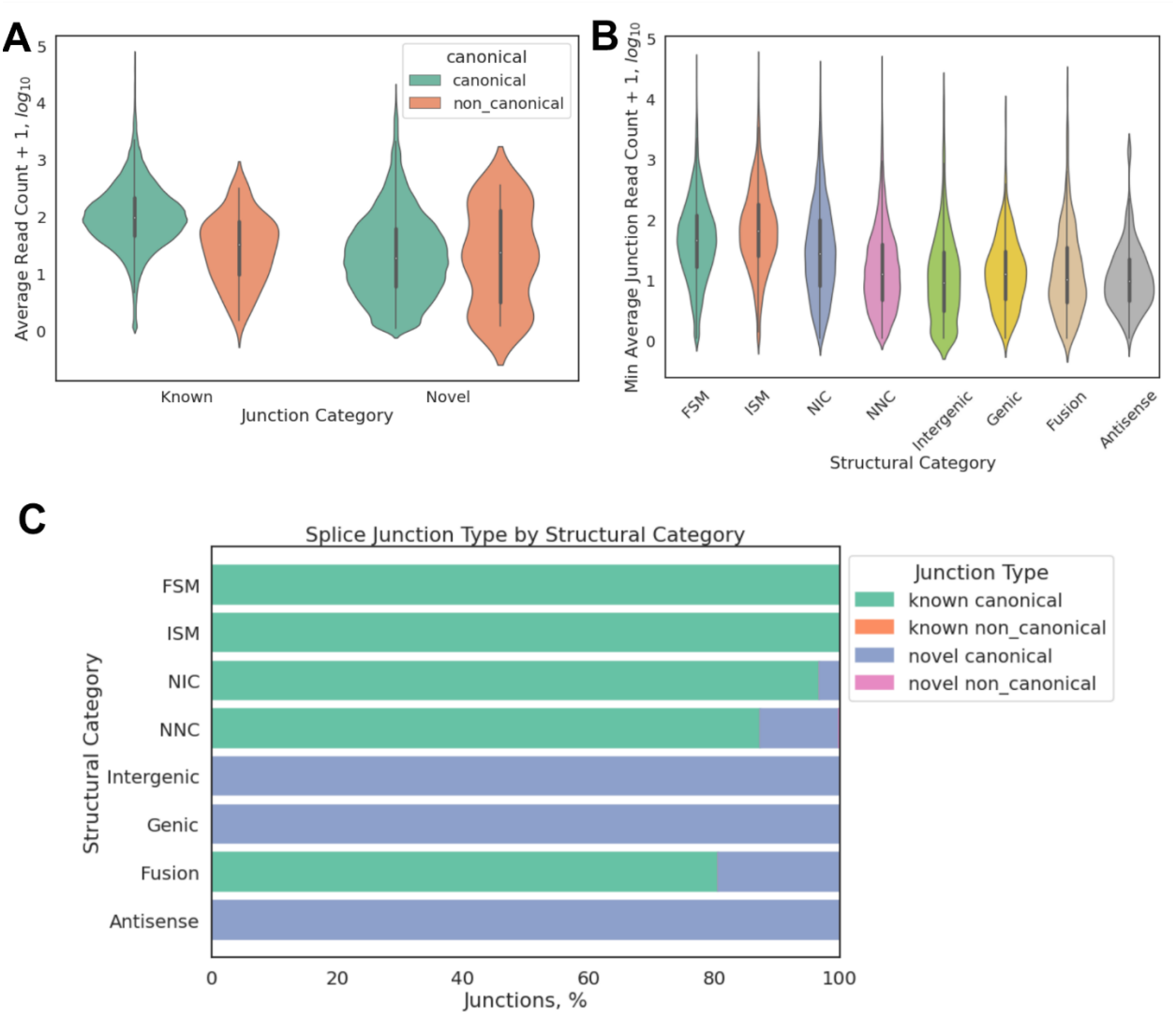

Figure 2.5 Splice junctions are better defined in the FAANG equine transcriptome.

(A) RNA-seq coverage measured as $\log_{10}$(ReadCount+1) at known or novel splice junctions, (B) minimum splice junction coverage transcripts by structural categories and (C) splice junction types by structural categories, known non-canonical junctions were observed in FSM, ISM, NIC, and NNC at 2.2%, 0.2%,

2.0%, and 1.2%, respectively; novel non-canonical junctions were only observed in NNC and intergenic isoforms at 2.8% and 2.5%, respectively.

Table 2.2 Portions of different splice junction types by structural categories

	% Known Canonical	% Known Non-canonical	% Novel Canonical	% Novel Non-canonical
Antisense	0.00	0.00	100.00	0.00
Fusion	80.57	0.00	19.43	0.00
Genic	0.00	0.00	100.00	0.00
Intergenic	0.00	0.00	99.97	0.03
NNC	87.35	0.01	12.61	0.03
NIC	96.81	0.02	3.17	0.00
ISM	100.00	0.00	0.00	0.00
FSM	99.98	0.02	0.00	0.00

FSM: full-splice match; ISM: incomplete-splice match; NIC: novel-in-catalog; NNC: novel-not-in-catalog

Sense-Antisense Transcripts

A total of 861 novel antisense transcripts were identified, with 2,742 isoforms annotated on the opposite strand. Overall, 3,246 transcripts on the plus strand that overlap at least 1 bp with a transcript on the minus strand were detected. Among these sense-antisense pairs of transcripts, 2,249 (69.3%) were coding-to-coding pairs, 954 (29.4%) coding-to-noncoding pairs, and 43 (1.3%) noncoding-to-noncoding pairs.

Tissue-specific Expression

Short-read RNA-seq data from 57 tissues (46 tissues from female animals and 23 tissues from male animals, with 12 tissues from both sexes, **Supplementary Table 2.1**) were used to quantify the Iso-seq

transcripts. 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 2.6A**). 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 (**Fig 2.6B**) and had different dominant major isoforms (isoform with highest relative expression of a given gene), depending on the tissue type (**Fig 2.6C-D**). Similar to humans, major isoforms in horses accounted for 30%-70% of the corresponding genes' total expression in any given tissue (**Fig 2.6E**). Notably, our RNA-seq data exhibited prominent clustering between sexes within the central nervous system (CNS) tissues (**Fig 2.7A**), while all other tissues clustered as expected regardless of sex. (**Fig 2.7B**).

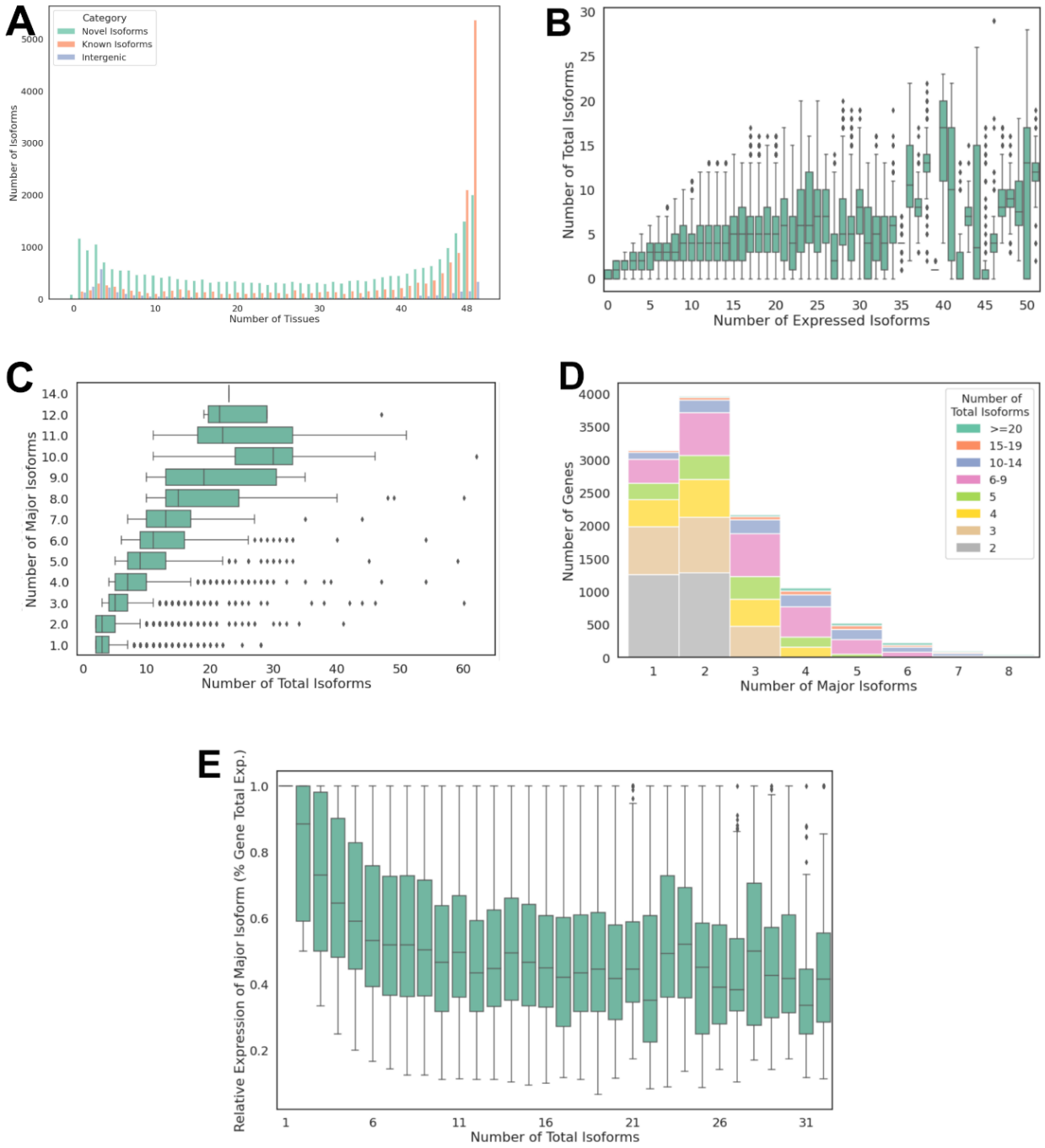

Figure 2.6 Short-read RNA-seq data mapped to the FAANG equine transcriptome identifies tissue-specific isoforms.

(A) Distribution of known vs. novel transcripts detected in different numbers of tissues, (B) number of expressed isoforms across all tissues vs. number of total isoforms per gene; boxes indicate IQR and whiskers indicate 1.5*IQR, (C) number of different major isoforms expressed across tissues vs. number of total isoforms annotated, (D) distribution of genes with different number of major isoforms and their total annotated isoforms and (E) relative expression of major isoforms in each tissue vs. total number of isoforms annotated. IQR: interquartile range, the range of second and third quartile of data.

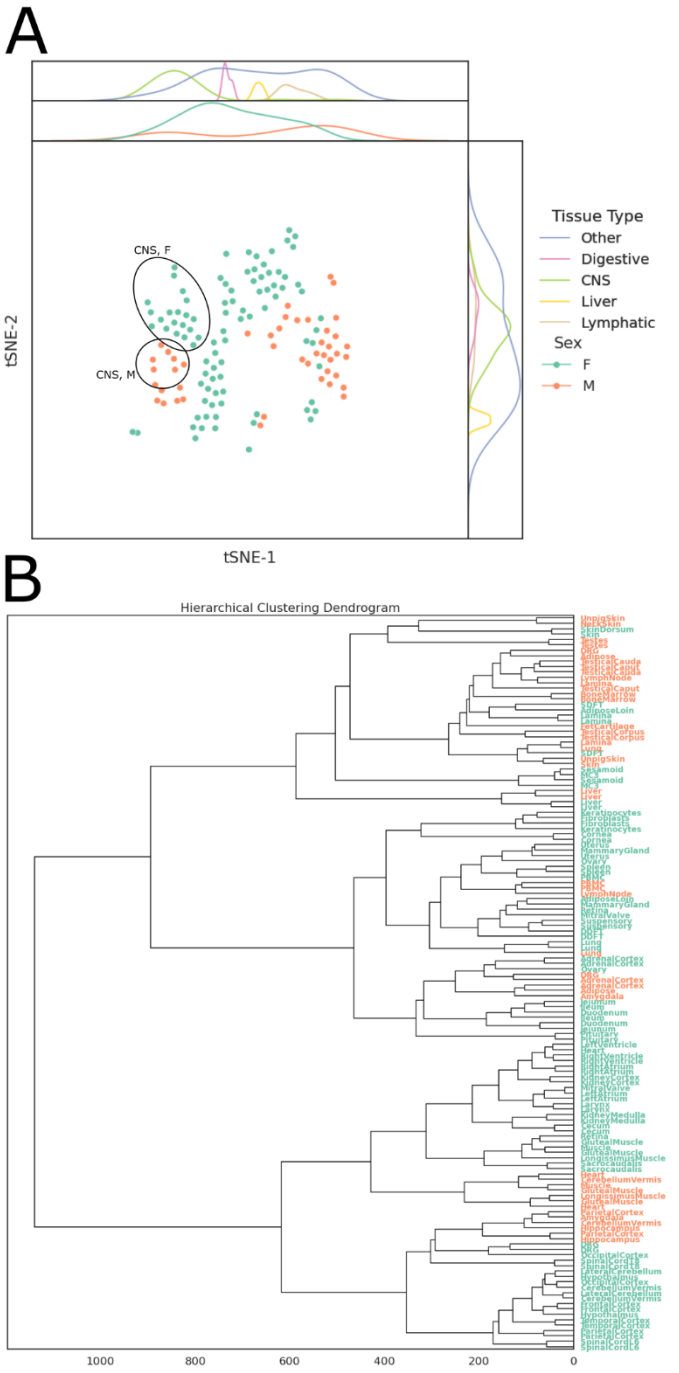

Figure 2.7 Sex-specific clustering of gene expression across tissue types.

(A) t-SNE plot of transcript levels (TPM) across samples and tissues and (B) agglomerative clustering of RNA-seq samples

Comparison with Ensembl and RefSeq Transcriptomes

To assess the completeness of the Iso-seq transcriptome, all available RNA-seq reads were aligned directly to each transcriptome and 3-23% improvement in numbers of properly paired reads in all tissues were observed, except for cerebellum vermis, duodenum, fibroblasts, keratinocytes, bone marrow, and epididymis (caput, corpus, and cauda).

To provide a comprehensive set of transcripts for the equine genome, we combined the Iso-seq transcriptome with Ensembl and RefSeq transcriptomes into a single annotation, termed the equine FAANG transcriptome. The FAANG transcriptome consisted of 153,492 transcripts (of which 128,723 were multi-exonic) from 36,239 genes, with a gene-to-isoform ratio of 4.2. This combined transcriptome contained a total of 26,631 coding genes, with 132,970 coding transcripts. RNA-seq alignments suggested an average 19.5% (8-45%) improvement in completeness compared to Ensembl and RefSeq transcriptomes across all sequenced tissues (**Fig 2.8**).

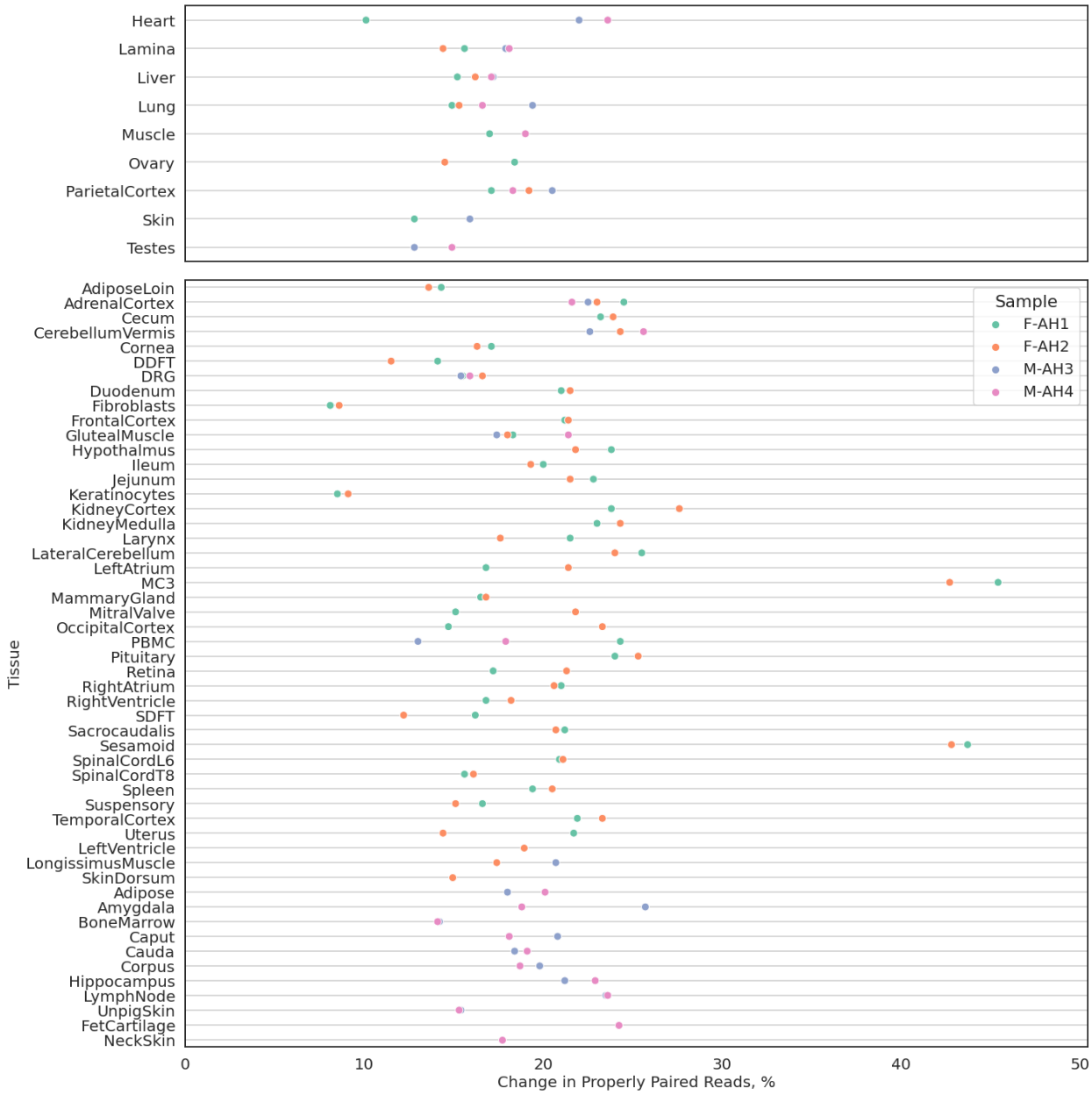

Figure 2.8 Comparison of FAANG, RefSeq and Ensembl equine transcriptomes.

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

Discussion

A comprehensive equine transcriptome with tissue-specific expression utilizing the rich tissue repository from the FAANG biobank and advanced long-read sequencing technology was developed. The FAANG

equine transcriptome consists of 36,239 unique genes with 153,492 transcripts, presenting a gene-isoform ratio of 4.2, or 5.0 when only protein-coding genes are considered. This is a substantial improvement as compared to the ratios of 2.0 from the Ensembl equine transcriptome or 2.8 from the RefSeq equine transcriptome and is more aligned to what is reported for the human genome [7]. We also demonstrated improved completeness of our transcriptome across over 40 tissues when assessed by companion short-read RNA-seq data, despite having only 9 tissue types included in our Iso-seq experiment.

The previous efforts to annotate the horse genome were limited by the number of tissue types available and sequencing lengths available at that time [27–29]. 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 [28]. 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) [29] and identified 36,876 genes with 76,125 isoforms. 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 [3,6]. 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 refine splice junction, TSS, and TTS annotation. As a result, we were able to identify 2,833 novel mono-exonic transcripts as well as improve TTS annotation for 11,098 known full-splice or incomplete-splice match transcripts.

In examining the transcriptional pattern of the horse genome, we revealed similar complexity in gene transcription to that of the human [7]. Specifically, we observed that genes with multiple isoforms tend to express more than one isoform simultaneously in any given tissue. The major isoforms (the isoform with highest expression) differed by tissue type. This aligns with the current understanding that isoforms, not genes, are directly associated with tissue-specific biofunctions and regulation occurs in a tissue specific manner. Most importantly, our data suggest that most known isoforms annotated in the Ensembl equine transcriptome are ubiquitous, while many novel isoforms identified in the Iso-seq transcriptome show tissue-specific expression, highlight the need for additional Iso-seq data across tissues. The addition of the novel isoforms identified here should aid the equine genetics community in advancing studies of complex traits.

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, a small portion of transcripts were still determined to be potentially 5' incomplete (**Figure 2.2A and B**). 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.

For the purpose of providing a comprehensive transcriptome, we focused on assessing the completeness of the FAANG equine transcriptome and overall complexities of tissue-specific transcription in the horse. However, the long read-length of Iso-seq data also provides unique opportunities for phasing exons. Therefore, future experiments will aim to use this data, coupled with whole genome sequencing and quantifiable RNA-seq of the same animals, to study allele-specific expression. Additionally, with the FAANG transcriptome and the large repository of RNA-seq data from a diverse set of tissues, future studies can focus on quantifying gene expression across tissues and conditions.

Methods and Materials

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 **Supplementary Materials 2.1 and 2.2**.

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 [18,19] 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 [30]. Circular consensus reads (ccs) were then constructed from demultiplexed subreads using PacBio Ccs program [31]. PolyA tails were trimmed from ccs reads using Isoseq3 [32]. 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 [32]. Clustered transcripts were aligned to the reference genome EquCab3 [2] using minimap2 [33] 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 [34] 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 [24] was then used to classify and annotate the transcriptome. Finally, the total transcriptome was filtered to remove nonsense-mediated decay transcripts, transcripts without short-read coverage support, and transcripts with a splice junction not covered by short-read RNA-seq data to generate the final FAANG equine transcriptome. To detect potential intra-primed transcripts, the percentage of adenines in a 20 bp window immediately downstream of the annotated transcription termination site (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 [35], matplotlib [36], seaborn [37], scipy [38], and scikit-learn [39].

RNA-seq analysis

Short-read RNA-seq data were trimmed to remove adapters and low-quality reads using trim-galore [40] and Cutadapt [41]. Read qualities were inspected using fastQC [42] and multiQC [43]. Trimmed reads were aligned to equCab3.0 using STAR aligner [44] with standard parameters (with --outSAMstrandField intronMotif --outSAMattrIHstart 0). PCR duplicates were marked using sambamba [45]. Mapping rates, qualities, and fragment lengths were assessed with samtools [46] and deeptools [47]. Aligned reads were used to assess completeness of transcriptomes using deeptools [47]. BWA MEM [48] was used to align the RNA-seq reads directly to transcriptomes and samtools [46] 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.

ATAC-seq analysis

ATAC-seq data from the 8 tissues (lamina, liver, left lung, left ventricle of heart, longissimus muscle, parietal cortex, testis, and ovary) collected from the same animals were generated and processed according to Peng et al. [22] Libraries were sequenced in 50 bp paired-end mode (PE50) on Illumina NovaSeq 6000. Aligned reads were used to quantify normalized read counts in 1Kb up- and down-stream of TSS sites annotated in the equine FAANG transcriptome.

Data Access

Short read 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 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.

Acknowledgements: The authors would like to acknowledge the four animals from which the FAANG tissues were collected.

Funding: This project was supported by the Grayson-Jockey Club Research Foundation, Animal Breeding and Functional Annotation of Genomes (A1201) Grant 2019-67015-29340 from the USDA National Institute of Food and Agriculture and the UC Davis Center for Equine Health with funds provided by the State of California pari-mutuel fund and contributions by private donors. Additional support for C.J.F. was provided by NIH L40 TR001136. Funding was also provided through a Priority Partnership Collaboration Award from the University of Sydney and University of California, Davis.

References

1. Wade, C. M. et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867 (2009).

2. Kalbfleisch, T. S. et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol 1, 197 (2018).
3. Aken, B. L. et al. The Ensembl gene annotation system. Database (Oxford) 2016, (2016).
4. Equus caballus Ensembl Annotation Release 105.3. Ensembl https://uswest.ensembl.org/Equus_caballus/Info/Annotation.
5. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733-745 (2016).
6. Equus caballus RefSeq Annotation Release 103. RefSeq https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Equus_caballus/103/.
7. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research 47, D766–D773 (2019).
8. Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).
9. Meller, V. H., Joshi, S. S. & Deshpande, N. Modulation of Chromatin by Noncoding RNA. Annu. Rev. Genet. 49, 673–695 (2015).
10. Shen, Y. et al. Insights into Enhancer RNAs: Biogenesis and Emerging Role in Brain Diseases. Neuroscientist 107385842110468 (2021) doi:10.1177/10738584211046889.
11. Moazzendizaji, S. et al. microRNAs: Small molecules with a large impact on colorectal cancer. Biotechnology and Applied Biochemistry bab.2255 (2021) doi:10.1002/bab.2255.
12. Wen, Z.-J. et al. Emerging roles of circRNAs in the pathological process of myocardial infarction. Molecular Therapy - Nucleic Acids S2162253121002456 (2021) doi:10.1016/j.omtn.2021.10.002.
13. Winkle, M., El-Daly, S. M., Fabbri, M. & Calin, G. A. Noncoding RNA therapeutics — challenges and potential solutions. Nat Rev Drug Discov 20, 629–651 (2021).
14. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22(9):1775–89 (2012).
15. Menet, J. S., Rodriguez, J., Abruzzi, K. C. & Rosbash, M. Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. eLife 1, e00011 (2012).
16. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. PLoS Comput Biol 5, e1000598 (2009).
17. Scott, E. Y. et al. Identification of long non-coding RNA in the horse transcriptome. BMC Genomics 18, 511 (2017).
18. Burns, E. N. et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim Genet 49, 564–570 (2018).
19. Donnelly, C. G. et al. Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. Front. Genet. 12, 650305 (2021).
20. Kingsley, N. B. et al. Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. Genes 11, 3 (2019).
21. Kingsley, N. B. et al. "Adopt-a-Tissue" Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare. Front. Genet. 12, 649959 (2021).
22. Peng, S., Bellone, R., Petersen, J. L., Kalbfleisch, T. S. & Finno, C. J. Successful ATAC-Seq From Snap-Frozen Equine Tissues. Front. Genet. 12, 641788 (2021).
23. Howe, K. L. et al. Ensembl 2021. Nucleic Acids Res 49, D884–D891 (2021).

24. Tardaguila, M. et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res (2018) doi:10.1101/gr.222976.117.
25. Gonzalez-Garay, M. L. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). in Transcriptomics and Gene Regulation (ed. Wu, J.) vol. 9 141–160 (Springer Netherlands, 2016).
26. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res 43, e78–e78 (2015).
27. Coleman, S. J. et al. Structural annotation of equine protein-coding genes determined by mRNA sequencing: Structural annotation of equine protein-coding genes. Animal Genetics 41, 121–130 (2010).
28. Hestand, M. S. et al. Annotation of the Protein Coding Regions of the Equine Genome. PLoS ONE 10, e0124375 (2015).
29. Mansour, T. A. et al. Tissue resolved, gene structure refined equine transcriptome. BMC Genomics 18, 103 (2017).
30. Pacific BioSciences. Lima. Version 2.0.0. 2020
31. Pacific BioSciences. Pbccs. Version 6.0.0. 2020
32. Pacific BioSciences. Isoseq3. Version 3.4.0. 2020
33. Li, H. New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574 (2021).
34. Tseng, E. cDNA_Cupcake. Version 28.0.0. 2021
35. Reback, J. et al. pandas-dev/pandas: Pandas 1.1.3. (Zenodo, 2020). doi:10.5281/ZENODO.3509134.
36. Caswell, T. A. et al. matplotlib/matplotlib v3.1.3. (Zenodo, 2020). doi:10.5281/ZENODO.3633844.
37. Waskom, M. seaborn: statistical data visualization. JOSS 6, 3021 (2021).
38. SciPy 1.0 Contributors et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020).
39. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
40. Krueger, F. Trim Galore! (2019).
41. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 17, 10 (2011).
42. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
43. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016).
44. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013).
45. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034 (2015).
46. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
47. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160–W165 (2016).
48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

Chapter 3 Tissue-specific annotation of open chromatin regions in the horse

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

Keywords: FAANG; horse; annotation; open chromatin; ATAC-seq

Abstract

High-quality genome and transcriptome assemblies are the beginning framework necessary to advance our understanding of the biology of the horse. In addition to these assemblies, annotation of the regulatory elements is needed to provide a comprehensive view of tissue specificity that will aid in answering complex biological questions. Active regulatory elements are typically characterized by chromatin accessibility, which allows transcription factors (TF) to access regulatory elements in the DNA, thereby controlling gene expression. Therefore, we assessed the chromatin accessibility, using the assay for transposase-accessible chromatins (ATAC-seq), in nine equine tissues (adipose, heart, lamina, liver, lung, ovary, testis, muscle, parietal cortex) from two healthy female and male horses. We annotated a total of 332,115 open chromatin regions across the nine tissues and correlated transcript abundance with transcription start site (TSS) accessibilities. The identified open regions were enriched in transcription start and termination sites (TSS/TTS), exons, and promoter regions and contained many known important TF binding sites, including CTCF and Sp/KLF family TFs. These data identified 190,815 open chromatin regions with tissue specific putative functional roles, paving the way for further hypothesis generation and testing of complex genetic diseases in the horse.

Introduction

Since the completion of the equine reference genome assemblies [1,2], steady progress has been made to identify tissue-specific transcription in the horse [3–8]. The Ensembl annotation for EquCab3.0 characterizes 59,087 transcripts, covering 46.5% of the genome (including introns, UTRs, and

pseudogenes) [6,9]. Similarly, the RefSeq annotation contains 77,102 transcripts, covering 46.0% of the genome [7,10]. Most recently, we compiled a multi-tissue comprehensive transcriptome containing 153,492 transcripts, covering 54.9% of the genome (see ***Chapter 2***) [8]. These findings align with many studies in other species, suggesting that much of the eukaryotic genome is transcribed, although many noncoding transcripts' functional roles remain unclear [11–16].

Nonetheless, many untranscribed regulatory elements, including enhancers and promoters, remain poorly characterized. Active regulatory elements are typically characterized by a lack of nucleosome binding and therefore, chromatin accessibilities are often used as a proxy for identifying active regulatory elements [17]. 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 [18]. Several efforts have been made to adapt the original ATAC-seq protocol to tissue [19] and cryopreserved nuclei [20] 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 [21, see *Addendum 1*].

To characterize tissue-specific transcription and regulation, the equine Functional Annotation of the Animal Genome (FAANG) initiative generated a biobank of over 80 tissues, cells, body fluids, and cryopreserved nuclei samples [22,23]. This rich repository of well-characterized tissue samples provides a unique opportunity to interrogate coordinated tissue-specific gene transcription and regulation. Indeed, RNA sequencing data from various tissues from this biobank has led to a much-improved transcriptome annotation for the equine genome [8]. Taking advantage of this valuable resource, we aimed to define tissue-specific open chromatin regions using cryopreserved nuclei of nine tissues (adipose, lung, liver, heart, longissimus muscle, cortex, lamina, ovary, and testis) in healthy female and male horses.

Results

Library Quality Assessment

Chromatin accessibility was profiled from nine equine tissues (adipose, heart, lamina, liver, lung, ovary, testis, muscle, cortex) of four biological replicates (female: AH1, AH2; male: AH3, AH4) (**Fig 3.1A**). Most libraries contained 60% to > 90% unique reads, with the exception of liver and cerebral cortex samples (**Fig 3.1B**). Data from the female liver samples were generated from our previous study, where excessive mitochondria contamination led to lower library complexities and resequencing was performed to reach desired unique read counts [21]. After removing polymerase chain reaction (PCR) duplicate reads, all libraries contained less than 20% of mitochondria reads (**Fig 3.1C**). Despite lower library complexities, both liver and cerebral cortex samples showed clear nucleosomal periodicities and high enrichment around transcription start sites (TSS) (**Fig 3.1D**). On the other hand, while few PCR duplicates were present in lamina libraries, no nucleosomal periodicities or substantial TSS enrichment were apparent in these libraries (**Supp. Fig 3.1**). These data suggested that the lamina libraries, while having high complexities, had very high background noise. However, deep sequencing of the female lamina samples improved library enrichment. Testis libraries also showed low TSS enrichment while having high complexities and apparent nucleosomal periodicities (**Supp. Fig 3.1**). This was speculated to be a result of high proportions of sperm cells in the testis samples. When restricted to a set of genes highly expressed in spermatozoa [25], TSS enrichment scores were significantly higher (AH3: 5.7, AH4: 4.4).

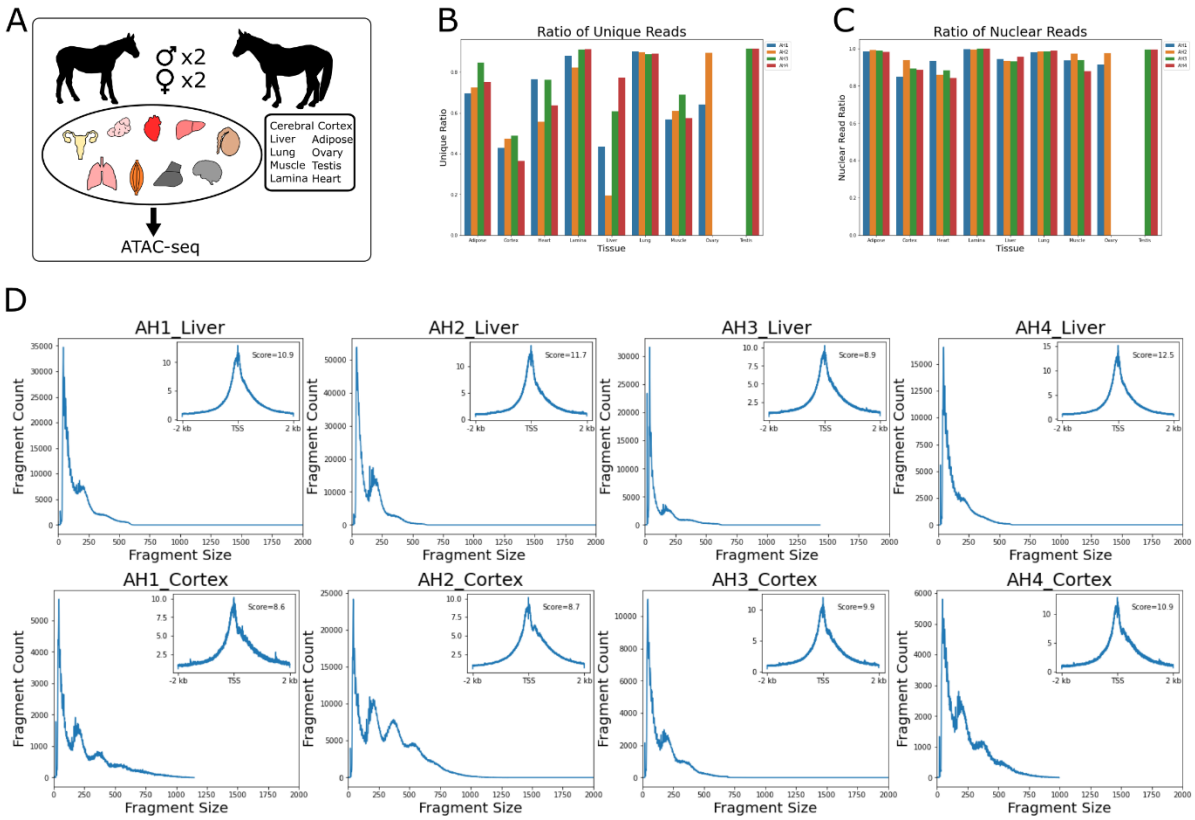

Figure 3.1 Experimental design and quality control.

(A) The overall experimental design for this project; (B) Ratio of unique reads across nine tissues; (C) Ratio of reads from nuclear genome after removing PCR duplicates; (D) Fragment size distributions and TSS enrichment plots (upper right inset plots) of liver and cerebral cortex samples.

Peak Calling

Since we were particularly interested in nucleosome free regions accessible to transcription factors, both ends of the aligned fragments were converted to BED format and peaks were called using MACS3 [26] single-end BED mode, with each read extended to 150 bp, centered around Tn5 transposition sites (5’ and 3’ ends of forward and reverse reads, respectively). After peak calling, peaks were merged iteratively, first within each replicate, then between two replicates of each sex, and eventually between two sexes of each tissue, except ovary and testis, following the paradigm proposed by Grandi et al. [27] Accuracies of the peak calling were assessed using published histone ChIP-seq data from the same tissues [28]. Briefly, open chromatin peaks were intersected with “true positive” (H3K4me1 or H3K4me3

peaks overlapping H3K27ac) and “true negative” (H3K27me3 peaks) peak sets to calculate true positive rates (TPR) and false positive rates (FPR). Since testes were not included in the histone ChIP-seq dataset from Kingsley et al. [28], testes’ libraries were not evaluated at this step. Area under curve (AUC) values of at least 0.6 were achieved for all tissues evaluated **(Fig 3.2A**, **Supp. Fig 3.2**, **Table 3.1**). 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 3.1**). To apply a filter with a similar retention rate for testis, we set a cutoff score of 93 (85th percentile) for testis, which resulted in 78,164 remaining peaks (**Table 3.1**). The filtered peak sets were then merged following the same iterative procedure, resulting in a union set of 332,115 non-overlapping peaks. To examine the distributions of these peaks across tissues, the union peak set was intersected with each tissue peak set. 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 ubiquitous across examined tissues. (**Fig 3.2B and C**)

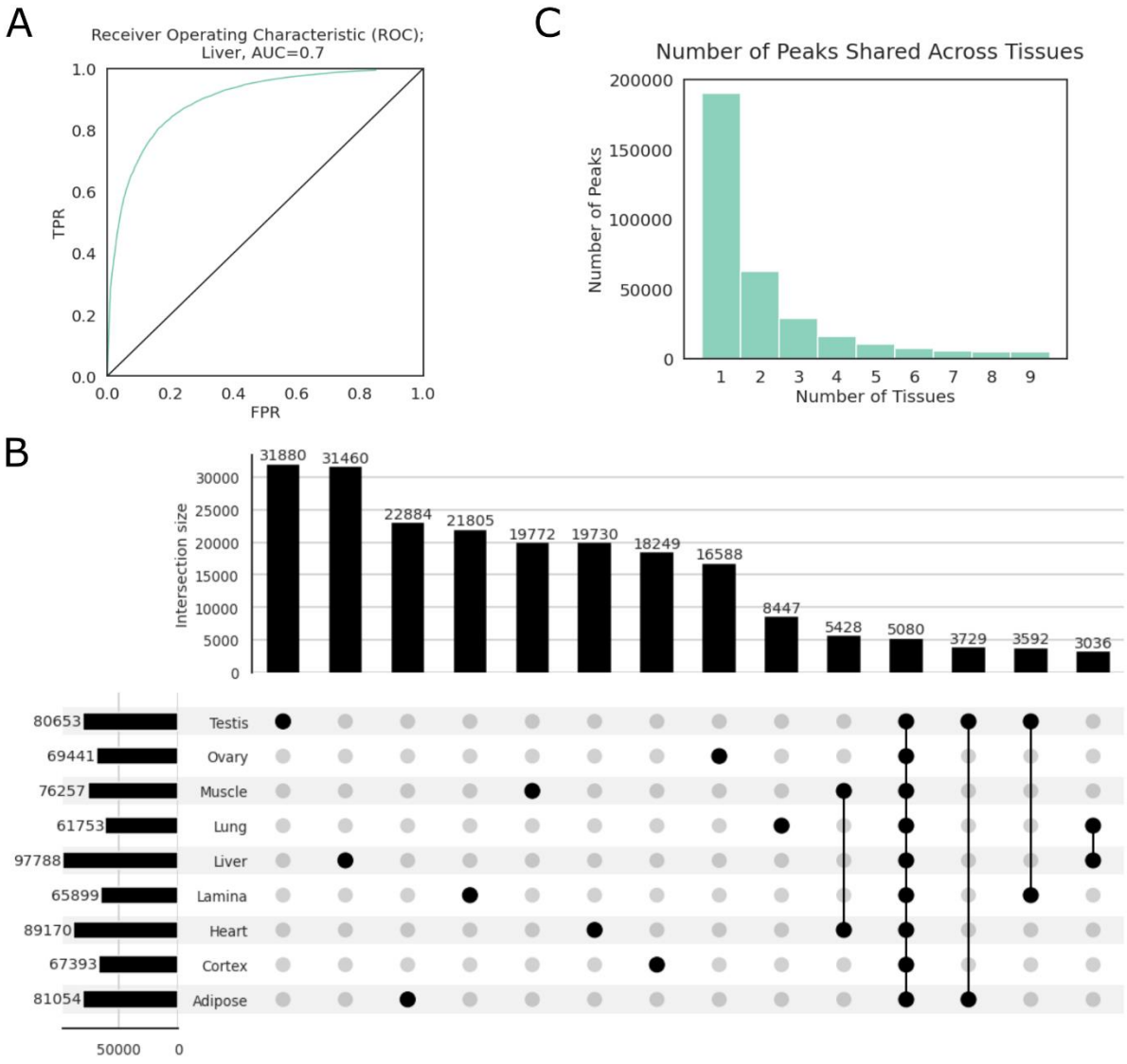

Figure 3.2 Peak metrics.

(A) ROC plot of liver peaks, as a representative example across tissues. TPR: true positive rate, FPR: false positive rate, AUC: area under curve; (B) Intersections of peaks between tissues, bottom right: top 14 intersections, black dots indicate that peaks in this intersection are found in a particular tissue, bottom left: histogram of total peaks in each tissue, top: histogram showing number of peaks in each intersection; (C) Number of peaks that were observed in a given number of tissues

Table 3.1 Peak metrics								
	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							
Ubiquitous	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								

Tissue Correlation

Using the union peak set, a count matrix of transposition events was constructed for all samples. This matrix was used to examine the pair-wise correlations between all samples (**Fig 3.3A**, **Supp Fig 3.3**). Replicates within most tissues were highly correlated, except lamina, consistent with previous observations of library qualities. Additionally, testis showed substantial differences from the rest of the tissues, likely due to a high proportion of spermatozoa with low transcriptional activities [29] in testis samples. Principle component analysis (PCA) revealed substantial tissue-specific structures, with the first principal component accounting for 25.2% of the variance in the count matrix (**Fig 3.3B**).

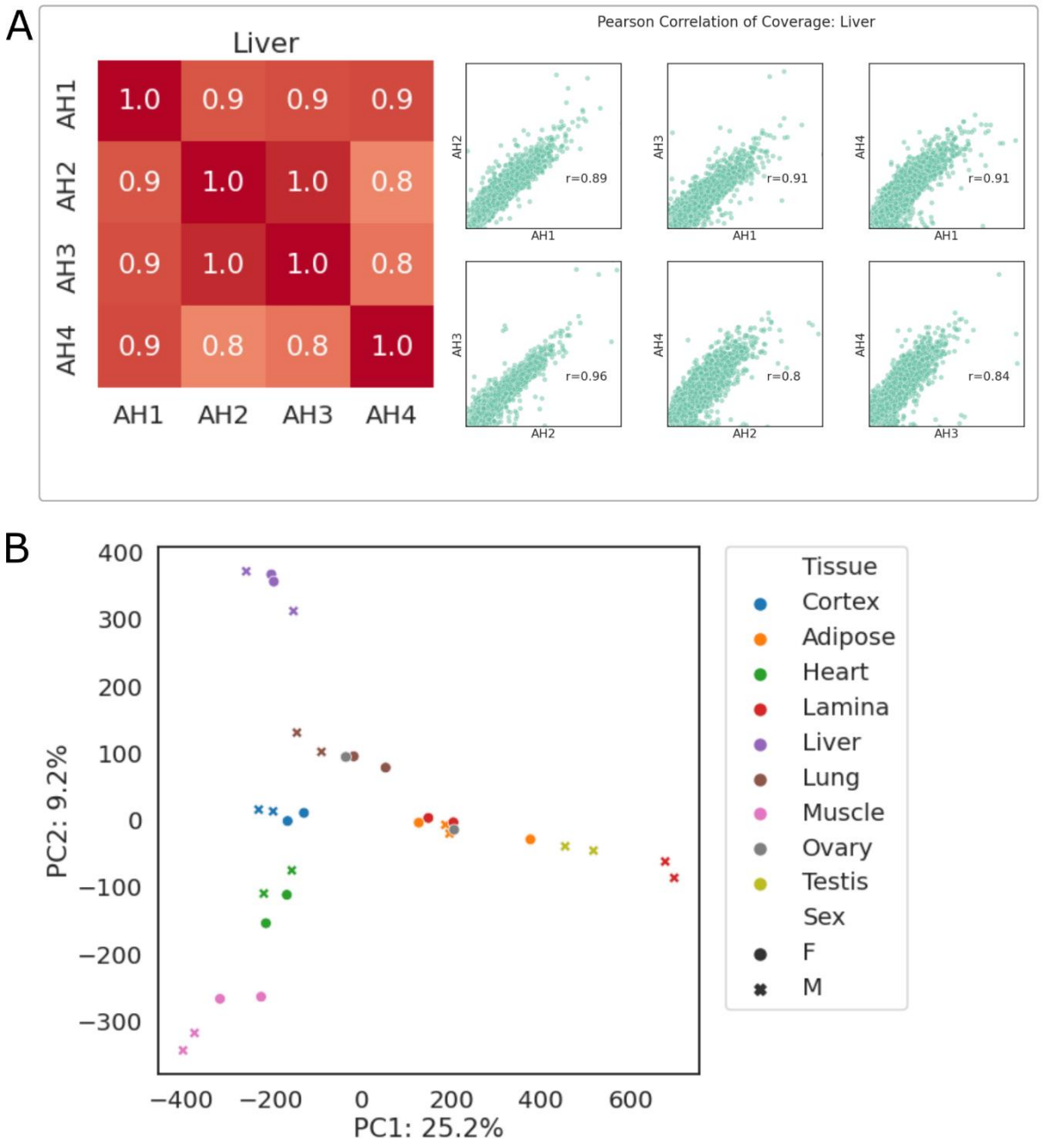

Figure 3.3 Cross- and within-tissue correlation.

(A) Heatmap and scatter plots of Pearson correlations among biological replicates of liver; (B) Principal component analysis (PCA) of transposition counts across all peaks

Peak Annotation

To explore potential functional roles of open chromatin regions, we annotated each peak by its proximity to annotated gene features [8]. The total union peaks were enriched in TSS, transcription termination sites (TTS), and exon regions (**Table 3.2**, **Supp Table 3.1**). This enrichment was more apparent in ubiquitous peaks (**Fig 3.4**). Gene ontology (GO) terms overrepresented in genes associated with these ubiquitous peaks were all essential housekeeping biological processes such as TOR signaling and kinase activity (**Supp Table 3.2**). 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 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. (**Supp Table 3.3**).

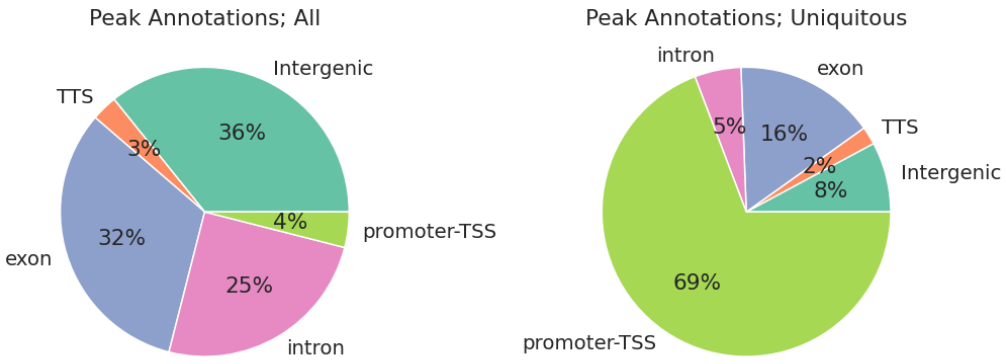

Figure 3.4 Peak annotations

Left: Composition of peaks by annotation in union peaks; Right: Composition of peaks by annotation in ubiquitous peaks

Table 3.2 Peak annotation		
	Union	
Annotation	**Number of peaks**	**Log_2 Enrichment**
TTS	10,602	0.39
Exon	116,529	0.38
Intron	74,505	-0.23
Intergenic	103,408	-0.41
Promoter-TSS	26,282	1.54
TTS: transcription termination site; promoter-TSS: 3kb up- and down-stream of annotated transcription start site (TSS)		

Differential Accessibility Analyses

In addition to tissue-specific peaks, we selected a direct comparison of cerebral cortex and heart, two neighboring tissues on the PCA plot (**Fig 3.3B**) with clear separation, to demonstrate the differentially accessible regions (DAR) and its correlation with differential gene expression. Overall, approximately 16% of peaks showed differential accessibility (FDR adjusted $p<0.05$), with 25,144 peaks (7.6%) more accessible in cerebral cortex and 29,206 peaks (8.8%) more accessible in heart (**Fig 3.5A**). To compare DAR with differentially expressed genes (DEG) between cerebral cortex and heart, peaks annotated as "promoter-TSS" (2 kb up- or down-stream of a TSS) were associated with their corresponding genes. The log_2 fold-change (log_2FC) of DAR was significantly correlated with log_2FC of DEG in the same cortex and heart samples (one-sided Wald test, $p<1 \times 10^{-5}$, Pearson correlation coefficient r=0.4, **Fig 3.5B**). 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 3.5C**). 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 (**Supp Tables 3.3, 3.4**). There were also 175 and 177 genes in Q2 and Q4, respectively. GO Terms related to synaptic activity were enriched in Q2 (**Supp Table 3.5**) while Q4 genes were overrepresented in actin-filament based processes.

Figure 3.5 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

Discussion

Using ATAC-seq data from nine equine tissues, we identified a total of 332,115 regions with open chromatin genome wide, with 59,024- 95,048 peaks identified in each tissue. Not surprisingly, open chromatin regions were enriched around important transcription sites, with promoter-TSS regions having $\log_2$ enrichment between 1.54 and 2.9, with the noted exception of lamina (**Supp. Table 3.1**). Open regions found in all nine tissues were even more highly enriched in transcription-related sites, with 69% of those located in promoter-TSS regions (**Fig 3.4**). This high level of TSS enrichment suggested functional elements related to essential biochemical functions that were highly accessible across all tissues that we assayed. Indeed, GO enrichment analyses revealed that genes associated with these ubiquitous regions were involved primarily in housekeeping functions.

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. [30], 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 [31], 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. 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. It is likely that many of these regions have important regulatory functions but, since little is known about the annotation of enhancers and CTCF binding sites in the horse genome, more work is needed to better characterize these specific peaks and the genes they regulate. 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 [32]. 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 [33].

In conclusion, we identified 332,115 accessible regions genome wide across nine tissues. These regions were shown to harbor both tissue-specific and housekeeping regulatory elements and therefore, should be of interest in studies of complex traits in the horse. To further corroborate the functional roles of

these important open chromatin regions, ChIP-seq data of important histone modifications and TF binding sites can be used to annotate many intergenic and intronic peaks, which is an important next step for the equine FAANG project.

Methods and Materials

RNA-seq analysis

Short-read mRNA-seq reads were trimmed to remove adapters and low-quality reads using trim-galore [34] and Cutadapt [35]. Read qualities were inspected using fastQC [36] and multiQC [37]. These reads were quantified against the equine FAANG transcriptome using salmon [38].

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 [22,23] were generated according to Peng et al. [21] Libraries were sequenced in 50 bp paired-end mode (PE50) on Illumina NovaSeq 6000. Reads were aligned to EquCab3.0 [2] 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 [39]. Remaining reads were shifted +4/-5 bp on plus/minus strand, respectively, to account for the 9 bp insertion introduced by Tn5 transposase [18] using deepTools [40]. Both forward and reverse reads of the final fragments were converted to bed format using bedtools [41] and peaks were called and refined using MACS3 [26,42] (-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 [27]. 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 [24]. 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 "ubiquitous".

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 [28]:

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 [43] (-size 250) with custom genome built from EquCab3.0 assembly [2] and FAANG transcriptome annotation [8]. GO enrichment analyses were performed using PANTHER [44] with default parameters.

Data Access

RNA-seq data can be accessed from ENA and SRA under the accession number PRJEB26787. ATAC-seq data can be accessed from ENA and SRA under the accession number PRJEB53037. Histone modification peaks can be accessed from FAANG data portal (https://data.faang.org/) under the accession number PRJEB35307

Funding: This project was supported by the Grayson-Jockey Club Research Foundation, Animal Breeding and Functional Annotation of Genomes (A1201) Grant 2019-67015-29340 from the USDA National Institute of Food and Agriculture and the UC Davis Center for Equine Health with funds provided by the State of California pari-mutuel fund and contributions by private donors. Additional support for C.J.F. was provided by NIH L40 TR001136.

Supplementary Figure 3.1 Library Qualities; fragment size distributions and TSS enrichment plots (upper right inset plots) of all tissues

Supplementary Figure 3.2 Peak Qualities; ROC plots of all tissues

Supplementary Figure 3.3 Cross- and within-tissue correlation; heatmap of Pearson correlations among all tissues and their biological replicates

References

1. Wade, C. M. et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867 (2009).
2. Kalbfleisch, T. S. et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol 1, 197 (2018).

3. Coleman, S. J. et al. Structural annotation of equine protein-coding genes determined by mRNA sequencing: Structural annotation of equine protein-coding genes. Animal Genetics 41, 121–130 (2010).
4. Hestand, M. S. et al. Annotation of the Protein Coding Regions of the Equine Genome. PLoS ONE 10, e0124375 (2015).
5. Mansour, T. A. et al. Tissue resolved, gene structure refined equine transcriptome. BMC Genomics 18, 103 (2017).
6. Equus caballus Ensembl Annotation Release 105.3. Ensembl https://uswest.ensembl.org/Equus_caballus/Info/Annotation.
7. Equus caballus RefSeq Annotation Release 103. RefSeq https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Equus_caballus/103/.
8. Peng, S. et al. Long-read RNA Sequencing Improves the Annotation of the Equine Transcriptome. 2022.06.07.495038 (2022) doi:10.1101/2022.06.07.495038.
9. Howe, K. L. et al. Ensembl 2021. Nucleic Acids Res 49, D884–D891 (2021).
10. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44, D733-745 (2016).
11. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447, 799–816 (2007).
12. Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).
13. Eddy, S. R. The C-value paradox, junk DNA and ENCODE. Current Biology 22, R898–R899 (2012).
14. Qu, H. & Fang, X. A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. Genomics, Proteomics & Bioinformatics 11, 135–141 (2013).
15. Elliott, T. A., Linquist, S. & Gregory, T. R. Conceptual and Empirical Challenges of Ascribing Functions to Transposable Elements. The American Naturalist 184, 14–24 (2014).
16. Palazzo, A. F. & Lee, E. S. Non-coding RNA: what is functional and what is junk? Front. Genet. 6, (2015).
17. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10, 161–172 (2009).
18. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC‐seq: A Method for Assaying Chromatin Accessibility Genome‐Wide. Current Protocols in Molecular Biology 109, (2015).
19. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods 14, 959–962 (2017).
20. 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).
21. Peng, S., Bellone, R., Petersen, J. L., Kalbfleisch, T. S. & Finno, C. J. Successful ATAC-Seq From Snap-Frozen Equine Tissues. Front. Genet. 12, 641788 (2021).
22. Burns, E. N. et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim Genet 49, 564–570 (2018).
23. Donnelly, C. G. et al. Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. Front. Genet. 12, 650305 (2021).
24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014).
25. Tomoiaga, D. et al. Single-cell sperm transcriptomes and variants from fathers of children with and without autism spectrum disorder. npj Genom. Med. 5, 14 (2020).

26. Liu, T. MACS: Model-based Analysis for ChIP-Seq. (2022).
27. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. Nat Protoc (2022) doi:10.1038/s41596-022-00692-9.
28. Kingsley, N. B. et al. Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. Genes 11, 3 (2019).
29. Ren, X., Chen, X., Wang, Z. & Wang, D. Is transcription in sperm stationary or dynamic? Journal of Reproduction and Development 63, 439–443 (2017).
30. Halstead, M. M. et al. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. BMC Genomics 21, 698 (2020).
31. Ko, J. Y., Oh, S. & Yoo, K. H. Functional Enhancers As Master Regulators of Tissue-Specific Gene Regulation and Cancer Development. Mol Cells 40, 169–177 (2017).
32. Xia, H., Dufour, C. R. & Giguère, V. ERRα as a Bridge Between Transcription and Function: Role in Liver Metabolism and Disease. Front. Endocrinol. 10, 206 (2019).
33. Stevanovic, M. et al. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. Front. Mol. Neurosci. 14, 654031 (2021).
34. Krueger, TrimGalore. Version 0.6.5 (2020).
35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 17, 10 (2011).
36. Andrews, S. FastQC: a quality control tool for high throughput sequence data. Version 0.11.8 (2010).
37. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048 (2016).
38. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14, 417–419 (2017).
39. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
40. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160–W165 (2016).
41. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
42. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008).
43. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–589 (2010).
44. Mi, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Research 49, D394–D403 (2021).

Chapter 4 Annotating tissue-specific chromatin states in the horse

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

Keywords: FAANG; horse; chromatin state; annotation; ChromHMM; enhancer

Abstract

Regulatory elements (REs) are segments of DNA that drive tissue- and developmental stage-specific gene expression. Identifying and annotating these REs is an essential step towards understanding the gene regulation network underlying complex phenotypes in the horse. Relying on the rich epigenomic dataset generated by the Functional Annotation of Animal Genome (FAANG) project, we catalogued several types of REs across nine tissues, including promoters, enhancers, and insulators. Through leveraging the FAANG transcriptome data, correlations of 84,613 REs were identified across tissues based on target gene expression. This dataset, along with the FANNG transcriptome assembly and open chromatin annotation from previous studies, are now available as UCSC Genome Browser tracks for the equine community.

Introduction

In eukaryote genomes, DNA is organized in a three-dimensional structure, where nucleosomes are dynamically unpacked in actively transcribed or regulatory regions [1–4]. 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 [5–7]. Therefore, annotating the genome by specifically defining these regulatory elements (REs) can provide significant context to understanding genetic variations contributing to many important traits in the horse.

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 [8–10]. 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 [11,12] with known downstream effectors that further regulate gene expression [13–15]. Additionally, H3K27ac is enriched around active elements and associated with higher levels of gene expression [16]. On the other hand, H3K27me3 is usually found around genes that are not active [17].

Since these four histone modifications denote important features of transcription regulation, the equine Functional Annotation of Animal Genome (FAANG) group assayed these marks using chromatin immunoprecipitation with sequencing (ChIP-seq) in nine tissue types from four animals, as previously reported [18,19]. Kingsley et al. [20] and Barber [21] reported tissue specific, as well as conserved putative regulatory sites, in these tissues across sexes as the data was available. However, the interactions between transcription factors and histone modifications are complex and multifaceted and combining these data set with others can advance the understanding of RE. For example, colocalization of H3K4me3 and H3K27me3 at bivalent promoters have been correlated with lineage differentiation [22], while colocalization of H3K4me1 and H3K27ac separate active enhancers from poised ones [17]. Therefore, defining these colocalized signals is necessary to better understand tissue specificity of active REs verses those that are poised.

In addition to histone modifications, three-dimensional structures of chromosomes also play an important role in gene regulation. CCCTC-binding factor (CTCF) is a main architectural protein that is integral to the formation of chromatin loops, which enables cis-regulatory interactions between DNA elements [23]. Given the complex combinatorial effects of various chromatin regulators on gene expression, various unsupervised methods [24,25] have been developed to interrogate the unique

regulatory states of genomic loci based on its epigenetic profiles, often termed chromatin states. In this study, we took advantage of a rich dataset of epigenetic modifications in the horse genome across nine tissue types (adipose, parietal cortex of brain, heart, lamina, liver, lung, muscle, ovary, testis) from the FAANG biobank. Here we report genome-wide, tissue-specific annotation of the horse genome and examine the correlations among each chromatin state's epigenetic profile, chromatin accessibility, and underlying transcription.

Results

Chromatin state discovery

Chromatin states were identified using four major histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K27me3) as well as CTCF binding. 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 4.1A**). Notably, the CTCF-bound active transcription start site (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 4.1B**). 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 4.1B, C**).

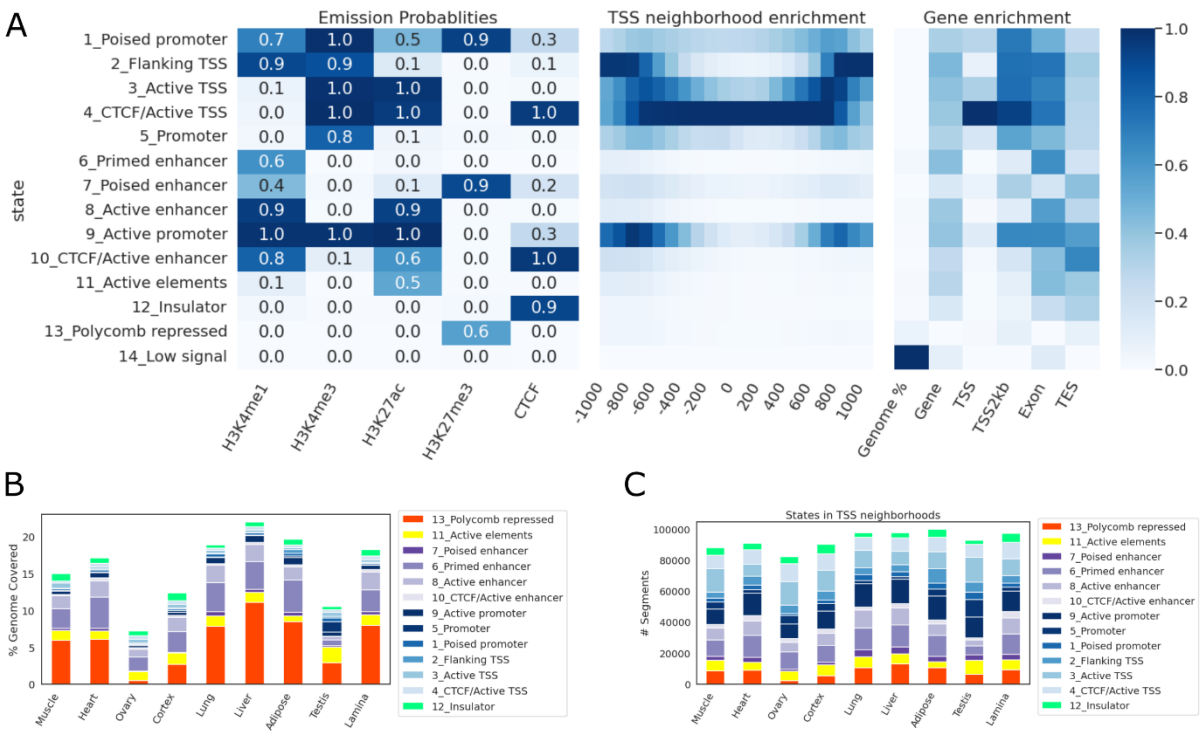

Figure 4.1 Chromatin states

(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

Tissue specificity

To examine the tissue specificity of each state, each state's segments from different tissues were merged into a union set of segments and intersected with each tissue. Overall, the majority of segments in each state were found in only one tissue type, with the exception of CTCF bound active TSS state (CTCF/Active TSS), where only a simple plurality of segments were tissue-specific (**Fig 4.2**). This state also harbored the greatest number of common segments across all nine tissues examined, followed by insulator and active promoter states. Additionally, promoter states demonstrated higher levels of ubiquity than enhancer states (**Fig 4.2**).

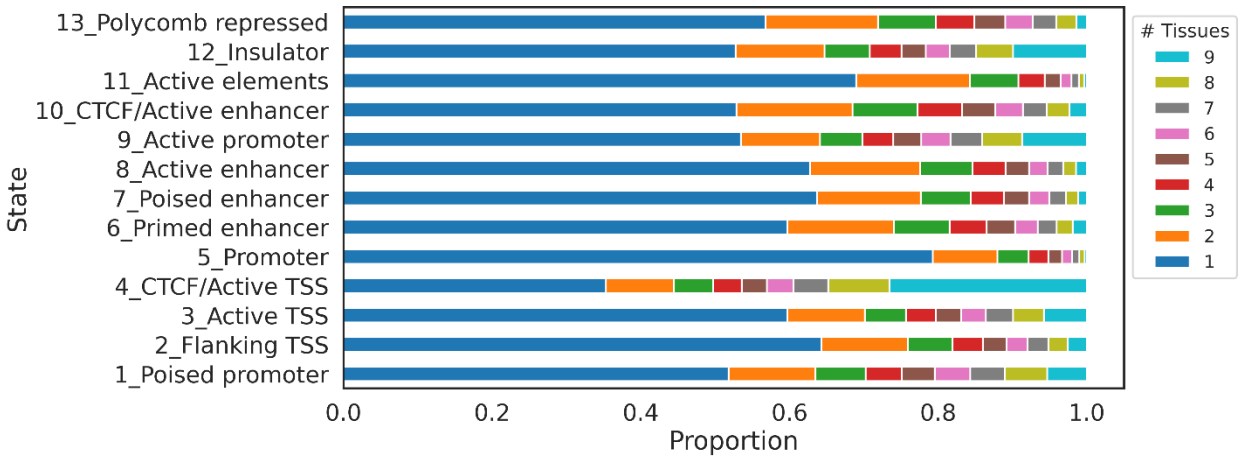

Figure 4.2 Tissue-specificity of states

The proportion of segments from each state that were identified in different numbers of tissues

Chromatin states predict gene expression

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 (***Chapter 2***) via quasi-mapping [26]. Transcript level quantification was then summarized to gene level using tximport [27]. For each tissue, genes were classified as high- or low-expression based on their aggregated transcripts per million (TPM) values (high: TPM $\geqslant$ 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 4.3**). CTCF bound active TSS state (state *4*) and showed a 59.4-fold enrichment around TSS of highly expressed (TPM $\geqslant$ 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 (**Fig 4.2**), 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) (**Fig 4.4**). 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 (**Supplementary Fig 4.1**). While less pronounced, it also accounted for 44% of CTCF-less active TSS state (state *3*) and 54% of poised promoter state (state *1*).

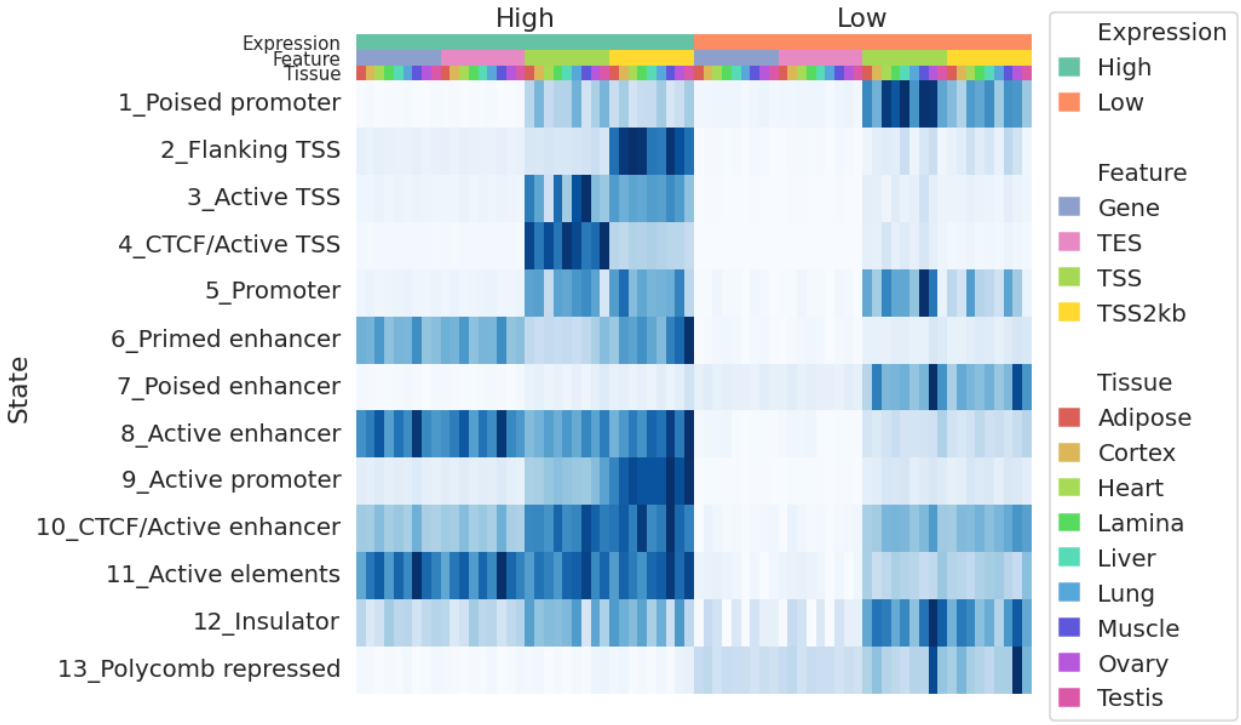

Figure 4.3 State enrichment in tissue specific genes

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

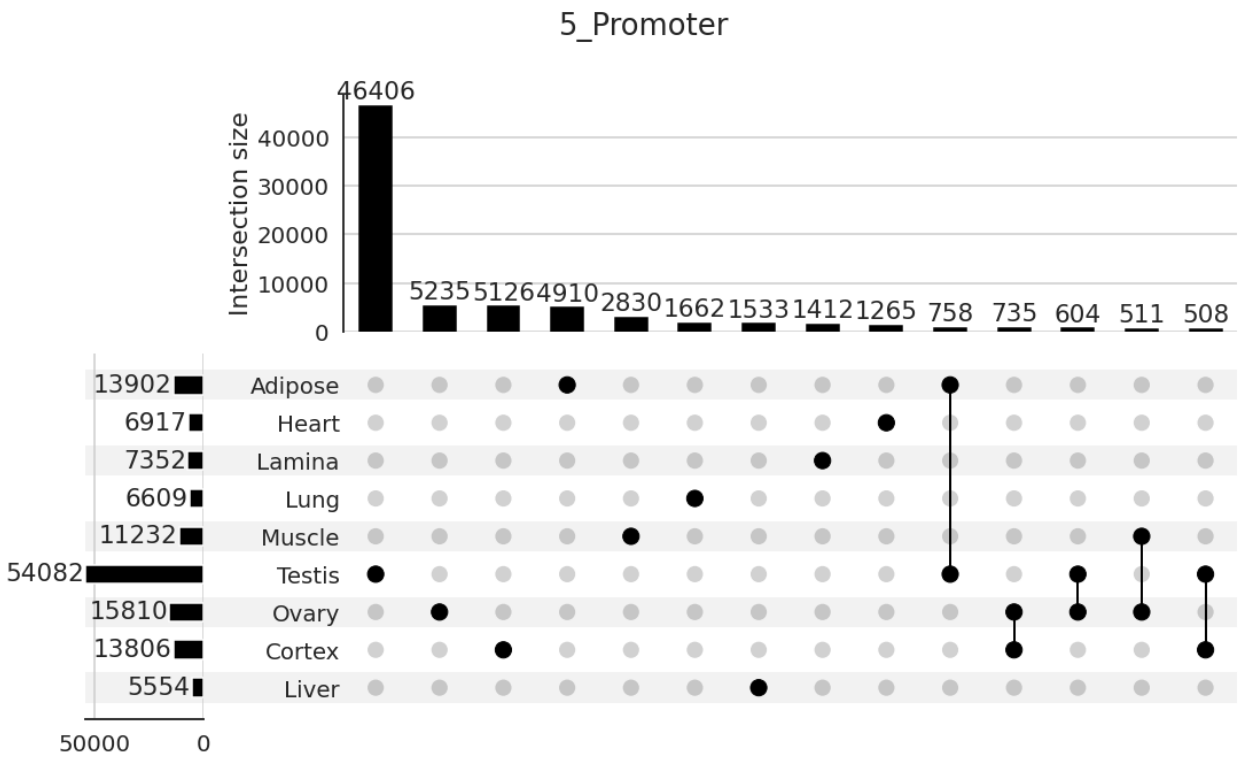

Figure 4.4 Promoter state shared across tissues

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 intersection with filled dots indicating that segments in this intersection were found in the corresponding tissue; Bottom left: bar plot indicates number of segments annotated as promoter state (state *5*) in each tissue

Chromatin states partially annotate open chromatin regions

To infer potential functions of open chromatin regions, especially those located in the intergenic regions, open chromatin peaks from ***Chapter 3*** were annotated based on overlap with annotated gene features as well as chromatin state segments in each tissue. First, we examined overlap between each chromatin region and different open chromatin states across tissues (**Fig 4.5A**). Open chromatin regions were categorized into 5 categories, based on their proximity or overlap with annotated gene features: overlapping promoter-TSS neighborhood (within 2kb up- and down-stream of TSS), overlapping transcription termination sites (TTS), overlapping exons, and overlapping introns. 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 4. 5A**).

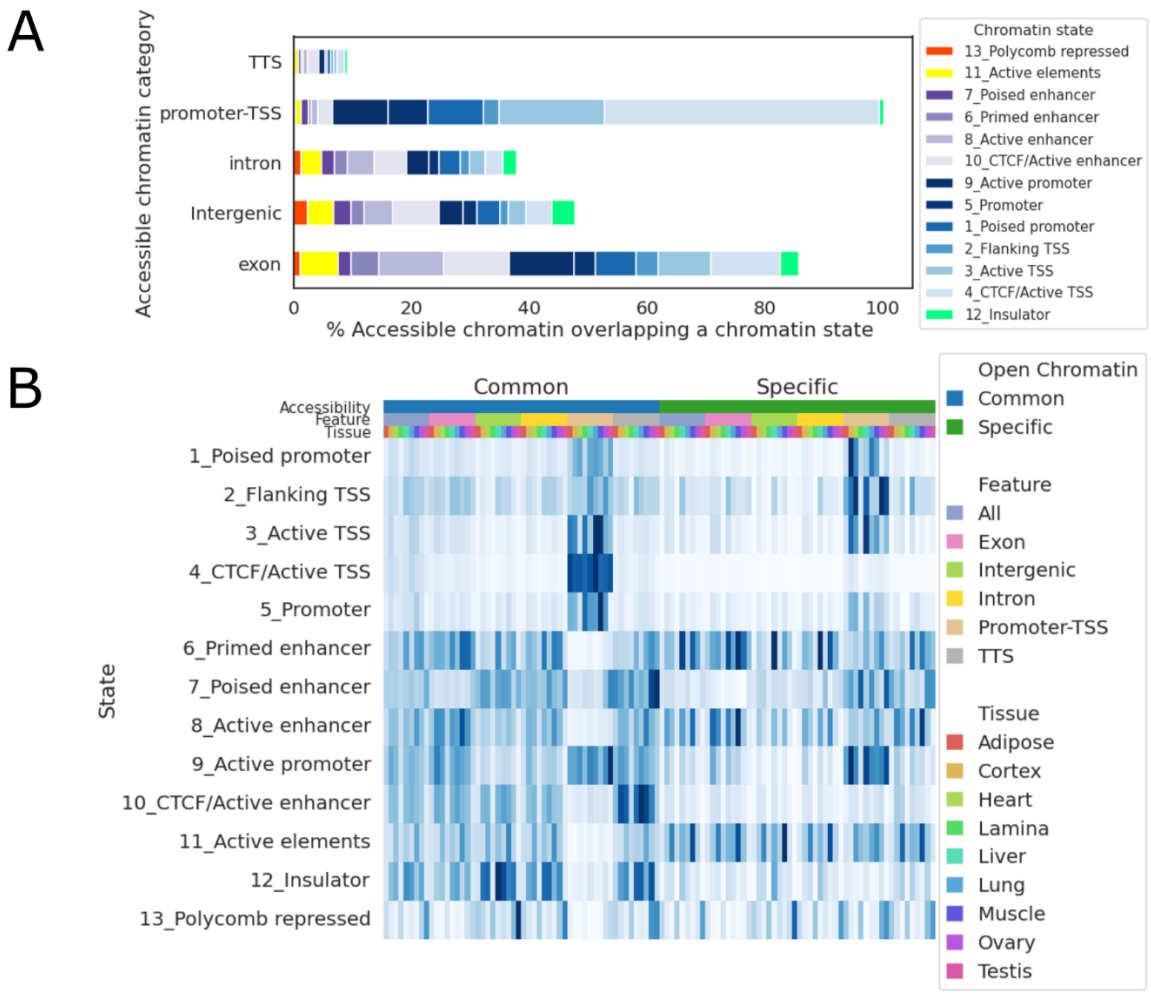

Figure 4.5 Chromatin accessibility across states

(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

Next, we compared chromatin state enrichment among shared and tissue-specific open chromatin regions (**Fig 4.5B**). 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).

Predicting target genes of REs

Since many enhancers interact with genes other than their nearest neighbors [28], we calculated Spearman correlation coefficients between ChIP-seq and RNA-seq data to predict target genes of REs. To identify topologically associated domains (TADs) within which enhancer-promoter interactions occur [29] in the absence of Hi-C data, we predicted CTCF-mediated chromatin loops using CTCF ChIP-seq data, as described by Oti et al [30]. 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 [31], 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. This was comparable to a previous study that identified 2,200 TADs, spanning 91% of the mouse genome using Hi-C data from mouse cell lines [32].

As demonstrated by Kern et al. [33], H3K27ac intensity of an RE 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 RE-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 RE-gene pairs remained as candidates. These REs were then annotated as genic, intergenic, or TSS-proximal based on their relative proximities. A majority of these candidate pairs had their REs outside of the gene bodies or TSS-proximal regions (intergenic REs, 66,051) while only a small portion of them had promoter-like relationship (REs in the TSS-proximal regions, 8,225). Intergenic REs were found at varying distances to TSS, with a median distance of 200 Kb and 79% of REs being within 1 Mb their target TSS. We also observed more REs (75%) located downstream of their target genes (**Fig 4.6**).

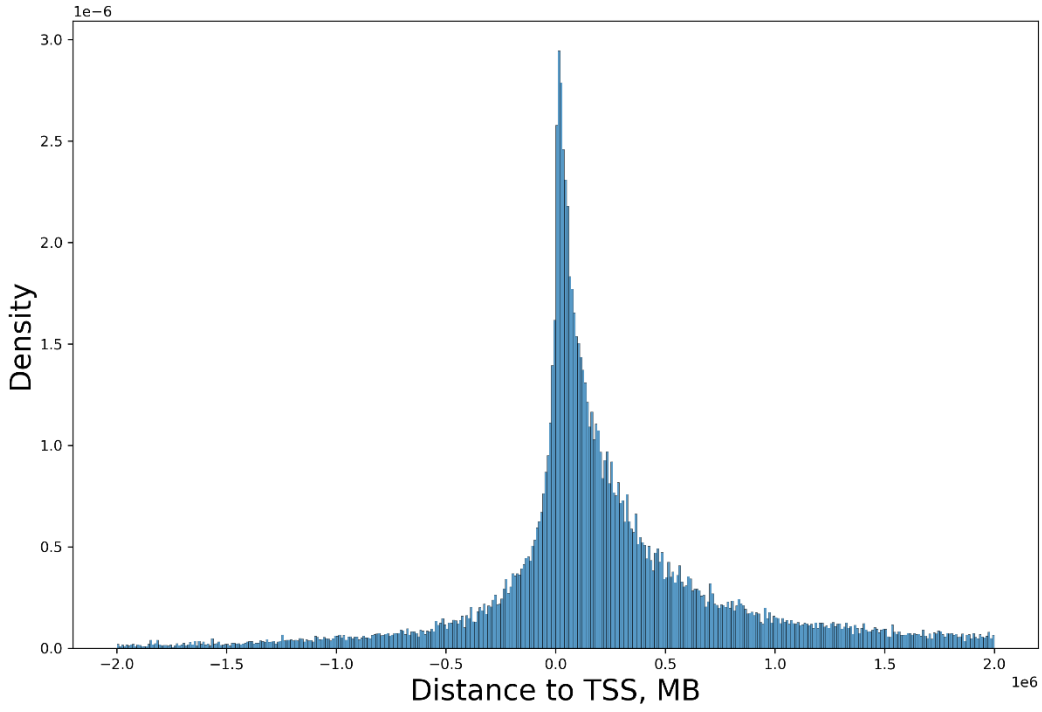

Figure 4.6 Distance from intergenic RE to target genes' TSS

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

A track hub of integrated FAANG dataset

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, in addition to the equine FAANG transcriptome, mRNA-seq data, as well as open chromatin regions (***Chapter 3***). **Fig 4.7** shows an example region from this track set, with two representative tissues displayed: brain and heart. Acyl-CoA thioesterase 11 (*ACOT11*) expression compared to the other tissues is outlined by the RNA-seq tracks. This gene was previously shown to be most abundantly expressed in heart [34] and our data corroborate this finding. The high expression of this gene was also supported by the active TSS states (states *3* and *4*) near the second annotated exon of *ACOT11* in heart, as well as the ATAC peaks near the same region.

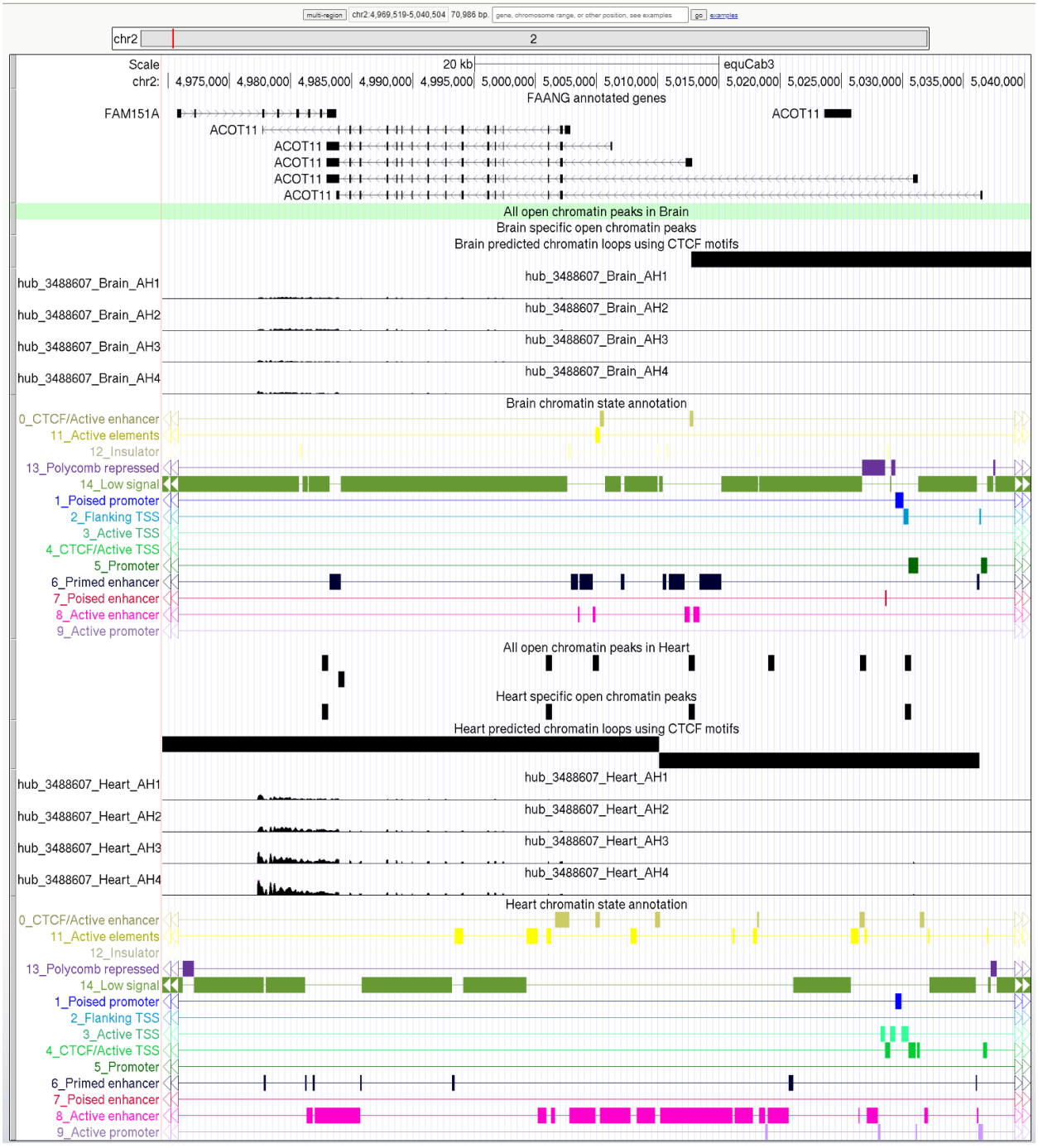

Figure 4.7 Equine FAANG UCSC tracks

An example region showing *ACOT11* and its surrounding regulatory elements in UCSC Genome Browser. Brain and heart are shown in order from top to bottom. Each tissue contains tracks showing ATAC peaks, predicted CTCF-mediated loops, RNA-seq read pileup, and chromatin states. Additional tracks can be enabled in the track settings.

Discussion

In this study, 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 REs. 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 [5,7,35]. 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 [36] 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 [37]. 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 [38].

Among the nine tissues assayed in this study, testis showed the most distinct regulatory landscape. It had the largest numbers of unique segments annotated as promoter state (state *5*), CTCF-less active TSS state (state *3*), and poised promoter state (state *1*), reflecting unique transcriptome complexities in testes, as has been previously demonstrated [39,40].

Overall, this study presents the first comprehensive overview of REs across a diverse range of tissues in the horse. Taken together with the previously reported tissue-specific transcriptome and chromatin accessibility map, these annotated states provide a critical resource enable a better understanding of the regulatory landscape impacting complex traits in the horse.

Methods and Materials

Chromatin state discovery

ChIP-seq data for histone modifications were obtained from previously published studies [20,21]. Additionally, ChIP-seq for CTCF was performed for the same nine frozen tissue samples at Diagenode Inc. Briefly, CTCF ChIP libraries were sequenced at 50bp single- and paired-end (female and male samples, respectively). Reads were aligned to EquCab3.0 [41] using bwa mem [42] with default parameters. Aligned reads were subsequently filtered to remove low-quality mapping, PCR duplicates, and mitochondria reads using SAMTools [43]. BAM files for all five marks were binarized using ChromHMM [24] 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.

Transcript analyses

RNA-seq data was obtained from and processed as described in ***Chapter 2***. In addition, transcript level TPM values were summarized to gene level using tximport [27]. 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 transcriptome from ***Chapter 2***.

Enrichment analysis

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

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

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. $N_{Ann \cap State}$ 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.

Open chromatin annotation

Open chromatin peaks for the same nine tissues were identified as previously discussed (see ***Chapter 3***). Each set of open chromatin peaks were annotated based on their proximity to or overlap with several geneic features (promtoer-TSS neighborhood, exon, intron, TTS), intergenic regions, and chromatin states, using annotatePeaks.pl from HOMER [44].

Data Access

CTCF ChIP-seq data can be accessed from SRA/ENA under project accession PRJEB41079. Histone ChIP-seq data were published by Kingsley et al. [20] and Barber [21]. All other data were discussed in previous chapters.

Funding: This project was supported by the Grayson-Jockey Club Research Foundation, Animal Breeding and Functional Annotation of Genomes (A1201) Grant 2019-67015-29340 from the USDA National Institute of Food and Agriculture and the UC Davis Center for Equine Health with funds provided by the

State of California pari-mutuel fund and contributions by private donors. Additional support for C.J.F.

was provided by NIH L40 TR001136.

Supplementary Figure 4.1 Active TSS and poised promoter states shared across tissues; Intersection plot showing number of segments annotated as A) Active TSS or B) Poised promoter states unique to each tissue and shared across tissues. Top: bar plot indicates sizes of each intersection; Bottom right: each column denotes a unique intersection with filled dots indicating that segments in this intersection were found in the corresponding tissue; Bottom left: bar plot indicates number of segments annotated as A) Active TSS (state *3*) or B) Poised promoter state (state *1*) in each tissue

References

1. Hansen, A. S., Iryna, P., Claudia, C., Tjian, R. & Xavier, D. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. eLife; Cambridge 6, (2017).
2. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64 (2017).
3. Sos, B. C. et al. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol 17, 20 (2016).
4. Liu, C. et al. An ATAC-seq atlas of chromatin accessibility in mouse tissues. Sci Data 6, 65 (2019).
5. Warburton, A., Breen, G., Rujescu, D., Bubb, V. J. & Quinn, J. P. Characterization of a REST-Regulated Internal Promoter in the Schizophrenia Genome-Wide Associated Gene MIR137. Schizophr Bull 41, 698–707 (2015).
6. Giorgio, E. 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 24, 3143–3154 (2015).
7. Gupta, R. M. et al. A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. Cell 170, 522-533.e15 (2017).
8. Zentner, G. E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. Nature Structural & Molecular Biology 20, 259–266 (2013).
9. Zhang, Y. et al. Overview of Histone Modification. in Histone Mutations and Cancer (eds. Fang, D. & Han, J.) vol. 1283 1–16 (Springer Singapore, 2021).
10. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. Exp Mol Med 49, e324–e324 (2017).
11. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39, 311–318 (2007).
12. Santos-Rosa, H. et al. Active genes are tri-methylated at K4 of histone H3. Nature 419, 407–411 (2002).
13. Lauberth, S. M. et al. H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. Cell 152, 1021–1036 (2013).
14. Bian, C. 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 30, 2829–2842 (2011).

15. Eberl, H. C., Spruijt, C. G., Kelstrup, C. D., Vermeulen, M. & Mann, M. A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics. Molecular Cell 49, 368–378 (2013).
16. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107, 21931–21936 (2010).
17. Boyer, L. A. et al. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441, 349–353 (2006).
18. Burns, E. N. et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim Genet 49, 564–570 (2018).
19. Donnelly, C. G. et al. Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. Front. Genet. 12, 650305 (2021).
20. Kingsley, N. B. et al. Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. Genes 11, 3 (2019).
21. 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 233, 143 (2022).
22. Mikkelsen, T. S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448, 553–560 (2007).
23. Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. PNAS 112, E6456–E6465 (2015).
24. Hon, G., Ren, B. & Wang, W. ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome. PLoS Comput Biol 4, e1000201 (2008).
25. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. Nature Methods 9, 215–216 (2012).
26. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14, 417–419 (2017).
27. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4, 1521 (2015).
28. Zhang, Y. et al. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature 504, 306–310 (2013).
29. Rao, S. S. P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).
30. Oti, M., Falck, J., Huynen, M. A. & Zhou, H. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. BMC Genomics 17, 252 (2016).
31. Zwillinger, D. & Kokoska, S. CRC standard probability and statistics tables and formulae. (Chapman & Hall/CRC, 2000). Section 14.7.
32. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).
33. Kern, C. et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nat Commun 12, 1821 (2021).
34. Fagerberg, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics 13, 397–406 (2014).
35. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

36. Khoury, A. et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat Commun 11, 54 (2020).
37. Kubo, N. et al. Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat Struct Mol Biol 28, 152–161 (2021).
38. Franco, M. M., Prickett, A. R. & Oakey, R. J. The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction1. Biology of Reproduction 91, (2014).
39. Soumillon, M. et al. Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Reports 3, 2179–2190 (2013).
40. Guo, J. et al. The adult human testis transcriptional cell atlas. Cell Res 28, 1141–1157 (2018).
41. Kalbfleisch, T. S. et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun Biol 1, 197 (2018).
42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
43. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
44. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576–589 (2010).

Concluding Discussion

In this thesis, 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. We showed, in three stages, either improvement to current equine genome annotation or inclusion of newly annotated regulatory elements. We anticipate these new resources will play a vital role to understanding how genetic variation in the horse contributes to equine biology and health.

In ***Chapter 2***, we outlined an approach to expand the equine transcriptome annotation. 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-to-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.

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.

Next, we identified potential regulatory states genome wide across tissues using signals from biochemical assays (histone modifications and CTCF ChIP-seq) and correlated them with accessible DNA elements identified from ***Chapter 3***. We observed that active regulatory states were enriched around

TSS-promoter regions, especially in genes with high expression. We also noted a significant enrichment of CTCF-bound active states among tissue-conserved accessible DNA elements, suggesting important housekeeping roles of constitutively bound CTCF and the chromatin structures that they maintain.

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 potential RE-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 thesis 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 cell-type differences in these tissues. Future studies utilizing single-cell-based technologies could help unravel such differences and identify cell-type-defining genes and REs. 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.

Addendum 1 Successful ATAC-Seq From Snap-Frozen Equine Tissues

Published in: Peng, S., Bellone, R., Petersen, J. L., Kalbfleisch, T. S. & Finno, C. J. Successful ATAC-Seq From Snap-Frozen Equine Tissues. *Front. Genet.* 12, 641788 (2021). https://doi.org/10.3389/fgene.2021.641788

Keywords: FAANG; horse; open chromatin; ATAC-seq

Abstract

An assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) has become an increasingly popular method to assess genome-wide chromatin accessibility in isolated nuclei from fresh tissues. However, many biobanks contain only snap-frozen tissue samples. While ATAC-seq has been applied to frozen brain tissues in human, its applicability in a wide variety of tissues in horse remains unclear. The Functional Annotation of Animal Genome (FAANG) project is an international collaboration aimed to provide high quality functional annotation of animal genomes. The equine FAANG initiative has generated a biobank of over 80 tissues from two reference female animals and experiments to begin to characterize tissue specificity of genome function for prioritized tissues have been performed. Due to the logistics of tissue collection and storage, extracting nuclei from a large number of tissues for ATAC-seq at the time of collection is not always practical. To assess the feasibility of using stored frozen tissues for ATAC-seq and to provide a guideline for the equine FAANG project, we compared ATAC-seq results from nuclei isolated from frozen tissue to cryopreserved nuclei (CN) isolated at the time of tissue harvest in liver, a highly cellular homogenous tissue, and lamina, a relatively acellular tissue unique to the horse. We identified 20,000–33,000 accessible chromatin regions in lamina and 22–61,000 in liver, with consistently more peaks identified using CN isolated at time of tissue collection. Our results suggest that frozen tissues are an acceptable substitute when CN are not available. For more challenging tissues such as lamina, nuclei extraction at the time of tissue collection is

still preferred for optimal results. Therefore, tissue type and accessibility to intact nuclei should be considered when designing ATAC-seq experiments.

Introduction

The completion of the equine genome assembly [1,2] has enabled research leading to novel discoveries concerning the health and reproduction of horses [3–5]. However, despite having the same genomic sequence, differential regulation of gene expression leads to tissue-specific profiles. A lack of understanding of gene regulation has largely stalled research of complex traits in horses. In humans and mice, the Encyclopedia of DNA Elements (ENCODE) project has provided an abundance of data for understanding gene regulation and its role in complex diseases and traits [6]. Unfortunately, limited resources are currently available in the horse. The Functional Annotation of Animal Genome (FAANG) initiative [7] is an international collaboration aimed to bridge this gap between genotype and phenotype. The equine FAANG project has successfully generated a biobank of over 80 tissues and bodily fluids of two reference animals [8]. RNA-seq of 32 tissues (unpublished, data access: PRJEB26787), as well as the identification of tissue specific histone marks for eight prioritized tissues, from this biobank has been performed (Kingsley et al., 2019). Additional projects are underway to identify tissue specific chromatin states to integrate all of these datasets and build a robust tissue specific functional annotation atlas in the horse [9].

An important component of gene expression and regulation is chromatin accessibility. Active genes and regulatory elements are typically found within or near regions of the DNA accessible to transcription factors. Therefore, identifying open chromatin regions is a crucial step to identify and categorize tissue specific regulatory elements in order to advance our understanding of complex traits in the horse. An assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) [10] is

commonly used to identify regions of open chromatin. A typical ATAC-seq protocol requires nuclei extracted from fresh tissues. Halstead et al proposed a modified ATAC-seq protocol to allow long-term storage of cryopreserved nuclei extracted from fresh tissues [11]. Still, the intensive efforts needed to prepare and cryopreserve nuclei during a large-scale tissue collection proves to be difficult. Alternatively, Corces et al successfully applied a modified ATAC-seq (Omni-ATAC) protocol on frozen human brain tissues [12]. However, the applicability of Omni-ATAC has not been tested in a wide variety of tissues in horse where nuclei extraction may prove challenging. Additionally, it has been shown that, in cultured cells, cryopreservation is preferrable to flash-freezing process in order to preserve native chromatin structures [13]. To our knowledge, no studies have investigated the effect of snap freezing on tissues for ATAC-seq library generation in comparison to cryopreserved nuclei preps. Additionally, the library preparation step is a major source of variation in RNA-seq studies [14], particularly at low read depth. As a result, RNA-seq data generated from different laboratories or at different times cannot often be directly compared. For a collaborative project, it is important to assess the effect of technical variations to better inform project planning and analytical decisions for data integration.

To address these gaps of knowledge in the applicability of ATAC-seq in snap-frozen horse tissues, and to provide a guide for future ATAC-seq studies to assess chromatin accessibility, we compared data from cryopreserved nuclei prepared from fresh tissue to that of nuclei extracted from snap-frozen tissues collected from the two mares from the initial equine FAANG biobank study [8]. In order for this comparison to be informative and applicable to a wide range of tissues, we utilized both liver, a highly cellular and homogenous tissue type, and lamina, a relatively acellular tissue unique to the horse. Equine laminae are highly vascularized interdigitated dermal and epidermal tissues in the equine foot that form the attachment between the hoof wall and the third phalanx. Inflammation of laminae in horses (i.e. laminitis) is a devastating disease that impacts many breeds of horses and often leads to euthanasia. Therefore, gene regulation in laminae is of particular interest to equine geneticists and

veterinary practitioners as this debilitating and life-threatening disease estimated to impact up to 34% of the horse population [15]. Laminitis is also the primary clinical consequence of equine metabolic syndrome (EMS) [16]. EMS is a complex syndrome that requires constant veterinarian care and diet control, impacting an estimated 18 to 27 percent of horse population [16]. Liver is the primary metabolic organ with a homogenously cellular structure. Detailed knowledge of gene expressions and regulations in healthy liver provides a baseline for studying impaired metabolism in horses with EMS. Additionally, to assess the effect of library preparation techniques, snap-frozen tissues and cryopreserved nuclei from this pilot study were sent to two different core laboratories for library generation and subsequent sequencing. We hypothesized that (1) ATAC-seq using frozen tissues would identify comparable peaks to those using cryopreserved nuclei from fresh tissues, (2) libraries generated from liver will have better quality than those from laminae, and (3) similar to what was found in RNA-seq studies there will be a significant amount of variation between the libraries generated by two laboratories.

Materials and Methods

Tissue collection and nuclei isolation

Liver and lamina tissues from two mares (AH2 and AH1) were collected as described in [8]. Briefly, two healthy adult Thoroughbred mares (AH1: 5 years old; AH2: 4 years old) were closely examined by veterinarians prior to tissue collection. Nuclei were isolated from liver and lamina tissues immediately following tissue collection and cryopreserved following protocols published in 11 with some modifications for lamina. Briefly, additional incubation periods with collagenase were added to assist in homogenization (see **Supplementary Materials** A1). These are referred to as cryopreserved nuclei (CN). Additionally, at time of collection, approximately 1 g aliquots of tissue were snap frozen in liquid nitrogen for nuclei extraction at a later time. These are referred to as frozen tissue-derived nuclei (FTDN).

ATAC-seq library preparation and sequencing

Both snap frozen tissues and cryopreserved nuclei were stored at -80°C for 3 years until shipped on dry ice overnight to two commercial laboratories (L1 and L2) for library preparation. Nuclei were extracted from frozen tissues using each laboratory's internally optimized protocol (See **Supplementary Materials A1**). Extracted Nuclei (FTDN) and cryopreserved nuclei (CN) were used to prepare ATAC libraries (Supplementary Methods and Supplementary Table 1). Libraries were sequenced on an Illumina HiSeq 4000, paired-end 2x75bp (L1) or NextSeq 500, paired-end 2x42bp (L2) with a targeted depth of 30 million read pairs.

ATAC-seq data analysis

Read QC was carried out using FastQC [17]. Adapters and low-quality ends were trimmed using TrimGalore [18] and Cutadapt [19]. Reads were then aligned to reference genome EquCab3 using BWA-MEM algorithm from BWA [20] using default parameters. Post-alignment filtering was employed to remove low mapping quality reads, mitochondrial reads, and PCR duplicates using Samtools [21] and Sambamba [22]. Genome coverage was analyzed using deepTools [23] Specifically, bamCoverage was used to convert bam files to bigwig files, using RPKM to normalize coverage with exact scaling (--normalizeUsing RPKM --exactScaling). Then multibigwigSummary was used to calculate average coverage across 1000bp windows (-bs 1000). plotPCA was used to calculate eigen values based on all genomic windows (--ntop 0) and top 2 principle components were plotted using matplotlib [24]. Custom scripts were used to analyze sample correlation, clustering, and correlation with ChIP-seq data and annotated genes using Python packages numpy [25], scipy [26], pandas [27], and matplotlib [24]. Open regions were identified using HMMRATAC (--threshold 2 --score fc -u 20 -l 10) [28] and MACS2 (-q 0.05 -B –broad -f BAMPE) [29]. Jaccard indices were calculated using pybedtools [30,31] for each pair of

biologic replicates with default parameters. More detailed pipeline is available at https://github.com/SichongP/FAANG_ATACseq.

Histone ChIP-seq data processing

Histone ChIP-seq data were downloaded from FAANG data repository (https://data.faang.org/home) under accession PRJEB35307. Histone marks were determined according to [32] and compared with open chromatin regions analyzed in this study for both liver and lamina.

ATAC-seq peak validation with histone marks

ATAC-seq peaks called by HMMRATAC and MACS2 were validated using histone ChIP-seq data following [28] with modifications to utilize available data in the horse. First, the following sets of peaks were generated from [32] data:

Real positive set (RP): peaks from either H3K4me1 or H3K4me3 that overlap H3K27ac peaks

Real negative set (RN): peaks from H3K27me3 data

Then, following metrics were calculated for each dataset:

$$\boldsymbol{TP = number\ of\ bases\ in\ called\ ATAC - seq\ peaks\ overlapping\ RP}$$

$$\boldsymbol{FP = number\ of\ bases\ in\ called\ ATAC - seq\ peaks\ overlapping\ RN}$$

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

$$\boldsymbol{Recall = \frac{TP}{RP}}$$

$$\boldsymbol{False\ Positive\ Rate\ (FPR)} = \frac{\boldsymbol{FP}}{\boldsymbol{RN}}$$

Increasing quality scores as produced by MACS2 or HMMRATAC were used as the cutoff score to filter peaks before the remaining peaks were used to calculate above metrics. Changes in the metrics as the cutoff score increased were used to identify the thresholds at which to filter final sets of open chromatin peaks.

RNA-seq data processing

RNA-seq reads from liver and lamina of the same two animals were available from a separate project under European Nucleotide Archive accession PRJEB26787. Briefly, RNA was isolated from liver or lamina tissues using Trizol chloroform phase separation followed by a column cleanup using Zymo Research Direct-Zol Mini columns. TruSeq mRNA libraries were prepared at Minnesota Genomics Center (Minneapolis, MN, USA) and sequenced at 125bp paired-end. These reads were quantified against Equcab3 Ensembl annotated genes [2,33] using Salmon [34] mapping-based mode. Transcript level counts were aggregated into gene level using the R package tximport [35] and final counts were normalized using the variance-stabilizing transformation method from DESeq2 vst function [36].

ATAC-seq peak validation with RNA-seq data

Ensembl annotated genes were classified as open or closed depending on whether their presumed promoter regions (1kb upstream of annotated gene start) overlapped with identified ATAC-seq peaks. These genes were then compared to their RNA abundance estimated using FAANG data.

Results

Libraries prepared by two laboratories (L1 and L2) using nuclei isolated from snap-frozen tissues (FTDN) or cryopreserved from tissues at time of collection (CN) from liver and lamina of two animals (AH1 and AH2, Thoroughbred adult mares) were sequenced at PE75 on an Illumina HiSeq 4000 (L1) or PE42 on an Illumina NextSeq 500 (L2). **Figure A1.1** shows a schematic of the experimental design.

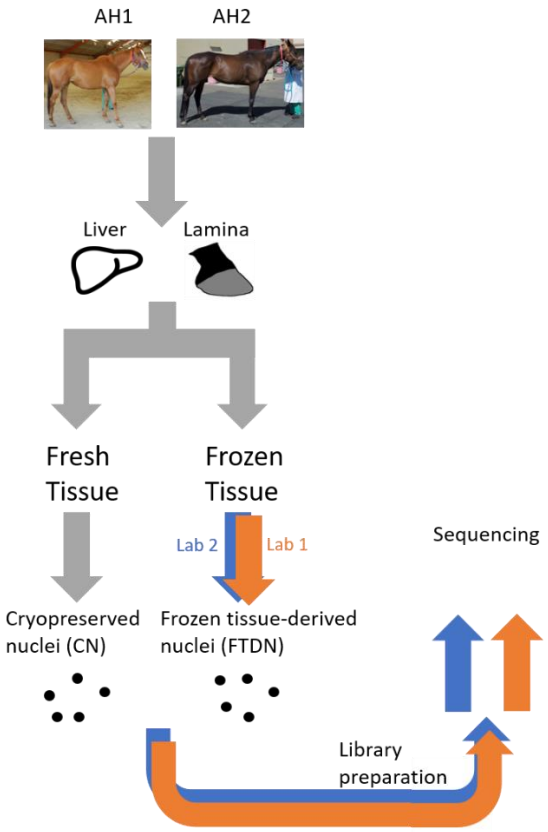

Figure A 1.1 A schematic of the experimental design

All samples were prepared at UC Davis prior to shipment to the core laboratories. Samples used were obtained from an equine biobank of two horses (AH1 and AH2), as previously described [8]

Library fragmentation

ATAC-seq libraries are expected to present a laddering pattern that corresponds to different nucleosome-bound fragments. **Supplementary Figures A1.1 and A1.2** show fragment size distributions

of ATAC libraries as determined by sequencing and Agilent Fragment Analyzer (L1) or TapeStation (L2) from L1 and L2, respectively. In general, liver libraries showed distinguishable laddering pattern while in lamina libraries, only the fragment size corresponding to nucleosome-free fragments was observed.

Sequencing read lengths

Since libraries from L1 and L2 were sequenced at different lengths (75bp and 42bp, respectively), we trimmed longer reads from L1 from 3' down to 42bp and compared read alignment statistics to those obtained using full length reads (75bp), after appropriate quality trimming. There were no significant changes in read alignment statistics, with less than 0.02% fewer reads aligned and less than 0.3% fewer reads identified as duplicates for each library after length trimming. Therefore, we proceeded with data analysis with using original full length reads from both laboratories.

Duplication rate and mitochondrial contamination

Overall, liver libraries have higher mitochondrial contamination than lamina libraries, likely due to higher metabolic activities in liver (**Supplementary Figure A1.3A**). Among liver samples, CN libraries prepared by L1 contained 56% and 81% duplicates, with 37% and 23% mitochondrial reads in AH1 and AH2, respectively. In comparison, the CN libraries from L2 contained 31% and 24% duplicates, with 23% and 10% mitochondrial reads from AH1 and AH2 respectively (**Supplementary Figure A1.3A**). It was suspected that the higher amount of mitochondrial contamination contributed to the higher duplication rate and led to lower library complexity. To test this hypothesis, resequencing was performed for the liver CN libraries from L1. The number of unique nuclear reads from AH2 largely remained unchanged despite increasing read depth 3-fold. For AH1, however, twice the number of unique nuclear reads was obtained after the total read depth was increased (**Supplementary Figure A1.3B**). Both the fingerprint plot and fraction of reads in peaks (FRiP) identified a decrease in enrichment for AH1 with increased

sequencing depth but little change for AH2 (**Supplementary Figure A1.3C** and **Supplementary Table A1.2**). This suggests that, in the AH1 library, while further sequencing increased the number of unique reads, it did not substantially improve peak detection. Lowered enrichment in the resequenced AH1 library suggests that a majority of additional unique reads are less enriched background reads. In the AH2 library, however, resequencing did not significantly improve library complexity, due to more cycles of amplification during library preparation and therefore, higher PCR duplication rate in the library.

Genome coverage and enrichment

To assess which part of the ATAC-seq protocol contributed more to library variations and complexities, we compared genome coverage and enrichment (**Figure A1.2**). Principle component analysis (PCA) revealed that liver libraries generally clustered closely together, while more variation was observed for the lamina libraries (**Figure A1.2A**). Within the lamina libraries, there is a clear clustering based on which laboratory prepared the libraries. The lamina libraries from L2 clustered closely with each other and with liver libraries while the lamina libraries from L1 clustered further away from liver libraries (**Figure A1.2A**). Heatmaps of the genome coverage Pearson correlation showed that liver CN libraries yielded well-correlated results, with the exception of that from AH2 by L1 (**Figure A1.2B**). This is consistent with low complexity of that library shown in Supplementary Figure 3. On the other hand, little correlation is observed among lamina library preparations (**Figure A1.2B**). Since no input libraries were used for ATAC-seq experiments 10, synthetic Jensen-Shannon distance (SJSD) was used, together with Area Under Curve (AUC) from fingerprint plots, to assess the enrichment of each library (**Figure A1.2C** and **Supplementary Table A1.3**). In general, liver libraries showed higher enrichment than lamina libraries. Within liver libraries, CN libraries were more enriched than FTDN libraries from L1, while both libraries from L2 showed similar enrichment. Within lamina libraries, both laboratories generated more enriched

libraries from CN than from FTDN. This is further exemplified in Figure 2D, showing the FRiP in each library.

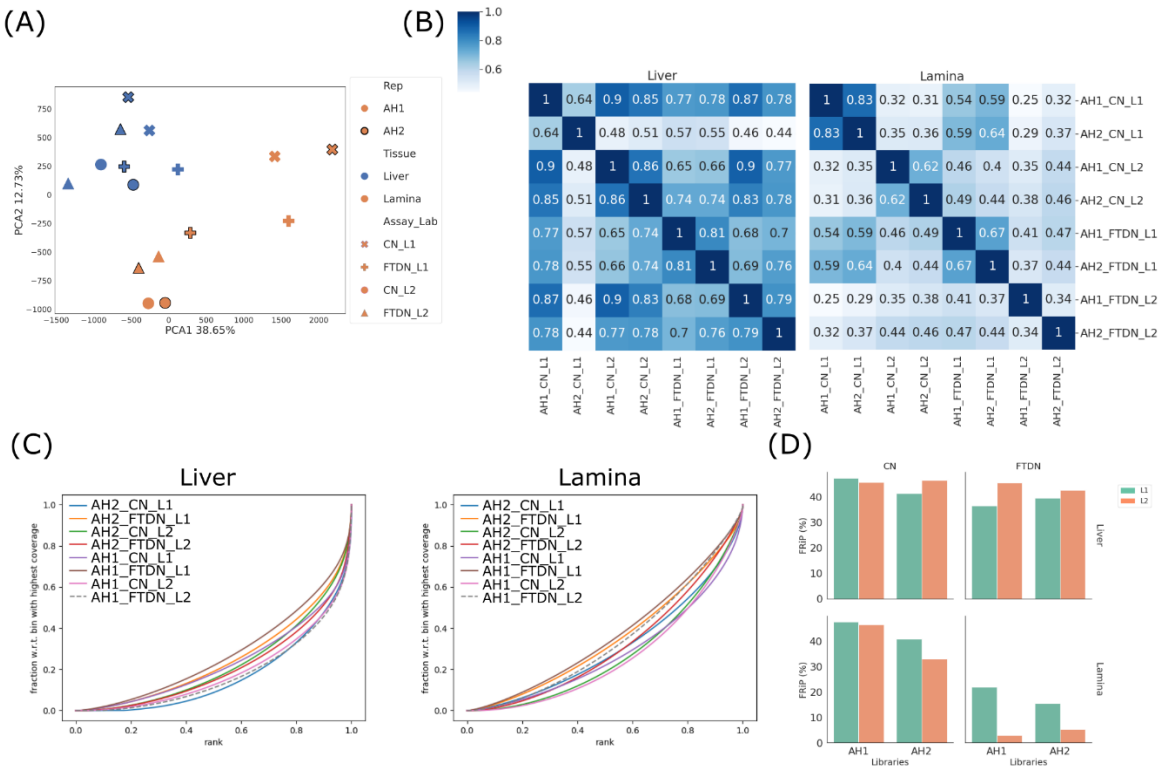

Figure A 1.2 coverage correlation between libraries.

Read depth was normalized across all libraries. (A) Principal component analysis of genome coverage, showing the first two principal components. (B) Pearson correlation of genome coverage in liver (left) and lamina (right) libraries. Linkage was calculated using Farthest Point Algorithm. (C) Fingerprint plot of genome coverage in liver (left) and lamina (right) libraries. (D) Enrichment as measured by FRiP in each library.

Peak Calling

To identify accessible chromatin regions, MACS2 29 and HMMRATAC 28 were used to call peaks and results from both programs were compared. To control for sequencing depth, all libraries were down-sampled to 60 million unique reads that are suitable for peak calling using sambamba view function. Using MACS2 (-q 0.05 -B –broad -f BAMPE), 31,000-721,000 peaks were identified. While using HMMRATAC (--threshold 2 --score fc -u 20 -l 10), 14,000-514,000 peaks were identified. Overall, using

HMMRATAC, peaks identified from lamina libraries had lower quality (fewer (Figure **A1.3A**) and shorter peaks (Figure **A1.3B**) with lower scores (**Figure A1.3C**)) than those from liver libraries. For liver libraries, CN generated comparable results to FTDN while, in lamina libraries, CN outperformed FTDN (Figure **A1.3D**). Similar results were obtained when peaks were called using MACS2 (**Supplementary Figure A1.4A-D**).

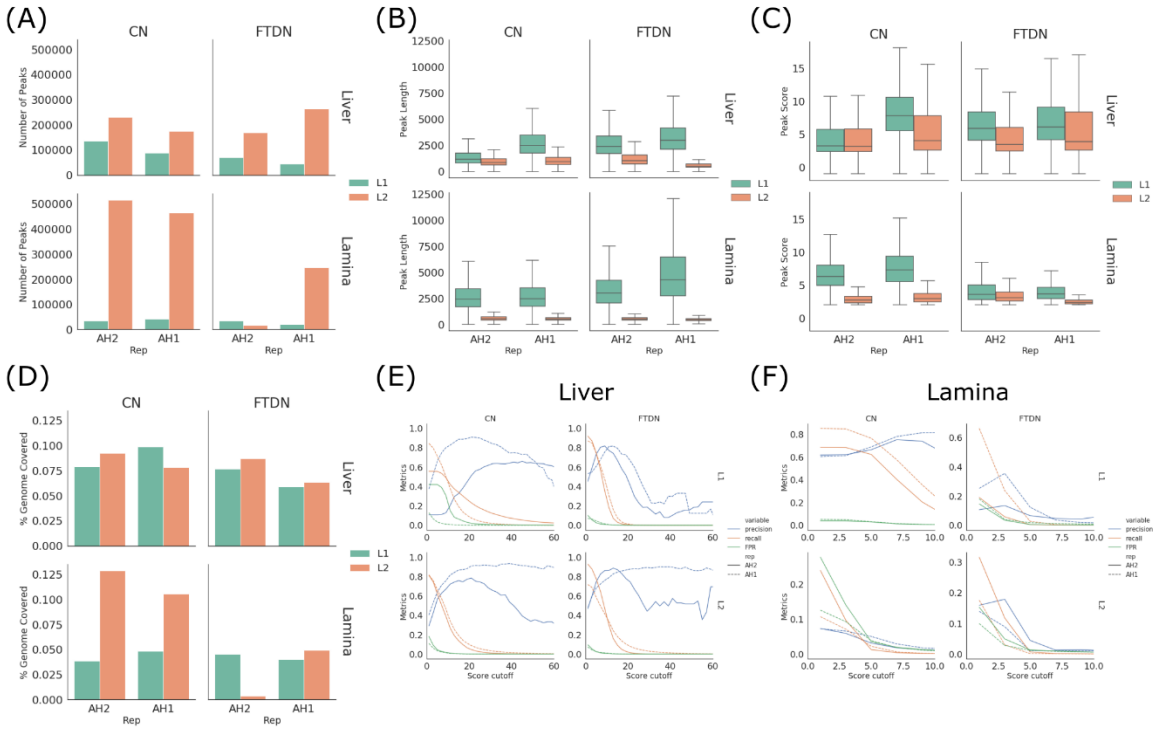

Figure A 1.3 HMMRATAC peak calling statistics

(A) Number of peaks, (B) peak length distribution, (C) peak score distribution, and (D) percent of genome covered by peaks for each library. (E-F) Peak metrics assessed using ChIP-seq dataset in liver (E) and lamina (F) libraries.

To better assess the quality of peaks, we used histone mark ChIP-seq data generated from the same samples as described in [32]. A set of metrics, precision, recall, and false positive rate (FPR), were generated for different cutoff scores as described in Methods. These metrics were then plotted against cutoff scores. Consistent with the observation of peak lengths and scores, peaks called using HMMRATAC from liver libraries had higher precision and recall rates and lower false positive rates

(**Figure A1.3E**) than lamina (**Figure A1.3F**). Consistent with observations of library quality, CN liver libraries of AH2 from L1 have lower recall and precision rates than that from L2 or that of AH1, despite having same unique read depth (**Figure A1.3E,F**). Comparing peaks identified by two programs, HMMRATAC identified peaks with higher recall and precision rates than MACS2 (**Supplementary Figure A1.4 E-F**).

ATAC-seq peak validation

Despite higher quality from L2 in liver AH2 CN library, L1 produced the only libraries from laminae with high quality peaks (**Figure A1.3F**). Therefore, to maximize usable data, libraries from L1 were chosen for all further analyses. HMMRATAC was used as it produced generally better metrics and because it allowed interrogation of nucleosome-bound regions vs. nucleosome-free regions for future studies.

A cutoff score, where the precision and recall lines intercept, was used for each sample set to filter peaks identified by HMMRATAC. Final peak counts are shown in **Table A1.1**. Consistent with previous observations, liver samples generated the most high-quality peaks, while CN libraries outperformed FTDN libraries. Using UpSetPlot [37] based on [38], we identified overlapping peaks in each dataset (**Figure A1.4A**). AH1 liver CN library generated the most unique peaks, consistent with the previous observation that this library has highest library complexity. Since 17,347 unique peaks were identified from this library only, a precision score of these unique peaks was calculated using histone ChIP-seq data mentioned above. A precision score of 18.4% was observed in these peaks, suggesting a high rate of false positive peaks. This further highlights the importance of replicates in an ATAC-seq experiment. FTDN libraries did not yield significant number of unique peaks that were not detected in CN libraries. Despite a relatively low quality of the lamina libraries, 12,256 unique peaks were detected from the lamina libraries.

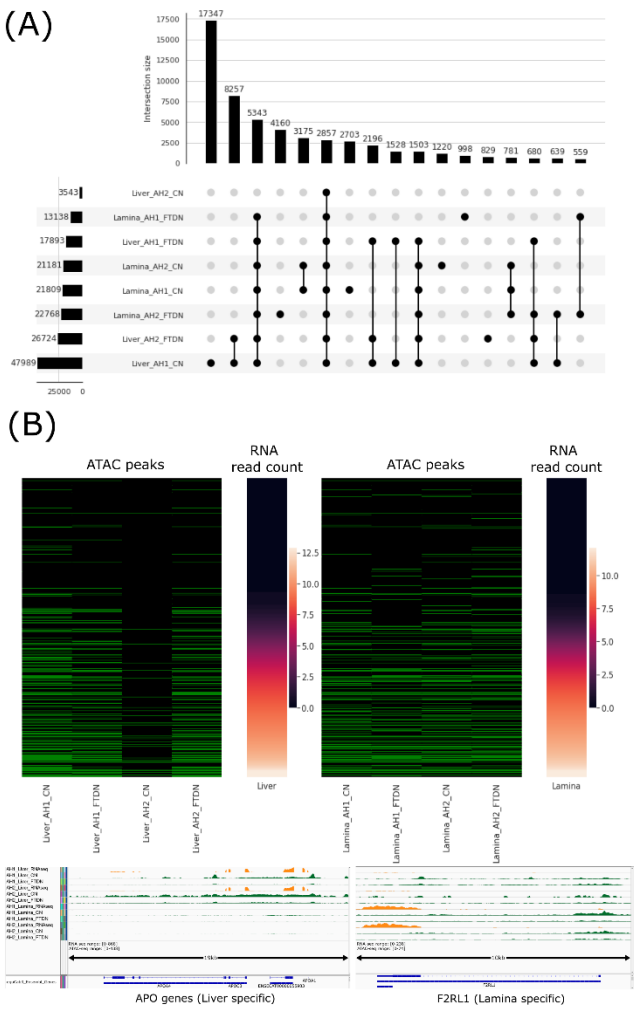

Figure A 1.4 Filtered ATAC-seq peaks

(A) Intersection plot of quality filtered peaks from each library. Bottom left panel shows filtered peak count in each library; bottom right panel shows different intersections (BedTools, 1bp minimum) of peaks where filled dots indicate presence of peaks in corresponding library; Top panel shows peak count in each intersection. (B) Relationship between promoter accessibility and gene expression (mean vst transformed count) in liver (top left) and lamina (top right). Green cell in ATAC peaks indicate presence of ATAC peaks and black cells indicate absence. Bottom panel shows bigwig tracks of RNA-seq and ATAC-seq read abundance (normalized using RPKM) near APO genes (left, liver specific) and F2RL1 (right, lamina specific) transcription start sites.

Table A 1.1 Cutoff used to filter peaks and metrics of filtered peaks

Filtered peaks and their corresponding cutoff scores in each library. AvePeakLen: Average peak length; MedianPeakLen: Median peak length; Bases Covered: number of bases covered by all peaks in a library; Confirmed Count: Overlapping peaks in both biological replicates; Jaccard Index: jaccard index of two biological replicates

Tissue	Rep	Nuclei Prep	Cutoff Score	Count	Ave PeakLen	Median PeakLen	Bases Covered	Confirmed Count	Jaccard Index
Liver	AH1	CN	6	61473	2937.0	2600	180,547,240	3646	0.05
Liver	AH2	CN	16	3810	2428.8	2090	9,253,670		
Liver	AH1	FTDN	6	22588	3701.6	3300	83,611,751	18596	0.35
Liver	AH2	FTDN	6	33782	3059.1	2650	103,343,612		
Lamina	AH1	CN	6	28418	3106.6	2650	88,284,203	23439	0.51
Lamina	AH2	CN	4	30906	2883.5	2480	89,117,724		
Lamina	AH1	FTDN	2	19886	5061.4	4300	100,651,092	17619	0.35
Lamina	AH2	FTDN	2	33762	3361.9	3010	113,504,835		

As an in silico validation of the results, peaks were overlapped with Ensembl gene annotation for EquCab3 2 at promoter regions (1kb upstream of annotated gene start) to classify each promoter as open or closed. These classified promoter regions were then compared to RNA abundance at the corresponding gene level (**Figure A1.4B**). In liver, AH1 CN identified more open promoters where RNA expression levels are high but the results from the two assays (CN and FTDN) were highly comparable for this sample in liver. Fewer peaks were identified from AH2 CN, due to low library quality and issues in repeat freeze thaw cycles as outlined in the discussion. In lamina, CN assays identified more open promoters than FTDN. Manual inspection of some highly abundant genes in liver and laminae validate accurate identification of open chromatin in each tissue (**Figure A1.4B**)

Overall, our results confirm that extracting nuclei from snap-frozen tissues for ATAC-seq library preparations negatively affects the library quality, resulting in fewer peaks detected. However, when cryopreserved nuclei from freshly collected tissue are not available, these data show that snap-frozen tissues can be used to prepare ATAC-seq libraries to give reliable peak calls, with the caveat that some regions of open chromatin will be missed. However, results from laminae suggest that for more challenging tissue types, fresh tissue extraction is a requirement.

Discussion

In this pilot study, we compared two tissues (liver and laminae, representing homogenous cellular and relatively acellular nuclei extraction, respectively) from the equine FAANG project for ATAC-seq library generation, using two nuclei extraction methods. Nuclei extracted and cryopreserved immediately after tissue collection and nuclei isolated from snap-frozen tissues were used to determine suitable methods for performing ATAC-seq to identify accessible chromatin regions in a wide variety of equine tissues for functional annotation. Similar to what was identified by [11], we determined that ATAC-seq can be used to characterize open chromatin in animal tissue but optimization is necessary to have a robust data set across tissues. Further, we found that while cryopreserved nuclei generally yield more peaks, frozen tissues can still be used to isolate nuclei and identify accessible regions. However, the quality of libraries generated by the frozen tissue protocol suffered when nuclei were extracted from a more challenging, relatively acellular tissue, such as laminae. Therefore, for challenging tissues, care should be taken at time of collection to prioritize those tissues for nuclei extraction and cryopreservation when possible.

We also showed that the frozen tissue protocol is more prone to variations introduced at the library preparation step. Specifically, FTDN liver libraries generated at two different laboratories only have a moderate correlation (0.68 for AH1 and 0.76 for AH2). Our analysis suggests that, similar to RNA-seq experiments, library preparation can introduce large variation that will impact subsequent data quality, specifically peak detection for ATAC-seq studies. However, since the two commercial laboratories used different internally optimized protocols, it is impossible to determine whether the variation was protocol-specific or lab-specific. Nonetheless, it is advisable for all ATAC-seq library preparations to be performed at a single site using the same protocols to minimize variability in datasets when trying to integrate information.

During library preparation, the cryopreserved nuclei aliquot from AH2 was partially thawed twice by L1 (first for an optimization experiment (data not shown) and then a second time to perform the data collection). The nuclei obtained during the second partial thawing were used in this study. Due to the precipitation of nuclei and contaminating mitochondria, this was likely the cause of low quality observed in that library preparation. The effect of different read lengths used by two laboratories was investigated and deemed to have no significant impact on read alignment. Our analysis suggested a detrimental impact on data quality by this practice and resequencing of this particular library also did not improve data quality nor was this resequencing effort able to identify more peaks. Therefore, it is advisable to avoid repeated partial thawing of cryopreserved nuclei aliquots.

Library fragment size screening using gel electrophoresis proved to be predictive of final fragment size distribution in sequencing results and data quality. As indicated in **Supplementary Figures A1.1 and A1.2**, a strong signature corresponding to nucleosome-free fragments without accompanying signatures for nucleosome-bound regions does not necessarily mean a high enrichment of nucleosome-free fragments. It could also indicate high levels of mitochondria contamination or fragmentation of chromatins before tagmentation, which are likely the cases in lamina libraries from L2.

We identified 20-33,000 accessible chromatin regions in lamina and 22-61,000 in liver, largely in line with observations of liver ATAC-seq from studies in other species [39–42]. As a preliminary study, we opted to include laboratory replicates in lieu of technical replicates in order to assess the effect of technical variations introduced during the library preparation step. Technical replicates would allow further validation of tissue specific open-chromatin. Following ENCODE standard [43] for ChIP-seq experiments, two biological replicates were collected for the FAANG project. However, more replicates would have allowed a more robust comparison between different protocols.

In this study, we demonstrated the feasibility of using snap-frozen tissues for ATAC-seq experiments for the equine FAANG project. For acellular tissues, more optimization is required for ATAC-seq experiments. We also showed that significant variation can be introduced during library preparation. This study provides important guidelines for planning future ATAC-seq experiments using equine FAANG tissues. We will use the guidelines established here to conduct ATAC-seq experiments on six other prioritized tissues in the mares. Furthermore, following these guidelines should enable the most meaningful integration of datasets across studies thus building a reliable functional tissue specific atlas of the equine genome which would advance our understanding of complex traits in the horse.

Data Availability

ATAC-seq data used in this study is available from the European Nucleotide Archive under the accession PRJEB41317. RNA-seq data from the liver and lamina tissues of the same two animals used in this study can be found from the European Nucleotide Archive under the accession PRJEB26787.

Disclosure: Funding was provided by the Grayson Jockey Club Foundation, USDA NRSP-8 and the UC Davis Center for Equine Health. Support for C.J.F. was provided by the National Institutes of Health (NIH) (L40 TR001136). None of the funding agencies had any role in the design of the study, analysis, interpretation of the data or writing of the manuscript. Salary support for S.P. was provided by the Ann T. Bowling Fellowship at the UC Davis Veterinary Genetics Laboratory.

Acknowledgments: The authors would like to acknowledge the Diagenode Epigenomic services team and Active Motif service team for providing partial financial support and completing the assays for this study. The preliminary results of this study were presented at Plant and Animal Genome conference in San Diego in 2019.

Supplementary Figure A1.1 Fragment size distributions of libraries from L1 as determined by sequencing and Fragment Analyzer.

Supplementary Figure A1.2 Fragment size distributions of libraries from L2 as determined by sequencing and tapestation.

Supplementary Figure A1.3 Duplication and mitochondrial contamination rates. **(A)** Total, mitochondrial, and unique nuclear read counts of all libraries; **(B)** Comparison between first sequencing run (left) and combined reads (right) from L1 liver CN libraries; **(C)** Fingerprint plot of L1 CN liver libraries.

Supplementary Figure A1.4 MACS2 peak calling statistics **A)** Number of peaks, **(B)** peak length distribution, **(C)** peak score distribution, and **(D)** percent of genome covered by peaks for each library. **(E-F)** Peak metrics assessed using ChIP-seq dataset in liver **(E)** and lamina **(F)** libraries.

References

1. Wade, C. M. et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science 326, 865–867 (2009).
2. Kalbfleisch, T. S. et al. Improved reference genome for the domestic horse increases assembly contiguity and composition. Commun. Biol. 1, 197 (2018).
3. Finno, C. J. & Bannasch, D. L. Applied equine genetics. Equine Vet. J. 46, 538–544 (2014).
4. Ghosh, M. et al. Transformation of animal genomics by next-generation sequencing technologies: a decade of challenges and their impact on genetic architecture. Crit. Rev. Biotechnol. 38, 1157–1175 (2018).
5. Raudsepp, T., Finno, C. J., Bellone, R. R. & Petersen, J. L. Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post‐genome era. Anim. Genet. 50, 569–597 (2019).
6. Qu, H. & Fang, X. A Brief Review on the Human Encyclopedia of DNA Elements (ENCODE) Project. Genomics Proteomics Bioinformatics 11, 135–141 (2013).
7. The FAANG Consortium et al. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. 16, 57 (2015).
8. Burns, E. N. et al. Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. Anim. Genet. 49, 564–570 (2018).
9. Giuffra, E., Tuggle, C. K., & FAANG Consortium. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. Annu. Rev. Anim. Biosci. 7, 65–88 (2019).
10. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC‐seq: A Method for Assaying Chromatin Accessibility Genome‐Wide. Curr. Protoc. Mol. Biol. 109, (2015).
11. 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).
12. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962 (2017).

13. Milani, P. et al. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. Sci. Rep. 6, 25474 (2016).
14. McIntyre, L. M. et al. RNA-seq: technical variability and sampling. BMC Genomics 12, 293 (2011).
15. Wylie, C. E., Collins, S. N., Verheyen, K. L. P. & Richard Newton, J. Frequency of equine laminitis: A systematic review with quality appraisal of published evidence. Vet. J. 189, 248–256 (2011).
16. Durham, A. E. et al. ECEIM consensus statement on equine metabolic syndrome. J. Vet. Intern. Med. 33, 335–349 (2019).
17. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
18. Krueger, F. Trim Galore! (2019).
19. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10 (2011).
20. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).
21. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
22. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034 (2015).
23. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160–W165 (2016).
24. Caswell, T. A. et al. matplotlib/matplotlib v3.1.3. (Zenodo, 2020). doi:10.5281/ZENODO.3633844.
25. Harris, C. R. et al. Array programming with NumPy. Nature 585, 357–362 (2020).
26. SciPy 1.0 Contributors et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272 (2020).
27. Reback, J. et al. pandas-dev/pandas: Pandas 1.1.3. (Zenodo, 2020). doi:10.5281/ZENODO.3509134.
28. Tarbell, E. D. & Liu, T. HMMRATAC: a Hidden Markov ModeleR for ATAC-seq. Nucleic Acids Res. 47, e91–e91 (2019).
29. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).
30. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27, 3423–3424 (2011).
31. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).
32. Kingsley, N. B. et al. Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. Genes 11, 3 (2019).
33. Cunningham, F. et al. Ensembl 2019. Nucleic Acids Res. 47, D745–D751 (2019).
34. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods 14, 417–419 (2017).
35. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research 4, 1521 (2015).
36. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).
37. Nothman, J. UpSetPlot. (2020).
38. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of Intersecting Sets. IEEE Trans. Vis. Comput. Graph. 20, 1983–1992 (2014).

39. Ackermann, A. M., Wang, Z., Schug, J., Naji, A. & Kaestner, K. H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol. Metab. 5, 233–244 (2016).
40. Liu, C. et al. An ATAC-seq atlas of chromatin accessibility in mouse tissues. Sci. Data 6, 65 (2019).
41. Foissac, S. et al. Multi-species annotation of transcriptome and chromatin structure in domesticated animals. BMC Biol. 17, 108 (2019).
42. Halstead, M. M. et al. A comparative analysis of chromatin accessibility in cattle, pig, and mouse tissues. BMC Genomics 21, 698 (2020).
43. Landt, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–1831 (2012).