4-2022

Annotating Gene Expression and Regulatory Elements in Tissues from Healthy Thoroughbred Horses and Identifying Candidate Mutations Associated with Perosomus Elumbis in an Angus Calf

Alexa Barber
University of Nebraska-Lincoln, alexa.barber@huskers.unl.edu

ANNOTATING GENE EXPRESSION AND REGULATORY ELEMENTS IN TISSUES FROM HEALHTY THOROUGHBRED HORSES AND IDENTIFYING CANDIDATE MUTATIONS ASSOCIATED WITH PEROSOMUS ELUMBIS IN AN ANGUS CALF

by

Alexa M. Barber

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Animal Science

Under the Supervision of Professor Jessica L. Petersen

Lincoln, Nebraska

April 2022

ANNOTATING GENE EXPRESSION AND REGULATORY ELEMENTS IN TISSUES FROM HEALHTY THOROUGHBRED HORSES AND IDENTIFYING CANDIDATE MUTATIONS ASSOCIATED WITH PEROSOMUS ELUMBIS IN AN ANGUS CALF

Alexa McKenna Barber, M.S.

University of Nebraska, 2022

Advisor: Jessica L. Petersen

Genome annotation has a direct impact on the success of genomic studies. Transcriptome analyses and chromatin immunoprecipitation and sequencing (ChIP-seq) have been used to functionally annotate genomes. These methods can identify protein-coding genes, non-coding transcripts, and cis-regulatory elements across the genome. The primary objective of the first study was to functionally annotate the equine genome through the assessment of nine tissues: adipose, brain, heart, lamina, liver, lung, skeletal, muscle, testis, and ovary. In the first project, 150 bp, paired-end RNA sequencing (RNA-seq) libraries were generated in stallion tissues and compared to previously generated mare RNA-seq libraries to quantify variation in gene expression due to sex and tissue type. On average, each tissue expressed (> 10 transcripts per million) over 8,000 genes, and adipose, liver, and skeletal muscle each had over 900 genes differentially expressed due to sex (P adj <0.05). In the second study, the peaks of four histone marks, H3K27ac, H3K4me1, H3K4me3, and H3K27me3, were examined to identify activated regions, enhancers, promoters, and silencers, respectively. Fifty base pair, paired-end ChIP-seq libraries were created for each histone mark in stallion tissues and compared to data from 50 bp single-end ChIP-seq libraries from mare tissues. On average, 77,000 activated

regions, 120,000 enhancers, 34,000 promoters, and 32,000 silenced regions were detected in each stallion tissue. Due to high correlations among sequencing depth, total peaks called, and tissue-unique peaks, regulatory elements unique to tissue types and sexes could not be well characterized.

The third study examined genomic variation associated with a congenital defect, perosomus elumbis, (PE) in Angus cattle. The affected calf was still-born, displaying lumbar aplasia, and arthrogryposis. Whole-genome sequencing of 31 Angus cattle identified a frameshift mutation in *PTK7* as a candidate variant for the development of PE in an Angus calf. Despite the implication of *PTK7* in similar phenotypes, additional research is needed to verify the etiology of PE in Angus cattle.

ACKNOWLEDGMENTS

To my advisor, Dr. Jessica Petersen: I appreciate the opportunities you gave me to work on a range of projects. Thank you for your patience as I developed my skills in epigenomic and bioinformatic research. I could not have accomplished what I have without your encouragement and support.

I greatly appreciate the guidance from my collaborators. To Dr. Carrie Finno, Dr. Rebecca Bellone, and Dr. Ted Kablfleisch: Thank you for your encouragement and help in analyzing data. Your expertise were immensely helpful in guiding me through the FAANG projects. To Nicole Kingsley: Thank you for taking the time to walk me through the ChIP-seq analyses and for your valuable input throughout the project. To Sichong Peng: I greatly appreciate your willingness to help me in every way you could.

Thank you to all the members of our lab group for your help during my time at UNL. To Anna Fuller: You are a great mentor inside the lab and out. You always made it enjoyable to be in the lab. Thank you for making me feel like I belonged. To Rachel Reith: I appreciate all of your help over the years. I've really enjoyed working beside you and learning from you. To Renae Sieck: Thank you for teaching me all sorts of things in the lab. I've really enjoyed our chats about life and work too.

To my friends and family: I would not have made it through my program without your love and support. I appreciate you being there for me through the good and bad days. Thank you for your patience and endless encouragement.

GRANT INFORMATION

The work in Chapters 2 and 3 was supported by the 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 work in Chapter 4 was partially supported by Angus Genetics Inc. All projects were completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

TABLE OF CONTENTS

CHAPTER 1: LITERATURE REVIEW

Equine genomics has vastly improved in the past two decades, yet the annotation of the equine genome lags behind that of mice and humans. The history of equine genomics following the development of the first reference genome is reviewed, including genome annotation, genomic tools, and major genomic studies. Some of the limitations of current genomic resources are discussed in horses and other livestock species. Many of these shortcomings in genome annotation have been addressed in the human genome by the ENCODE Project Consortium. The role of histone modifications in functionally annotating regulatory elements is discussed in addition to the major accomplishments of the ENCODE Project. Lastly, the Functional Annotation of Animal Genomes Project (FAANG), and the state of functional annotation in the equine genome are reviewed to provide context for the need of RNA-sequencing (RNA-seq) and chromatin immunoprecipitation and sequencing (ChIP-seq) studies in stallion horses.

Part I. Tools for Equine Genomics

Generation of the Equine Reference Genome

The Human Genome Project aimed to sequence the entirety of the human genome (Collins & Fink, 1995). This effort took over a decade to complete and resulted in the first reference genome for the study of human traits and diseases (International Human Genome Sequencing Consortium, 2001). The finalized reference genome covered about 99% of euchromatic regions and consisted of only 341 gaps but came at a cost of $450 million (International Human Genome Sequencing Consortium, 2004; Spencer, 2001). Despite the cost, the Human Genome Project had an immense impact on human health,

biotechnology, and the understanding of genomics as a whole (Hood & Rowen, 2013). In part due to the technological advances that were cultivated by the Human Genome Project, the cost of whole genome sequencing has dramatically declined, thus allowing for the development of reference genomes for other species of interest. One such species is the horse. The equine industry has had a large economic impact in the U.S. for decades, growing from a $39 billion industry in 2005 to a $50 billion industry in 2017 (American Horse Council, 2018). The study of disease and performance traits in these valuable animals was accelerated by the generation of the first equine reference genome, EquCab2.0 (Wade et al. 2009).

EquCab2.0 was generated using DNA from a single Thoroughbred mare, Twilight, and sequenced to 6.8x coverage using bacterial artificial chromosome (BAC) libraries and Sanger sequencing (Wade et al., 2009). The genomic sequence outlined in EquCab2.0 consists of 2.5-2.7 gigabases (Gb) with 95% of sequence being assigned to one of the 32 chromosomes (Wade et al., 2009). The equine genome was predicted to have 20,322 protein-coding genes with over 81% demonstrating orthology to human genes (Wade et al., 2009). While EquCab2.0 provided a solid foundation for equine genomic studies, the limitations of this reference genome have become apparent.

In 2015, the EquCab2.0 reference sequence was compared to the original 28 million Sanger reads used in the assembly and new 40x coverage Illumina short read data from Twilight (Rebolledo-Mendez et al., 2015). Over 1.9 million variants were identified between Twilight's Sanger reads and the reference genome with this number increasing to nearly 4 million when including the Illumina short read data (Rebolledo-Mendez et al., 2015). Of these variants, 4% were homozygous in the Sanger reads and 18% were

homozygous in the Illumina reads (Rebolledo-Mendez et al., 2015). Variants found in the homozygous state represent loci where neither of Twilight's alleles match the reference. This suggests that erroneous base calls could have contributed to the final reference sequence. Beyond these single nucleotide variants, 42,304 gaps that cover 2.2% of the genome were identified (Kalbfleisch et al., 2018). Between missing and miscalled bases, the limitations of EquCab2.0 warranted the generation of an improved reference genome.

Utilizing the foundation of EquCab2.0 and new sequencing approaches, such as PacBio long read sequencing and Hi-C proximity ligation, the newest equine reference genome, EquCab3.0, was published in 2018 (Kalbfleisch et al., 2018). Substantial improvements were observed in contiguity, completeness, and mapability. In this assembly, all but one chromosome is covered by a single scaffold and gaps were reduced from 2.2% (EquCab2.0) to 0.34% in EquCab3.0 (Kalbfleisch et al., 2018). Mapabilty also improved by 2.15% and 0.44%, for RNA sequencing (RNA-seq) and whole genome sequencing (WGS) datasets, respectively (Burns et al., 2018; Kalbfleisch et al., 2018). Unlike EquCab2.0, EquCab3.0 also phased haplotypes to select the most common allele shared amongst four other Thoroughbreds for the reference sequence at loci where Twilight is heterozygous (Kalbfleisch et al., 2018). Due to the more recent release of EquCab3.0, many of the genomic tools currently utilized in the equine community were developed using the EquCab2.0 reference genome.

Development of the Three Equine SNP Arrays

Genetic variation within a species plays an important role in phenotypic variation amongst individuals. Mutations in the genome can be associated with favorable and/or

deleterious traits. To explore variation in the equine genome, single nucleotide polymorphisms (SNPs) were identified in Twilight and representative horses from seven breeds. Twilight was heterozygous at ~750,000 loci and an additional ~400,000 SNPs were identified across the seven other breeds (Wade et al., 2009; McCue et al., 2012). On average, one SNP can be found every 2000 base pairs (bp) in the equine genome (Wade et al., 2009). These SNPs can be useful for tagging variation across the genome and identifying polymorphisms associated with traits of interest.

SNP arrays consist of many SNP loci that serve as markers for regions of variation in the genome. These arrays are dependent on the idea of linkage disequilibrium, or the concept that loci that are close together on a chromosome are inherited together at higher frequencies than expected if they were inherited independently (not linked). With LD present across the genome, the SNPs chosen for a SNP array are assumed to tag nearby variants that may be associated with traits of interest. Therefore, genotypes derived from these arrays are often used in genome wide association studies (GWAS).

An equine SNP array was developed in 2009 to allow for more affordable genotyping and genome-wide association studies. The EquineSNP50 BeadChip was developed using SNPs documented in the EquCab2.0 reference genome. This SNP chip contained ~53,500 SNPs that were validated by at least one heterozygous call amongst 351 horses that were successfully genotyped on the array (McCue et al., 2012). The SNPs fall approximately every 43 kilobases (kb) across the 31 autosomes and every 49 kb across the X chromosome (McCue et al., 2012). Assessing 14 breeds, McCue (2012) determined 49,603 (91.1%) of the SNPs on the array to be informative, or having a minor

allele frequency (MAF) greater than 0.05; however, the number of informative SNPs dropped substantially when analyzing some breeds alone (McCue et al., 2012). For example, only 37,053 or 68% of SNPs were informative in the Norwegian Fjord (McCue et al., 2012).

Shortly after the development of the EquineSNP50 BeadChip, the EquineSNP70 BeadChip was released consisting of approximately 46,000 informative SNPs from the 50K array and ~19,000 new SNPs (Chassier et al., 2018; Schaefer & McCue, 2020). The 70K array has one SNP approximately every 35kb and provides markers in gaps previously identified in the 50K array (Schaefer & McCue, 2020). Despite the improvements made in the EquineSNP70 BeadChip, the SNP density was only moderate, and areas of uneven genomic coverage remained (Schaefer et al., 2017). Furthermore, LD was found to decay rapidly and reached $r^2 < 0.2$ within 50kb when considering 14 horse breeds (McCue et al., 2012). Based on variation in haplotypes and LD across breeds, Wade et al. (2009) suggested that at least 100,000 SNPs would be required for effective GWAS.

The third and newest commercially available SNP array was made available in 2017 and contains 670,805 SNPs identified in the whole genome sequence of 153 horses across 24 breeds (Schaefer et al., 2017). These SNPs on the MNEc670K array were derived from a larger discovery array including over 2 million SNPs. The MNEc670K array was designed to include SNPs that tagged common haplotypes across the genome in 15 breeds (Schaefer et al., 2017). The number of SNPs required to recreate all-breed specific haplotypes varied by breed, with ponies, draft horses, and Quarter horses, requiring over 350,000 tagging SNPs and Thoroughbreds and Icelandic horses requiring

less than 150,000 tagging SNPs (Schaefer et al., 2017). The final MNEc670K array includes ~220,000 SNPs tagging haplotypes in four or more breeds, ~70,000 found on earlier SNP arrays, and ~7,000 SNPs in the highly studied major histocompatibility complex (MHC) region (Schaefer et al., 2017). The 670K array has at least 8 SNPs across every 50 kb region in the genome with approximately 3.7 kb between each SNP (Schaefer et al., 2017). Genome coverage was vastly improved by the 670K SNP array.

Genome Wide Association Studies in the Horse

All three equine SNP arrays have been applied to a wide variety of studies. Some applications include studies of fertility (Raudsepp et al., 2012; Gottschalk et al., 2016), racing performance (Binns et al., 2010), conformation (Singer-Hasler et al., 2012; Frischknecht et al., 2015), domestication (Schubert et al., 2014), and breed variation (McCue et al., 2012; Petersen et al., 2013). SNP arrays have also been employed to identify quantitative trait loci (QTL) associated with complex disease, such as osteochondrosis (OC) (Schaefer & McCue, 2020).

OC is characterized by the failure of ossification in the cartilage of growing bones and can impair the performance of young horses making it a disease of particular interest (McCoy et al., 2016). A variety of studies have associated genetic risk loci on multiple chromosomes with OC in Hanoverian warmbloods, Standardbreds, and Thoroughbreds (Dierks et al., 2007; Lampe et al., 2009; Lykkjen et al., 2010; Corbin et al., 2012; McCoy et al., 2016). Yet, little consensus among risk loci exists across studies. Although environment certainly contributes to OC, the high prevalence of the disease in certain breeds, like the Standardbred, suggests that genetic association efforts are not futile

(McCoy et al., 2016). Improvement of sequencing technologies and genome annotation may allow for the identification of functional candidate mutations among the QTLs identified by GWAS in the future.

In addition to complex traits, the equine SNP arrays have been useful in identifying loci associated with congenital defects. A study by Drögemüller (2014) into congenital hepatic fibrosis (CHF) of Franches-Montagnes identified a single SNP tagging a 952kb haplotype in affected horses that contained the polycystic kidney and hepatic disease 1 gene (*PKHD1). PKHD1* has been previously implicated in similar phenotypes in humans (Drögemüller et al., 2014). Despite the inability to identify a perfectly associated causative variant with subsequent whole genome sequencing, further research into the PKHD1 gene relative to hepatic fibrosis may be warranted (Drögemüller et al., 2014). Another condition that appears shortly after birth in affected horses is equine guttural pouch tympany (GPT). This disease is characterized by abnormal distention of air-filled tubes in the head of horses that results in labored breathing, difficulty swallowing, and pneumonia (Metzger et al., 2012). In studies of Arabians and German Warmbloods risk loci were identified on two different chromosomes, but putative causal mutations were not found (Metzger et al., 2012). These studies exemplify common outcomes of GWAS where genomic regions can be successfully associated with a trait but the function of the genome in that region to result in the studied outcome is not clear. In some cases, additional sequencing of associated genes has led to putative causal mutations within protein coding genes.

SNP arrays have proven successful in identifying strong functional candidates in some congenital defects when combined with Sanger sequencing and WGS. Lavender

Foal Syndrome (LFS) is a neurologic disorder accompanied by a coat color dilution present in Egyptian Arabian horses (Brooks et al., 2010). GWAS with SNP genotyping and subsequent Sanger sequencing, identified a single base pair deletion in the *MYO5A* associated with LFS (Brooks et al., 2010). *MYO5A* was linked to similar disorders in mice and humans, and the single base pair deletion in horses disrupted a highly conserved region of the gene (Brooks et al., 2010). The use of Sanger sequencing to examine GWAS hits provided a strong functional candidate for LFS (Brooks et al., 2010). Similar methods were used to identify a putative mutation for Naked Foal Syndrome (NFS) in Akhal-Teke horses (Bauer et al., 2017). Bauer (2017) used 670K SNP data to associate haplotypes with NFS. Subsequent whole genome sequencing of two cases and two controls identified a nonsense mutation in *ST14* associated with NFS (Bauer et al., 2017). *ST14* had previously been implicated in hair follicle development in mice making the nonsense mutation in *ST14* a strong functional candidate for the hairless phenotype of NFS foals (List et al., 2003; Bauer et al., 2017). Although some traits of interest have been successfully associated with strong functional candidates, putative causative variants have not been identified for many other important traits.

Limitations of Reference Genomes and SNP Arrays

Reference genomes and SNP arrays provide a strong foundation for genetic studies; however, there are limits to the usefulness of these resources. Differences between the population being studied and the reference sequence as well as shortcomings in the annotation of the reference can inhibit the productivity of genetic studies.

Structural variants in the genome often exist between breeds of the same species. One such example was demonstrated in cattle.

The most current *Bos taurus* reference genome, ARS-UCD1.2, is based on a Hereford cow that is highly inbred (Rosen et al., 2020). This high-quality reference genome contains a single scaffold for each of the 30 bovine chromosomes and only 459 gaps across the 2.6 Gb sequence (Rosen et al., 2020). Despite the 200-fold increase in continuity and 10-fold increase in accuracy compared to the previous reference from the same Hereford cow, UMD3.1.1 (Zimin et al., 2009), the ARS-UCD1.2 reference still has difficulties capturing variants of interest in distantly related cattle breeds (Rosen et al., 2020).

Development of reference genomes for other breeds of interest has been underway and demonstrates the genomic differences between breeds. The development of two haplotype-resolved reference genomes for Angus and Brahman cattle through trio binning identified genetic differences between Angus, Brahman, and Hereford cattle. The number of structural variants identified among six Brahman and five Angus cattle was dependent on whether the cattle were mapped to their corresponding breed's reference sequence (Low et al., 2020). As expected, a greater number of structural variants, such as deletions, duplications, and inversions, were observed when mapping cattle to the opposite breed's reference sequence (Low et al., 2020). When comparing the Brahman (UOA_Brahman_1) and Angus (UOA_Angus_1) reference genomes to the Hereford ARS-UCD1.2 reference genome, 0.4% of UOA_Angus_1 and 0.8% of UOA_Brahman_1 consisted of structural variants (indels, expansions, and contractions) compared to ARS-UCD1.2 (Low et al., 2020). The work by Low et al. (2020) demonstrate that genetic

variation does exist between cattle breeds and that reference genome selection impacts the ability to accurately identify variation across breeds.

Another study examining the impact of reference genome selection on WGS studies of Brown Swiss cattle identified differences in annotation between ARS-UCD1.2 and UOA_Angus_1. Few differences in mapping accuracy and SNP calling were identified when mapping WGS from Brown Swiss cattle to ARS-UCD1.2 and UOA_Angus_1; however, the annotation of variants using Ensembl's Variant Effect Predictor (VEP) showed significant differences between the two reference assemblies (Lloret-Villas et al., 2021). Nearly 10% more SNPs and indels were annotated as intergenic in ARS_UCD1.2 than UOA_Angus_1 (Lloret-Villas et al., 2021). SNPs and indels were found in intronic regions 10% more often in UOA_Angus_1 than ARS_UCD1.2 (Lloret-Villas et al., 2021). Minor differences were observed between the two reference genomes when considering variants in exons (Lloret-Villas et al., 2021). Signatures of selection occur when selection pressure results in the loss of variation at loci near causative variants and can be identified by alleles close to fixation or alleles recently fixed within a population (Lloret-Villas et al., 2021). When considering signatures of selection, 40 regions of selection were identified in ARS_UCD1.2 compared to 33 regions in UOA_Angus_1, but little overlap between loci was observed between the two references (Lloret-Villas et al., 2021). This demonstrates that reference genome can significantly impact the outcome of some genetic studies. Furthermore, the study by Lloret-Villas et al. (2021) suggests that a portion of chromosome 13 is inverted in the UOA_Angus_1 reference. These differences in annotation can be especially

inhibitory to studies where candidate variants are filtered by their predicted impact on gene function.

Genome wide association studies can also be limited by the annotation of reference genomes. Most reference genomes are annotated using a variety of methods, including comparing sequences from other species, utilizing transcriptome data, and *ab initio* gene prediction based on the sequence itself. Gene structures are frequently predicted by two main genome databases, National Center for Biotechnology Information (NCBI) and Ensembl. The predicted protein-coding gene lists for the EquCab2.0 equine reference genome consisted of 20,322 genes from Ensembl and 17,610 genes from NCBI (Coleman et al., 2010). To identify 5' and 3' ends as well as exons, introns, and splice junctions, Coleman et al. (2010) generated RNAseq data from 8 tissues to clarify the structure of protein-coding genes predicted by Ensembl and NCBI. After generating almost 300 million sequence tags, Coleman (2010) refined the structure of 11,356 genes. When considering loci that did not represent overlapping genes or pseudogenes, 89% of genes predicted by Ensembl and NCBI displayed expression in at least one of the eight studied tissues (Coleman et al., 2010). Ultimately, a consensus gene set consisting of 20,302 protein coding genes was defined; however, these protein coding genes only comprise about 1.28% of the genome (Coleman et al., 2010). A later study examining RNAseq data from 43 tissues in the horse identified 68,594 transcripts in which 71% of the transcripts overlapped previously annotated genes (Hestand et al., 2015). Of the 20,039 transcripts that did not align to previously annotated loci, over 90% contained a single exon, suggesting that some of these transcripts may correspond to non-coding

RNAs (ncRNAs), unannotated small open reading frames (smORFs), or gene fragments that were improperly constructed in the equine reference genome (Hestand et al., 2015).

In the past, annotation of genomes primarily focused on protein coding genes; however, of the ~20,000 protein coding genes found in most mammalian species, the protein coding sequence comprise less than 2% of the genome (Coleman et al., 2010; The ENCODE Project Consortium, 2012). This is particularly inhibitory to GWAS as one study found 88% of human trait associated loci fell within intronic and intergenic regions (Hindorff et al., 2009). Many other functional elements exist within the genome outside of protein-coding genes, including ncRNAs, transcription factor binding sites, transcriptional regulatory elements, and DNA methylation sites (The ENCODE Project Consortium, 2012). The large-scale annotation of these functional elements was undertaken by the Encyclopedia of DNA Elements (ENCODE) Project, and the discoveries made in this project have drastically changed our understanding of genome function.

Part II. Progress Toward Functional Annotation

The ENCODE Project: Overview and Pilot Phase

The human ENCODE Project began in 2003 with the intent of annotating all functional elements in the human genome and was subdivided into three phases: pilot, technology development, and production (The ENCODE Project Consortium, 2004). In the pilot phase, the consortium aimed to identify procedures that could accurately and economically characterize large portions of the human genome (The ENCODE Project Consortium, 2004). Concurrently, the technology development phase aimed to develop

laboratory and computational procedures to address the gaps in technology discovered in the pilot phase (The ENCODE Project Consortium, 2004). Between these two phases, the most efficient technologies and protocols for functional annotation would be determined to allow for a comprehensive and economical assessment of the entire human genome in the production phase (The ENCODE Project Consortium, 2004).

The pilot phase of ENCODE assessed suitable methods for large scale functional annotation by focusing on a 30 Mb region of the genome (~1%) split into forty-four 0.5-2 Mb regions (The ENCODE Project Consortium, 2004). The ENCODE Project Consortium (2004) manually chose approximately half of these regions to represent stretches of genome containing well characterized genes or regulatory elements with large amounts of comparative sequence data to leverage preexisting knowledge. The remaining target regions were selected with an algorithm that ensured selection of representative regions in terms of gene content and non-exonic conservation between humans and mice (The ENCODE Project Consortium, 2004). The consortium examined a variety of technologies in the pilot phase including transcript microarray assays, chromatin immunoprecipitation microarray hybridization (ChIP-chip), computational gene calling methods, and expression reporter assays. These technologies were employed with the intent of identifying genes and their cis-regulatory elements (promoters, enhancers, repressors, and silencers), transcription start and end sites, transcription factor binding sites, DNA methylation sites, accessible chromatin, chromatin modifications, and conserved regions across species (The ENCODE Project Consortium, 2004).

A synthesis of the results from the pilot phase were published by the ENCODE Consortium in 2007, including transcriptome analyses, novel transcription start site

annotation, regulatory element identification, DNA replication regulation, and evolutionary constraint analysis across mammalian species (The ENCODE Project Consortium, 2007). Only 2% of transcripts identified were found in all sample types, whereas 40% of transcripts were found in only one sample type (The ENCODE Project Consortium, 2007). Rapid amplification of cDNA ends (RACE) was used to clarify the 5' ends of transcripts. RACE products were hybridized with tiling arrays and added to complement the previous datasets. Over 70% of the bases in the ENCODE region were contained within unspliced, primary transcripts that were identified in multiple assays (The ENCODE Project Consortium, 2007). When assessing transcription start sites (TSSs), ~2,700 novel TSSs were identified and supported by a similar presence of transcription factors, histone modifications, and DNase I accessibility at known and novel TSSs (The ENCODE Project Consortium, 2007). These chromatin structural modifications were also found to be able to predict the location and activity of TSSs with up to 91% accuracy (The ENCODE Project Consortium, 2007). The presence of histone modifications was also correlated with the signal of DNA replication. Activating histone modifications such as histone 3 acetylation and histone 3 lyisne 4 mono- and tri-methylation were negatively correlated with replication signals, while repressive modifications, such as histone 3 lysine 27 trimethylation, were positively correlated with replication signals (The ENCODE Project Consortium, 2007). The presence of activating histone marks fell within open chromatin between 81-93% of the time as noted by DNase I hypersensitive sites (DHSs); however, 29-57% of DHSs lacked activating histone modifications (The ENCODE Project Consortium, 2007). To assess if these regulatory elements were in conserved regions of the genome, regions of evolutionary constraint

were examined. Evolutionary constraint is defined as regions of the genome that reject mutations and that can be identified by assessing the frequency of intraspecies polymorphisms (The ENCODE Project Consortium, 2007). Approximately 50% of non-coding functional elements are found in unconstrained regions across mammals, suggesting variation in these elements both within and between species (The ENCODE Project Consortium, 2007).

Overall, the pilot phase of ENCODE, studying just 1% of the human genome, provided a wealth of new knowledge regarding genome function. As much as 74% of the genome is transcribed and transcription is tissue-specific in many cases. The presence of transcription factors and histone modifications were symmetrical around TSSs, suggesting the functional relevance of regulatory elements both upstream and downstream of the TSS (The ENCODE Project Consortium, 2007). Further evidence was provided to suggest that histone modifications can be used to identify regulatory elements and transcriptional activity. Lastly, almost half of functional non-coding elements were located within non-conservative regions of the genome, warranting further study of these regulatory elements both across tissue types and in other species. The findings of the pilot phase foreshadowed the impact of the overall ENCODE project which has resulted in over 7,400 published studies to date.

Transcriptome Analyses in the ENCODE Project

Transcription results in RNA products that collectively makeup the transcriptome. The transcriptome can be defined by just the messenger RNAs (mRNAs) produced from the transcription of protein coding genes or as all RNAs within the cell, including non-

coding RNAs. As RNA is transcribed by RNA polymerase II, a modified guanine cap is added to the 5’ end of the RNA which functions to prevent its degradation. Messenger RNA is also polyadenylated at the 3’ end, and both the 5’ cap and 3’ poly-A tail are involved in translation (Gertsel et al., 1992). Long non-coding RNAs (lncRNAs) can also be modified to include 5’ caps and 3’ poly-A tails (Guttman et al., 2009). Not all lncRNAs, however, are polyadenylated (Cheng et al., 2005; Yang et al., 2011). Various technologies leverage these post-transcriptional modifications for transcriptome analyses. For example, cap analysis gene expression (CAGE) captures transcripts by targeting the 5’ cap and poly-A tail selection is often used to focus sequencing effort on mRNA transcripts (Carninci et al.,1996; Zhao et al., 2018).

The early phases of the ENCODE project used a variety of technologies to characterize the transcriptome, including CAGE, RNA paired end tagging (PET), and tiling arrays (The ENCODE Project Consortium, 2007). The methods used in the later phases of the ENCODE project employed newer technologies, such as massively parallel sequencing, capable of generating larger amounts of data. A newer approach to transcriptome analysis is RNA-seq which employs next-generation sequencing (NGS). RNA-seq allows for an in depth look at gene structure as transcripts are fragmented and fitted with adapters on one or both ends of the fragment. Between 30 and 400 base pairs are sequenced from each adapter creating reads that map across the length of the transcript (Wang et al., 2009). RNA-seq provides an advantage over earlier technologies as it can help identify intron/exon boundaries and different isoforms associated with a gene (Wang et al., 2009).

A large degree of the understanding of transcription across the genome is derived from the results of the ENCODE project. In the production phase where the entirety of the human genome was assessed, Djebali (2012) explored transcription in 15 cell lines using a variety of methods. Within the 15 cell lines, 62 and 74% of the genome was contained within processed and primary transcripts, respectively (Djebali et al., 2012). The processed transcripts from a single cell line covered 22% of the genome on average and no single cell line possessed more than 57% of the transcripts identified across all 15 cell lines (Djebali et al., 2012). Djebali (2012) determined approximately 50% of protein-coding transcripts to be ubiquitously expressed in all 15 cell lines, while only 7% were cell line specific. Djebali (2012) found the opposite to be true for lncRNAs of which nearly 30% of transcripts were cell line specific and only 10% were found in all studied cell lines. Many genes can produce multiple isoforms, yet a single isoform generally comprises the majority of transcripts in a given condition (Djebali et al., 2012). Some distal enhancer sequences were found to be transcribed, but both the degree of transcription and chromatin modifications associated with the enhancer regions were found to be cell line specific (Djebali et al., 2012). Djebali's 2012 findings increased the GENCODE annotation of the human genome to include 45% more transcripts and 80% more genes, many of which were mono-exonic. Overall, this work emphasized the importance of assessing the transcriptome across a variety of cell types and demonstrated that the majority of the genome is contained within primary transcripts. The findings from the ENCODE project have been integrated into an annotation called GENCODE.

The GENCODE annotation is a gene set comprised of genes manually annotated by the Human and Vertebrate Analysis and Annotation (HAVANA) group and genes

automatically annotated by Ensembl (Harrow et al., 2012). The GENCODE 7 release, published alongside the ENCODE paper, summarizing the results from the production phase, included 20,687 protein-coding genes, 9,640 lncRNAs, and approximately 10,000 pseudogenes (Harrow et al., 2012). Over 140,000 alternative transcripts were proposed in the GENCODE gene set compared to the RefSeq and UCSC annotations, yet many of these alternative transcripts were missing either their 5' or 3' ends (Harrow et al., 2012). When assessing the predicted exon-exon junctions in transcripts and the presence of novel transcripts, RT-PCR and sequencing validated 82% of the identified exon-exon junctions and 73% of novel transcripts (Harrow et al., 2012). The GENCODE 7 release provided a solid foundation for genomic studies in humans, yet the GENCODE annotation has been consistently updated since its original release. As of 2021, the GENCODE gene set consists of 19,954 protein coding genes, 17,957 lncRNAs, 14,767 pseudogenes, 7,569 small RNAs, 645 immunoglobin/T cell receptor genes, and over 230,000 transcripts (Frankish et al., 2021). Overall, the ENCODE project significantly impacted genomics studies by emphasizing the variation in gene expression present across cells and tissues and the pervasive presence of non-coding genes in the genome. The findings from this project have inspired transcriptomic studies in other species, including the horse.

Transcriptome Studies in the Horse

In the horse, most transcriptome studies have been limited in the tissues assayed either determining differential expression associated with traits of interest or improving the annotation of the reference genome. Studies aimed at determining the impact of

exercise on gene expression have assessed the transcriptome of blood and skeletal muscle in Thoroughbred and Arabian horses (McGivney et al., 2010; Park et al., 2012; Capomaccio et al., 2013; Ropka-Molik et al., 2017). Some studies focused on differential gene expression between untrained and trained muscles (McGivney et al., 2010; Ropka-Molik, 2017), while others assessed changes in gene expression immediately following exercise ranging from 30 min of trotting (Park et al., 2012) to endurance races of nearly 100 miles (Capomaccio et al., 2013). Genes involved in the immune system, the cell cycle, signal transduction, and lipid metabolism were found to be differential expressed immediately following exercise (Park et al., 2012; Capomaccio et al., 2013). Genes displaying differential expression following long term training include those involved in metabolism, muscle growth and development, and mitochondrial function (McGivney et al., 2010; Ropka-Molik et al., 2017). Some of these studies also recognized transcripts in unannotated regions of the genome suggesting both the presence of noncoding genes and the limitations of the EquCab2.0 reference annotation at the time they were published (Park et al., 2012; Capomaccio et al., 2013).

Further studies have examined the transcriptome relative to reproduction. Sperm were previously believed to have minimal transcriptomes reflecting that of the testis; however, Das and others (2013) identified 202 transcripts in sperm that were not expressed in the testis. Further examination of the sperm transcriptome using RNA-seq, identified over 19,000 transcripts present in the sperm compared to ~6,600 identified using a microarray (Das et al., 2013). Many of the transcripts consisted of micro RNAs (miRNAs), genes on the Y chromosome, and those involved in sperm specific functions (Das et al., 2013). Although ~13,000 transcripts did not align to annotated elements in

EquCab2.0, Das (2013) demonstrated the complexity of the sperm transcriptome. Iqbal and colleagues (2014) characterized the transcriptomes of equine embryonic cells and assessed differential expression between the inner cell mass (ICM) and trophectoderm (TE). Over 10,000 genes were expressed in both ICMs and TE with 1201 transcripts exclusive to ICM and 705 transcripts unique to TE (Iqbal et al., 2014). Genes overexpressed in ICM were related to cell differentiation, cell migration, and regeneration while genes overexpressed in TE corresponded to placental development and cellular transport (Iqbal et al., 2014). Together, these studies demonstrate some of the benefits of RNA-seq technology, including greater detection of transcripts and the ability to generate high-quality RNA libraries from cells within a single embryo; however, the lack of progression in genome annotation prevented the characterization of over two-thirds of the transcripts expressed in sperm (Das et al., 2013).

Gene expression in the equine immune system has been assessed in various lymphoid tissues and leukocytes. One study assessed the transcriptome related to six immune related cells and tissues, including lymphocytes, spleen, lymph node, liver, jejunum, and kidney. The consensus transcriptome from these cells and tissues showed little overlap with Ensembl's annotation of gene structure, and over 8,000 novel isoforms were identified (Moreton et al., 2014). Furthermore, 91 families of paralogs were expanded in the horse compared to the human, with 83 of 91 determined to be simple duplications (Moreton et al., 2014). Peripheral blood mononuclear cells (PBMCs) are another form of immune cell that modulate both the innate and adaptive immune system. Examining the transcriptome of 561 PBMC cultures from 85 Warmblood horses, 42,602 predicted genes were expressed (Pacholewska et al., 2015). Over 7,500 unannotated

transcripts were identified with 57% demonstrating homology to expressed sequence tags found in other species (Pacholewska et al., 2015). This study identified 543 novel transcripts with high coding potential with 61 of these novel coding genes unique to the horse (Pacholewska et al., 2015). Both studies provided information regarding gene expression in immune cells and tissues and once again identified opportunities to improve the annotation of EquCab2.0.

Many other equine transcriptome studies have taken place. One study assessed the differential gene expression between the Korean Jeju horse and Thoroughbred horses. Five tissues were selected for the comparison including skeletal muscles of the rump and thigh, liver, heart, and lung (Srikanth et al., 2019). Over 5,400 genes were differentially expressed between Jeju and Thoroughbred tissues with genes involved in process such as angiogenesis and cell adhesion, muscle cell differentiation, fat metabolism, and molecular signaling pathways (Srikanth et al., 2019). Of the differentially expressed genes, 71 were in regions identified as signatures of selection between the two breeds including genes related to body size, muscle fiber type, and mitochondrial function (Srikanth et al., 2019). Another unique study examined the difference in gene expression between fetal, adult, and embryonic stem cell derived tenocytes. This study identified 542 differentially expressed genes between the fetal and adult tenocytes when cultured in 3D; however, only 10 genes were differentially expressed when the tenocytes were cultured in a 2D monolayer (Paterson et al., 2020). This is interesting as it suggests that different methods of cell culture can greatly impact gene expression and that the transcriptome of cultured cells may not reflect genes expressed *in vivo*. These studies exemplify the range of applications of transcriptome analysis.

In 2017, equine transcriptome data available from a variety of sources was compared to a new transcriptome derived from 59 samples across 8 tissues (Mansour et al., 2017). The transcriptome built by Mansour et al. (2017) matched most closely with that derived from PBMCs by Pacholewska et al. (2015) followed by the annotation available from NCBI. About 50% of the 76,125 transcripts identified by Mansour et al. (2017) were shared across the transcriptomes generated by Pacholewska et al. (2015), Hestand et al. (2015), NCBI, and Ensembl (Mansour et al., 2017). The study by Mansour et al. (2017) also assessed differences in gene expression across tissues; however, there is likely technical bias present in this comparison as different methods were used to generate libraries across tissues, including single-end vs paired end reads, rRNA depletion vs polyA+ selection, and stranded vs unstranded libraries (Mansour et al., 2017). Work utilizing the transcriptome from Mansour et al. (2017) assessed lncRNAs in the equine genome (Scott et al., 2017). Nearly 21,000 lncRNAs were identified across 8 tissue types; however, more lncRNAs likely exist due to the fact that lncRNAs are frequently expressed in a tissue-specific manner (Djebali et al., 2012; Scott et al., 2017). Altogether, these studies provided expansion and validation of the annotation of the equine genome.

Together, these studies of the equine transcriptome have bolstered the annotation of the equine genome and provided a foundation for the study of differential gene expression across tissues, breeds, and phenotypes. Despite the advancements made via these studies, limitations in experimental design and methodology have prevented an unbiased comparison of gene expression across a wide variety of tissues. Many studies have examined only a single tissue or cell type, while others have combined numerous

tissue samples to create a single transcriptome. In studies where gene expression across tissues was compared, differences in library preparation and read depth have confounded biological differences between tissues (Mansour et al., 2017). Additional research that prioritizes the reduction of technical bias will be necessary to better understand tissue-specific gene expression. Although transcriptomic studies can identify differences in gene expression, not all cis-regulatory elements modulating these changes can be identified through RNA-seq. Other methods, such as chromatin immunoprecipitation and sequencing (ChIP-seq), can be utilized to identify histone modifications associated with cis-regulatory elements such as enhancers, promoters, and polycomb repressors.

Overview of Histone Modifications

DNA within the cell is condensed and stored as chromatin. The basic unit of chromatin is the nucleosome which consists of 147 bp of DNA coiled around histone proteins. Nucleosomes are connected by short stretches of DNA, termed linker DNA, which creates the primary chromatin structure resembling beads on a string (Figure 1.1). Each nucleosome organizes about 200bp of DNA when accounting for both the wrapped and linker DNA (McGinty & Tan, 2015). Folding and coiling of the primary chromatin creates secondary and tertiary chromatin structures that are highly compact. Most transcriptionally active regions of the genome are believed to fall in loosely packaged DNA, termed euchromatin, while transcriptionally repressed regions of the genome are believed to exist mostly in the condensed heterochromatin (Morrison & Thakur, 2021). Beyond chromatin unpacking, nucleosomes are often temporarily removed in regions of active gene expression (Lee et al., 2004). Although gene expression is more common in

areas of loosely packed DNA, some studies suggest that transcription can still occur in regions of tertiary chromatin (Zhou et al., 2007; Hu et al., 2009). Both the location of nucleosomes and the compaction of chromatin are dynamic allowing for changes in transcriptional programs. The location and degree of chromatin compaction can be altered by the binding of various proteins including chromatin remodeling ATPases, transcription factors, and histone modifying enzymes (Mellor, 2005). Here, we will focus on the impact of various histone modifications on transcription.

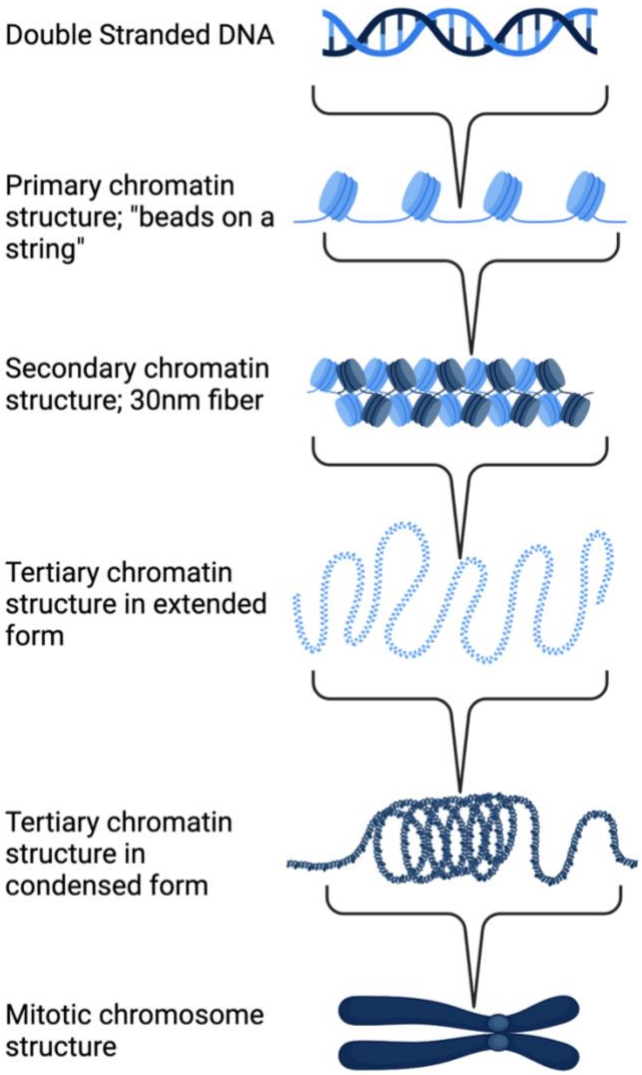


Figure 1.1. DNA Packaging and Chromatin Structure
The primary structure of chromatin is the most accessible (euchromatin) and resembles beads on a string. The secondary structure is defined as the 30-nm fiber. The tertiary structure refers to the looping and coiling of the chromatin fiber. The secondary and tertiary structures are referred to as heterochromatin. Figure created with BioRender.com

Each nucleosome contains a histone protein of eight subunits with two copies each of H2A, H2B, H3, and H4 (Arents et al., 1991). Each histone subunit has a N-terminal tail that protrudes from the nucleosome core consisting of 25-59 amino acids (Grant, 2001; Nurse

et al., 2013) (Figure 1.2). The N-terminal tails of the H4 and H2A subunits fall outside the nucleosome while the N-terminal tails of the H2B and H3 subunits fall between the DNA gyres on the nucleosome (McGinty & Tan, 2015). The N-terminal tails of histones play a key role in chromatin compaction. Histone tails can interact with DNA and other histone cores to pull nucleosomes together (Arya et al., 2009; Nurse et al., 2013). The positive charge on the histone tails helps overcome the repulsive forces generated between the negatively charged DNA allowing for tighter packing of nucleosomes (Arya et al., 2009). Studies have found that the removal of histone tails impairs oligomerization of nucleosomes and demonstrated that the N-terminal tails of all histone subunits are involved to some degree in the assembly of chromatin structures (Tse & Hansen, 1997; Gordon et al., 2005; Nurse et al., 2013). Therefore, it is not surprising that chemical modifications to the N-terminal tails impact both chromatin structure and transcription. Many post-translational modifications to lysine and arginine residues along the N-terminal tails have been identified. The impact of acetylation and methylation along histone tails was first explored by Allfrey et al. (1964). Deposition of acetyl and methyl groups along histone tails can impact transcription through alteration of the electrostatic environment of the nucleosome and the recruitment of various transcription factors (Bannister & Kouzarides, 2011). The biological role of four commonly studied histone modifications and the techniques used to study their locations will be discussed.

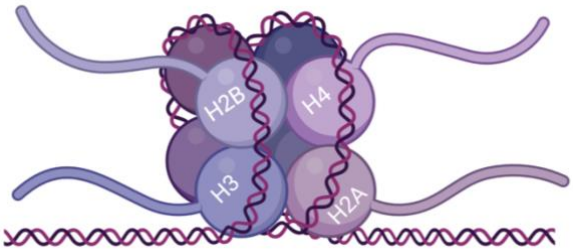


Figure 1.2. Nucleosome Structure
A drawing depicting the octamer structure of a nucleosome. Each subunit has N-terminal tail extending from the nucleosome. Figure created with BioRender.com

The common nomenclature for post translational histone modification includes naming modifications with the histone subunit (e.g.H3), the residue (e.g. lysine 27 or K27), and the chemical modifier (e.g. ac for acetylation or me3 for trimethylation). Although many modifications have been identified across all 4 types of histone subunits, the focus of this review will include 4 commonly studied H3 modifications (H3K27ac, H3K4me1, H3K4me3, and H3K27me3). The N-terminal tail of H3 is the longest of all the subunits, consisting of 59 amino acids, and is of particular interest as the histone tail interacts directly with nucleosomal DNA (Nurse et al., 2013).

Histone Acetylation: H3K27ac

Acetylation of various components of histone proteins and its impact on genomic function have been studied since the 1960s. Acetylation of histone proteins is regulated by histone acetyltransferases (HATs) and histone deacetylases (HDACs). HATs place acetyl groups on histone proteins and can be separated into two categories, Type A HATs and Type B HATs. Type B HATs are involved in acetylating histone subunits in the cytosol, while Type A HATs are found in the nucleus and acetylate histones already in nucleosomes (Gujral et al., 2020). HATs use acetyl CoA as a cofactor to transfer acetyl groups to the lysine residues of the histone tails (Bannister & Kouzarides, 2011). Common Type A HATs include p300, CBP, and GCN5 (Bordoli et al., 2001; Xue-Franzén et al., 2013; Gujral et al., 2020). HDACs function to remove acetyl groups from both histone proteins and nonhistone proteins (Seto & Yoshida, 2014). The presence of HDACs were first identified in 1969 by Inoue & Fujimoto, but the first HDAC was not isolated and cloned until 1996 (Inoue & Fujimoto, 1969; Taunton et al., 1996). Since

their discovery, 18 HDACs have been identified in humans (Seto & Yoshida, 2014). HDACs use either zinc or NAD(+) to remove acetyl groups from proteins (Finnin et al., 1999; Imai et al., 200; Seto & Yoshida, 2014). Together, HATs and HDACs regulate chromatin structure and transcription.

Acetylation of both the histone core and histone tails can impact nucleosome structure and chromatin folding. Acetylation of the histone cores and tails reduces the number of turns of DNA (Bauer et al., 1994) and induces a reduction in the degree of chromatin folding independent of H1 linker binding protein, which functions to stabilize the nucleosome (Garcia-Ramirez et al., 1995; Wang et al., 2001). The reduction in chromatin folding resulting from acetylation of the histone tails is proposed to be due to a partial neutralization of the positive charge on the histone tail that impairs the interaction of the tail with the linker DNA between nucleosomes (Garcia-Ramirez et al., 1995; Bannister & Kouzarides, 2011). Although the chemical effect of acetylation of individual residues in histone tails has not been thoroughly explored, the impact of H3K27ac on transcription and gene function has been examined closely.

Histone acetylation was implicated in increasing transcriptional activity as early as 1964 (Allfrey et la., 1964). Many studies broadly examined H3 acetylation in relation to gene expression and regulation. The initial phase of the ENCODE project identified enrichment of H3 acetylation at the TSS of genes and demonstrated greater enrichment at genes that were active and near CpG islands (The ENCODE Project Consortium, 2007). H3K27ac, among other H3ac, was also demonstrated to localize at TSSs (Wang et al., 2008). Enhancers are another cis-regulatory element that play a key role in increasing gene expression. The presence of H3K27ac has been shown to differentiate active

enhancers from poised enhancers (Creyghton et al., 2010). Furthermore, H3K27ac is often identified at enhancer clusters, termed super-enhancers (Hnisz et al., 2013). A recent study demonstrated that p300/CBP mediated acetylation is directly responsible for activating enhancers and initiating transcription at enhancer-regulated genes (Narita et al., 2021). The study determined that acetyltransferase activity was required for transcription factor and RNA polymerase II recruitment to enhancer-regulated genes (Narita et al., 2021). CBP has been shown to be essential for H3K27ac, while knock down of GCN5 showed little effect on H3K27ac (Tie et al., 2009). This suggests that H3K27ac, among other modifications, plays a key role in enhancer activation and gene expression. Furthermore, the presence of H3K27ac without modifications frequently found at enhancers (H3K4me1) demonstrated similar expression levels to regions containing both H3K27ac and H3K4me1 (Creyghton et al., 2010). Together, these studies provide strong evidence to support that H3K27ac is involved in activating gene expression across the genome.

Histone Methylation: H3K4me1, H3K4me3, & H3K27me3

Similar to histone acetylation, histone methylation was first explored in the 1960's (Allfrey et al., 1964; Murray et al., 1964), and the enzymes responsible for the placement and removal of methyl groups have been well characterized. Lysine methyltransferases (KMTs) are the enzymes responsible for adding methyl groups to lysine residues in histone tails, such as H3K4 and H3K27. KMTs such as SETD1A/B, MLL1-4, SETD7, and PRDM9 are responsible for mono- and tri-methylation of H3K4me3, and EZH1 & EZH2, within the polycomb repressive complex 2 (PCR2), are

involved in trimethylation of H3K27me3 (Husmann & Gozani, 2019). Methyl groups are transferred from other molecules, such as *S*-adenosyl-L-methionine, to the lysine residues (Kwon et al., 2003). Removal of the methyl groups is performed by lysine demethylases, of which two types exist: LSD demethylases and JMJC demethylases (Kooistra & Helin, 2012). LSD demethylases, including LSD1, can remove mono- and di-methylation at H3K4 and H3K9 (Shi et al., 2004; Kooistra & Helin, 2012), but are incapable of removing trimethylation. JMJC demethylases, including RBP2 and JMJD3, remove trimethylation at lysine residues, including H3K4 and H3K27 (Christensen et al., 2007; De Santa et al., 2007; Lan et al., 2007). The placement and number of methyl groups added to histone tails is tightly regulated and involved in both gene activation and repression.

H3K4me3 are associated with active promoters and has been identified at the TSS of genes across multiple species (Santos-Rosa et al., 2002; Schneider et al., 2004; The ENCODE Project Consortium, 2007; Barski et al., 2007; Schuettengruber et al., 2009). Santos-Rosa et al. (2002) suggested that H3K4me3 was present at the promoter of active genes and absent from inactive genes in yeast; however, this pattern does not hold true in humans. A study by Barksi et al. (2007) identified H3K4me3 marks associated with TSSs, some enhancers, and silent promoters. Over 90% of the genome found to be associated with RNA polymerase II also overlapped H3K4me3 (Barski et al., 2007). Additional research supports the presence of H3K4me3 at both active and inactive promoters (Schneider et al., 2004). Barksi et al. (2007) also identified colocalization of H3K4me3 with repressive H3K27me3, which resulted in lower expression at these loci. Although most studies have identified H3K4me3 to be enriched at active promoters, there

is mixed evidence as to whether H3K4me3 can predict the activity of the genes to which it localizes (Schneider et al., 2004; Barski et al., 2007).

Most studies have not shown a direct role of H3K4me3 on transcription, but rather an indirect role through the recruitment of chromatin remodeling enzymes and transcription factors, such as CHD1 and NURF (Lee et al., 2004; Flanagan et al., 2005; Sims III et al., 2005; Li et al., 2006; Parvi et al., 2006; Wysocka et al., 2006; Sims III et al., 2007). An indirect role of H3K4me3 also supports previous observations of H3K4me3 at both active and inactive promoters (Schnieder et al., 2004; Barski et al., 2009). Although H3K4me3 alone may not be indicative of active expression, H3K4me3 is a useful mark for identifying promoters.

H3K4me1 has also been identified at promoter regions (Barski et al., 2007). H3K4me1 is often found in a bimodal pattern around the TSS of active genes (Barski et al., 2007; Heintzman, et al., 2007; The ENCODE Project Consortium, 2007; Bae et al., 2020); however, Cheng et al. (2014) demonstrated that H3K4me1 may have repressive properties when occupying the TSS in the absence of H3K4me3, and Bae et al. (2020) determined H3K4me1 to predict poised promoters when present at the TSS in a unimodal pattern. Distal enhancer regions are commonly occupied by H3K4me1 with H3K4me3 less abundant at these loci (Heintzman et al., 2007; Heintzman et al., 2009; Creyghton et al., 2010). The presence of H3K4me1 is highly cell type specific (Koch et al., 2007; Heintzman et al., 2009), which may contribute to variation in the number and location of H3K4me1 marks identified across studies. Active enhancer regions can also be inhabited by H3K27ac which is likely due to the interaction of the enzymes that deposit H3K4me1 (MLL3/MLL4) and H3K27ac (CBP/p300) (Zhang et al., 2020; Lai et al., 2017); however,

the presence of H3K27ac is not required for H3K4me1 marked enhancers to be active (Creyghton et al., 2010; Zhang et al., 2020). Similar to the function of H3K4me3, H3K4me1 modulates transcription through the binding of chromatin remodeling enzyme, such as BAF (Local et al., 2018). In addition to BAF, H3K4me1 modulates the recruitment of the cohesion complex which is directly involved in enhancer promoter interactions (Kagey et al., 2010; Yan et al., 2018). H3K4me1 is a useful marker for identifying enhancers in the genome, especially since these active regions can lie thousands of base pairs away from the genes they regulate (Heintzman et al., 2007).

Unlike the previously discussed histone modifications, H3K27me3 is primarily associated with repressed, or transcriptionally silent regions of the genome (Boyer et al., 2006; Barski et al., 2007; Hosogane et al., 2016). H3K27me3 can also be present with activating marks, such as H3K4me3, to create bivalent or poised promoters (Berstein et al., 2006); however, PCR2, the enzyme that deposits H3K27me3, is suppressed by activating histone marks, such as H3K4me3 and H3K27ac (Schmitges et al., 2011). As with the other histone methylation, H3K27me3 impacts transcription through the recruitment of chromatin remodeling enzymes. H3K27me3 attracts cPRC1 and BAH which induce chromatin compaction and create regions of facultative heterochromatin (Bierne et al., 2009; Grau et al., 2011; Isono et al., 2013). Facultative heterochromatin refers to condensed regions of the genome associated with H3K27me3 and silenced genes (Bierne et al., 2009). Unlike constitutive chromatin that exists at the centromere and telomeres, facultative heterochromatin is more dynamic. Regions silenced by H3K27me3 cover broad ranges of the genome as binding of PCR2 to H3K27me3 increases deposition of H3K27me3 in the surrounding histones until active regions are reached (Oksuz et al.,

2018; Schmitges et al., 2011). Proper regulation of polycomb repression is imperative to both embryonic development and the maintenance of adult stem cells (Boyer et al., 2006; Lee et al., 2006; Bogliotti et al., 2012; Koppens et al., 2016). Although other histone modifications are involved in gene repression, H3K27me3 is the most commonly assayed modification for identifying repressed regions of the genome.

The presence of all four histone modifications, H3K27ac, H3K4me3, H3K4me1, and H3K27me3, can be examined independently to quantify enhancers, promoters, and silencers; however, the overlap of these histone modifications is often assessed to create a comprehensive atlas of chromatin states. This method can be particularly helpful in determining the activity levels of regulatory elements, as many studies have concluded that "active" marks such as H3K4me3 and H3K4me1 can be found in both active and inactive promoters and enhancers (Schnieder et al., 2004; Parvi et al., 2006; Barski et al., 2009). The ENCODE project was one of the first studies to assess these four histone marks across a range of cell lines. This project has established a foundation for the methods of assessing such histone marks and has provided immense support for the value of exploring histone modifications.

The Role of ChIP-seq in Functional Annotation

Two main methods were used to assess histone modifications in the ENCODE project: ChIP-chip and ChIP-seq (Figure 1.3). Chromatin immunoprecipitation (ChIP) allows for the identification of DNA-protein interactions by crosslinking the DNA and protein. The DNA is then sheared to create small fragments that can be bound by antibodies specific to histone modifications. These antibodies are used to pull down the

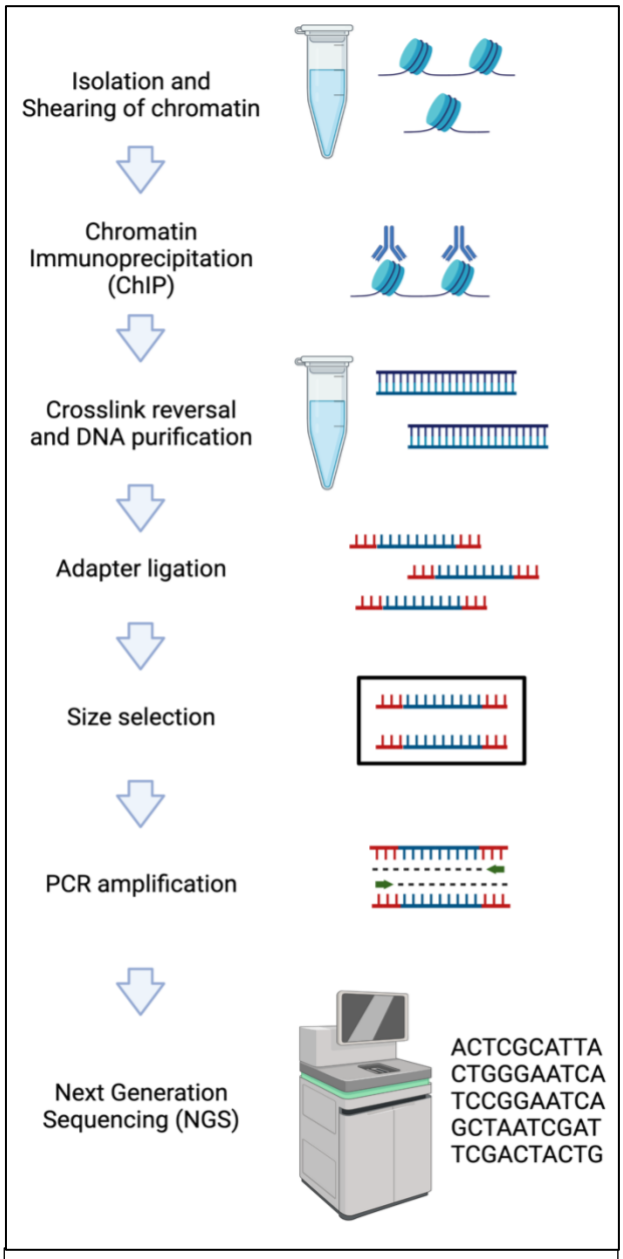


Figure 1.3. Library Preparation for ChIP-seq Experiments
Proteins are cross-linked to DNA and chromatin is isolated and sheared. Antibodies are used to precipitate regions of DNA associated with specific proteins. Cross-links are reversed and DNA is purified and amplified for sequencing using either next generation sequencing (NGS) (ChIP-seq) or microarrays (ChIP-chip). Figure created with BioRender.com

DNA sequences associated with the histone modification after which the crosslinking is reversed. The only difference between ChIP-chip and ChIP-seq is how the DNA associated with the histone modifications is assessed. In ChIP-chip, the DNA fragments are hybridized to microarrays with a fixed number of oligo binding sites. ChIP-seq employs next generations sequencing which can allow for more precise mapping of histone modifications due to the ability of the reads to map to any location in the genome. ChIP-chip and ChIP-seq libraries are often corrected for noise by using input samples that undergo crosslinking and fragmentation, but not immunoprecipitation. Both techniques have been successfully employed by the ENCODE project, and other groups have determined both methods produce quality libraries (The ENCODE

Project Consortium, 2007; Ho et al., 2011; The ENCODE Project Consortium, 2012). However, ChIP-seq is now the standard as next generation sequencing is well developed. ChIP-seq is also capable of identifying a greater number of peaks and narrower peaks than ChIP-chip (Ho et al., 2011). The results of ChIP-seq studies from the ENCODE project have made a profound impact on the understanding of regulatory elements and gene expression across the genome.

Through the integration of ChIP data, models that use histone modifications to predict transcription and cis-regulatory elements have been developed (Wang et al., 2008; Karlić et al., 2010; Dong et al., 2012). Using a set of 38 histone modifications, Karlić (2010) created a model that could predict gene expression in CD4+ T-cells with a Pearson correlation coefficient of $r = 0.77$. Using the same dataset, Wang (2008) identified a 'backbone' of 17 histone modifications, including H3K4me1, H3K4me3, and H3K27ac, that was positively correlated with gene expression. Further, Wang (2008) identified a cluster of four modifications, including H3K9me2, H3K9me3, H3K27me2, and H3K27me3, that was associated with gene silencing. Another study by Dong et al. (2012) was able to predict gene expression with up to 83% accuracy in seven cell lines using 11 histone modifications. While large sets of histone modifications can predict gene expression with greater accuracy, Karlić (2010) demonstrated that a model based on H3K27ac, alone, predicted gene expression with 72% accuracy. These studies clearly demonstrate the value of histone modification data in prediction gene expression. The ENCODE project takes these concepts a step farther to identify how various histone modifications interplay to modulate the activity of cis-regulatory elements, such as enhancer, promoters, and silencers.

In the production phase of the ENCODE project, 12 histone modifications, including H3K27ac, H3K4me1, H3K4me3, and H3K27me3, were assessed across 6 cell lines (The ENCODE Project Consortium, 2012). One group identified 36,589 putative enhancers marked by H3K4me1 in HeLa cells and 24,566 enhancers in K562 cells, yet only 22% of enhancers were found in both cell lines (Heintzman et al., 2009). The ENCODE Project Consortium (2012) assessed promoter-like, enhancer-like, and repressive regions partially characterized by the presence of H3K4me3, H3K4me1, and H3K27me3, respectively. ENCODE identified 339,124 enhancer regions, 70,292 promoter regions, and hundreds of thousands of quiescent or repressed regions across 46 cell lines (The ENCODE Project Consortium, 2012). Both Heintzman (2009) and the ENCODE Consortium (2012) identified a large degree of cell type specificity across chromatin states; however, repressive regions were most commonly found in all six cell types, supporting previous studies that suggest the importance of H3K27me3 in regulating genes involved in early development (The ENCODE Project Consortium, 2012; Boyer et al., 2006). Together, these studies demonstrate the value of examining histone marks in relation to genome function.

Overall, the ENCODE project identified evidence of over 80% of the genome participating in biochemical reactions related to transcription and chromatin function (The ENCODE Project Consortium, 2012). The project depicted the pervasive nature of transcripts and regulatory elements in the genome. Further, the ENCODE project demonstrates the need for assessing regulatory elements on a tissue-by-tissue basis. The wealth of knowledge derived from the ENCODE project has vastly improved the study of human genomics; however, as depicted in the study, gene activity and regulatory

elements are not reliably conserved across tissues and species. Therefore, projects to functional annotation genomes of other species have followed in the footsteps of the ENCODE project.

The Functional Annotation of Animal Genomes (FAANG) Project

The FAANG consortium was established in 2014 with the intent of functionally annotating the genomes of domestic species. The broad aim of the project is to enhance the understanding of how the genome contributes to phenotype (The FAANG Consortium, 2015). Similar to the ENCODE project, the FAANG project was intended to be, and is, highly collaborative. Open access data portals have been generated to allow for easy access of datasets as they are produced (FAANG.org). The project began with groups of researchers dedicated to studying livestock species, such as cattle, sheep, pigs, and chickens (The FAANG Consortium, 2015). To allow for interspecies comparisons, key tissues and assays were selected to provide a standardized core of research on each species. Key tissues that were prioritized include skeletal muscle, adipose, liver, and tissues from the reproductive, immune, and nervous systems (The FAANG Consortium, 2015). The core assays include RNA-seq, ChIP-seq of H3K4me1, H3K4me3, H3K27ac, and H3K27me3, and ATAC-seq (assay for transposase-accessible chromatin sequencing) (The FAANG Consortium, 2015). Since its establishment in 2014, the FAANG project has grown to include many other species. Research groups specific to cattle (Fang et al., 2019), sheep (Davenport et al., 2021), pigs (Pan et al., 2021), chickens (Kern et la., 2021), various farmed fish, and horses (Kingsley et al., 2020) have been successfully

bolstering the knowledge of genomic function in important domestic species, yet most of these projects are still underway. Such is the case with the equine FAANG project.

The genomic resources available for the horse provide a starting point for genomic research into equine health and performance. Yet, limitations in the annotation of the reference genome have repeatedly demonstrated the need for functional annotation of the equine genome. The work outlined below aims to fulfil these needs and represents a portion of the results from the equine FAANG project. The use of RNA-seq and ChIP-seq in identifying gene expression and regulation in two Thoroughbred stallions will be discussed in detail. The results of this research will allow for functional characterization of the genome across eight different tissues and will provide an invaluable resource to the equine community.

CHAPTER 2: CHARACTERIZING THE TRANSCRIPTOME OF EIGHT TISSUES IN HEALTHY THOROUGHBRED HORSES

Introduction

Mammalian genomes contain approximately 20,000 protein coding genes with up to twice as many non-coding transcripts (Frankish et al., 2021). Many of these genes and transcripts are differentially expressed across cell types to allow for diverse biological functions throughout the body. Beyond normal tissue function, genes can be differentially expressed in response to environmental stimuli and to allow for sexual dimorphism. To determine changes in transcription under various conditions, an understanding of normal gene expression in healthy tissues is imperative. Thus, many studies in humans have aimed to characterize normal gene expression across a variety of cells and tissues.

Assessing more than 50 cell lines and tissues, together, many groups have characterized genes commonly expressed across tissues and those exclusive to one or a few tissue types in humans (Ramsköld et al., 2009; Djebali et al., 2012; Melé et al., 2015). Ramsköld and others (2009) identified between 11,000 and 15,000 genes expressed in each of 16 assayed tissues. Over 8,000 genes were ubiquitously expressed in the 16 tissues, yet tissues such as the testis and brain expressed over 6,000 genes beyond those ubiquitously expressed, many of which were unique to either the brain or testis (Ramsköld et al., 2009). Similarly, Melé et al. (2015) found thousands of genes differentially expressed across 43 human cell types and tissues. Further, The ENCODE project suggests that as many as 7% of protein coding genes and 29% of long non-coding RNAs (lncRNAs) were specific to one of 15 studied human cell lines (Djebali et al., 2012). Together, these studies demonstrate the vast number of genes expressed across

tissues. As the number of genes determined to be ubiquitously expressed or cell-type specific is highly dependent on the tissues being studied, the transcriptome of a species is best characterized through the examination of a variety of tissues.

Although sex may play a lesser role in differential gene expression than tissue type, a significant number of genes are differentially expressed across sexes. One study identified 92 protein coding genes and 43 lncRNAs that demonstrated global sex-biased expression across 43 human cells and tissues with many of the differentially expressed genes residing on the sex chromosomes (Melé et al., 2015). Another 753 tissue-specific genes also demonstrated sex-biased gene expression (Melé et al., 2015). Over 6,500 genes were differentially expressed by sex in at least one of 53 human tissues, yet only a small number of genes were differentially expressed due to sex in all tissues (Gershoni & Pietrokovski, 2017). Many organs, such as the brain and heart, exhibit a large degree of sex-biased gene expression, suggesting the importance of accounting for both tissue type and sex in study design (Mayne et al., 2016). In the cases where multiple tissues cannot be examined or sex cannot be balanced between treatment groups, understanding how tissue type and sex impact gene expression in healthy tissues can be valuable for removing bias from applied differential expression analyses.

Although the transcriptome has been well characterized in a variety of tissues in humans, studies characterizing the transcriptome of a variety of healthy tissues is lacking in the horse. Most transcriptome studies in the horse have been used to aid the annotation of the equine genome, thus many projects examined a limited number of tissues or pooled many tissues to generate only a single transcriptome (Coleman et al., 2010; Hestand et al., 2015; Pacholewska et al., 2015). As part of the equine Functional Annotation of

Animal Genomes (FAANG) project, this work aims to provide a resource describing the transcriptome of seven common tissues, testis and ovary in healthy, adult Thoroughbred horses. To do so, we generated poly-A^+ RNA-seq libraries for adipose, parietal cortex, left ventricle of the heart, lamina, liver, lung, skeletal muscle, and testis from two healthy Thoroughbred stallions. These data were combined with similar libraries from two healthy Thoroughbred mares to characterize the transcriptome in healthy tissues and identify differentially expressed genes across sexes. This project was designed to provide needed insight into normal gene function of commonly studied tissue types and to be publicly available for use in equine genetic studies.

Materials and Methods

RNA Isolation

Tissues were provided by the FAANG biobank and were collected from two Thoroughbred stallions between the ages of three and four years old (Donnelly et al., 2021). The methods used to collect and store tissues are described in Donnelly et al. (2021); in brief, samples utilized for RNA-sequencing were flash frozen in liquid nitrogen and stored at -80˚C until RNA isolation. Of the 102 tissues in the biobank, 8 tissues, including those prioritized by the FAANG Consortium (The FAANG Consortium, 2015) and those important to equine health, were chosen for study including abdominal adipose, parietal cortex, left ventricle of the heart, lamina, liver, left lung, *Longissimus dorsi* muscle, and left testis. All surfaces and tools were cleaned with RNAseZap (Invitrogen, Waltham, Massachusetts) to reduce RNAses in the workspace. Approximately 70mg of each tissue (100mg of adipose) was minced with a razor on dry

ice. Tissue samples were then homogenized in 1ml of Trizol™ (Invitrogen, Waltham, MA, USA) using a Kinematic Polytron™ (Luzern, Switzerland) on ice in 30 second bursts. After homogenization, the samples were incubated at room temperature for 5 minutes before the addition of 200μL of chloroform. Tough tissues, including adipose and testis, were incubated for 10 minutes or more at room temperature before undergoing a dirty spin. For the dirty spin, the adipose and testis samples were centrifuged at 12,000 x g for 10 minutes, and the supernatant was transferred to a clean 2.0mL tube before the addition of chloroform. After the addition of chloroform, the samples were vortexed, incubated for 2-3 minutes at room temperature, and spun at 12,000 x g for 15 minutes. The clear supernatant of each sample was added to 600mL of ethanol and placed on a spin column. RNA was isolated using the Zymo Research Direct-zol RNA Miniprep kit (Irvine, CA, USA) according to manufacturer guidelines with minor revisions as described. DNA was removed from samples using a 15 minute on-column DNase I treatment. In addition to the 1-minute spin outlined in the Zymo protocol, the columns were spun for an additional two minutes before elution in DNase/RNase free water to remove any residual wash buffers. The speed and number of bursts used to homogenize samples, their elution volume, and the presence of additional steps are recorded in **Table 2.1**. RNA was quantified using an Agilent Bioanalyzer 2100 Eukaryote Total RNA Nano chip (Santa Clara, CA, USA) and RNA integrity (RIN) values were used to determine RNA quality.

RNA Libraries from two Thoroughbred Mares

Tissues from two healthy Thoroughbred mares between the ages of four and five years old were previously collected for the equine FAANG project (Burns et al., 2018). The sequencing data corresponding to RNA isolated from the same tissues in two mares, with ovary replacing testis, were retrieved from the European Nucleotide Archive (ENA) under project PRJEB26787. These data are also available in the FAANG data portal under BioSampleIDs SAMEA104728877 and SAMEA104728862. The protocol used to isolate RNA from the mare tissues is similar to what is described above. The mare data consisted of 125bp paired-end, stranded, poly(A+) selected libraries (Illumina TruSeq).

Data Analysis

RNA from the stallion tissues was sent to Admera Health (South Plainfield, NJ, USA), prepped using the TruSeq kit (Illumina, San Diego, CA, USA), and sequenced on a NovaSeq 6000 Sequencing System (Illumina, San Diego, CA, USA). Libraries include stranded, poly(A+) selected 150 bp paired-end reads. After data were received, adaptors were removed and reads were trimmed using Trim-Galore (Kruger, 2019) and Cutadapt (Martin, 2011). FastQC (Andrews, 2010) and mutliQC (Ewels et al., 2016) were used to assess read quality. Trimmed reads were mapped to the EquCab3.0 reference genome using STAR aligner (Dobin et al., 2013) under the default parameters. PCR duplicates were marked with sambabma (Tarasov et al., 2015), and mapping rates, qualities, and read lengths were assessed with samtools (Li at al., 2009) and deeptools (Ramírez et al., 2016). The generated bam files were filtered to remove unmapped reads, alternative alignments, PCR or optical duplicates, and reads that failed Illumina quality checks using

samtools (Li et al., 2009). Further, reads were selected to include only those that were mapped and properly paired.

Gene Expression Profiles and Pathway Analysis

The transcripts corresponding to each gene were quantified using Subread's featureCounts (Liao et al., 2014). The Refseq annotation of EquCab3.0 (GCF_002863925.1) was used to quantify the expression of 30,647 genes. Transcripts were counted as pairs and required to be at least 49bp in length. Gene length was determined by featureCounts using the length of the gene's exons and transformed to transcripts per kilobase million (TPM) to account for both gene length and sequencing depth. TPM was calculated by first determining the reads per kilobase (RPK) for each gene. RPK was calculated by dividing a gene's read count by its length in kilobases. The RPK of each gene was summed and divided by 1 million to get a scaling factor for each sample. The RPK values were then divided by the sample's scaling factor to get TPM.

Genes were considered expressed in an individual if greater than 10 TPM were found at that locus. This threshold is considered "medium expression" by the European Bioinformatics Institute's Expression Atlas (EMBL-EBI, Hinxton, Cambridgeshire, UK; https://www.ebi.ac.uk/gxa/FAQ.html#). In the case of the ovary and testis, genes were considered expressed if both biological replicates from the corresponding sex expressed the gene at greater than 10 TPM. Comparisons of gene expression across the remaining tissues was performed with genes expressed at greater than 10 TPM in all four biological replicates. Genes were considered expressed in only one sex if expression was greater than 10 TPM in both replicates of that sex and below 1 TPM in both replicates of the

opposite sex. Highly expressed genes were defined as genes with greater than 1000 TPM, again, following the guidelines outlined by EMBL-EBI's Expression Atlas (Hinxton, Cambridgeshire, UK; https://www.ebi.ac.uk/gxa/FAQ.html#). Genes expressed in all biological replicates for a given tissue were analyzed for KEGG pathway enrichment using David Bioinformatics Database's Functional Annotation Tool (Huang et al., 2009a; Huang et al., 2009b). The gene names were converted to Entez gene IDs by the David Bioinformatics Database, and the background list was the default for *Equus caballus*. Significantly enriched KEGG pathways were defined as having a false discovery rate (FDR) of less than 0.05.

Differential Expression Analysis

Differential expression analysis was performed using the raw read counts generated by featureCounts (Liao et al., 2014). The R package, DESeq2 (Love et al., 2014), was used for the analysis. Comparisons were made across sexes within each tissue with female expression considered reference. Genes with fewer than 10 transcripts identified across all individuals were discarded to remove genes with little or no expression in a given tissue. Differentially expressed genes were filtered to maintain only those with an Benjamini-Hochberg adjusted p-value of less than 0.05. Differentially expressed genes were categorized as overexpressed in the stallions if the $\log_2$ fold change was positive and overexpressed in the mares if the $\log_2$ fold change was negative.

Results

RNA Quality, Read Annotation, and Data Availability

The two stallion replicates are denoted as AH3 and AH4. The RIN values for the stallion samples ranged from 7.6 to 9.7 with an average of 8.8 (**Table 2.2**). The sequencing depth, including the mare libraries, ranged from 28.3 million to 73.3 million paired reads. The average sequencing depth across both mares and stallions was 39.4 million paired reads. The bam files generated for this project can be accessed at https://equinegenomics.uky.edu/faangHorses_RNASeq.html. On average, 92.9% of reads mapped to a single location in the genome. After filtering to remove PCR duplicates, unmapped reads, and low-quality reads, the average library across mares and stallions consisted of 25.2 million read pairs.

Gene Expression Profiles

An average of 72.6% of uniquely mapped read pairs were assigned to genes within the EquCab3.0 RefSeq annotation, while an average of 23.4% of read pairs mapped outside of annotated exons. The total number of gene-assigned reads ranged from 6.8 million reads in a mare heart sample to 31.1 million reads in a mare brain sample. On average, each sample had 18.1 million reads assigned to annotated genes.

The number of genes expressed varied by tissue, but on average, 8,068 genes displayed at least medium expression (>10 TPM) across tissues (**Figure 2.1**). Skeletal muscle expressed the fewest genes at 6,330 or 21% of the genes in the RefSeq annotation. The greatest number of genes expressed in a tissue shared across all replicates were found in the brain. The brain expressed 9,884 genes comprising 32% of the genes in the RefSeq annotation. The testes and ovaries expressed 10,604 and 8,922 genes, respectively. Excluding reproductive tissues, the greatest number of sex-specific genes

were identified in the female brain followed by the female liver with 27 and 13 genes, respectively (Supplementary Table 2.1). Together, 11,115 genes were expressed across the nine studied tissues comprising 36.2% of annotated genes. There were 4,218 genes expressed in all nine tissues with 4,700 genes expressed at greater than 10 TPM in only one studied tissue. Adipose had the fewest tissue-specific genes at 58 while the testis had the greatest number of tissue-specific genes at 2,288 (**Figure 2.1**).

Genes with high levels of expression (> 1000 TPM) were also examined. On average, each tissue had 88 genes that were highly expressed. The heart had the greatest number of highly expressed genes at 132 genes and the brain the fewest (35 genes). Each tissue had a small number of highly expressed genes unique to that tissue. The greatest number of tissue-specific, highly expressed genes were found in the liver, with the fewest found in the lamina. There were 93 highly expressed genes unique to the liver and only 6 highly expressed genes unique to the lamina **(Table 2.3)**. The unique, highly expressed genes frequently had functions specific to that tissue type, such as complement genes (*C1S* and *C1R*) in the liver, keratin 14 (*KRT14*) in the lamina, and synaptosome associated protein 25 (*SNAP25*) in the brain (Supplementary Table 2.2). Only two genes were highly expressed across all nine tissue types: miRNA-703 (*MIR703)* and tumor protein, translationally-controlled 1 (*TPT1*).

KEGG Pathway Enrichment of Expressed Genes

Pathways were considered enriched if the FDR was less than 0.05. Each tissue had between 130 and 182 enriched pathways (**Table 2.4).** Ninety-eight of these pathways were enriched in all studied tissues. The number of pathways unique to a tissue ranged

between zero and 23 (**Table 2.4**). The liver and brain had the most unique pathways with 23 unique to the liver and 18 unique to the brain. No pathways were unique to lamina or muscle. Pathways unique to a tissue are often specific to that tissue's function. Genes involved in the regulation of lipolysis in adipocytes were enriched in the adipose, and genes involved in three cardiomyopathy pathways were enriched in the heart. The pathways shared between all 9 tissues and those unique to only a single tissue can be found in Supplementary Table 2.3.

Sex-Based Differential Expression Analysis

To understand the impact of sex on gene expression, genes differentially expressed between mares and stallions were assessed in each of the seven shared tissues (**Figure 2.2**). The tissue with the most differentially expressed genes was adipose in which 1,765 genes were overexpressed in the stallions and 1,617 genes were overexpressed in the mares. The lung had the fewest differentially expressed genes with 27 overexpressed in the stallions and 10 overexpressed in the mares (**Figure 2.2**). Only four genes were differentially expressed across all 7 shared tissues including HLA class I histocompatibility antigen, alpha chain G (*HLA-G*), centriole and centriolar satellite protein (*OFD1*), an uncharacterized pseudogene (*LOC111771383*), and an uncharacterized ncRNA (*LOC102150010*). *HLA-G*, *OFD1*, and *LOC102150010* were also differentially expressed between the testis and ovaries. *HLA-G*, *OFD1*, and *LOC111771383* were all overexpressed in the stallions while *LOC102150010* was overexpressed in the mares. All genes differentially expressed between sexes were

identified in at least two tissue types. The genes with the greatest log_2 fold change in expression between mares and stallions are displayed in Supplementary Tables 2.4.

A principal component analysis was performed in DESeq2 to assess clustering among samples. Regardless of which tissue was used as the reference for gene expression, the principal component plot remained static. All samples clustered closely based on their tissue type, but sex appeared to have a minimal impact on clustering (**Figure 2.3**). The first principal component accounted for 42% of the variability between samples and parsed liver, heart, and muscle from the remaining tissue types. The heart and muscle samples clustered relatively closely, but the liver samples were isolated. The second principal component accounts for 24% of the variability among samples and further separated the heart, muscle, and liver samples. Tissues that remained closely clustered include adipose, brain, ovary, testis, lamina, and lung.

Discussion

Characterizing the transcriptome is valuable for directing genomic studies; however, the number of genes identified as expressed can be highly dependent on the tissues examined and the definition of expression. In a study examining 24 tissues in mice and humans, 60-70% of annotated genes were expressed at greater than 0.3 reads per kilobase million (RPKM) (Ramsköld et al., 2009). In this study, only 36% of annotated genes in the equine genome were considered expressed across nine tissue types. Due to the low power resulting from the small number of biological replicates in this study, the threshold for gene expression was set at a conservative 10 TPM to increase the likelihood that the genes transcribed in the studied tissues would be representative of

genes expressed in that tissue in all adult horses. The drastic difference between the percentage of genes expressed across tissues between this study and Ramsköld's study can be partially explained by the differences in the number of tissues examined, the threshold for expression, and an increase in gene annotation in the last decade. Further, the different methods used to normalize read counts can impact the ability to compare relative gene expression across tissues.

Both RPKM and TPM correct for sequencing depth and gene length (Zhao et al., 2021). Correcting for these factors is important because a larger total read count will increase the number of transcripts that map, and longer transcripts often create more fragments during library preparation which will increase the number of reads aligning to those genes. However, RPKM and TPM are not the same. Unlike RPKM, the total count of TPM adjusted transcripts is the same across samples which improves the ability to compare gene expression across samples within a study; however, direct comparisons of gene expression across studies can be limited by differences in library preparation (Zhao et al., 2020). Although there are limitations to all normalization methods (Zhao et al., 2020; Zhao et al., 2021), TPM was chosen to allow for comparison across tissues in this study.

On average, 8,068 genes were identified to be expressed across the assayed tissues. The fewest genes were expressed in the *Longissimus dorsi* with the greatest number of genes expressed in the testis and then the brain. Each tissue in this study expressed 20-35% of the genes in EquCab3.0 RefSeq annotation. Ubiquitously expressed genes made up a large proportion of the genes expressed in each tissue. The majority of the genes expressed in the muscle (67%), heart (60%), and liver (57%) where genes that

were ubiquitously expressed. The testis and brain had the smallest percentage of genes that were ubiquitously expressed, 40% and 43%, respectively, and the largest proportion of genes that were tissue-specific. These findings mirror those found by Ramsköld and others (2009) in which skeletal muscle had the fewest expressed genes and the testis had the greatest number of expressed genes. Ramsköld also found the transcriptomes of skeletal muscle and liver to consist largely of ubiquitously expressed genes while the testis and brain had a greater variety of expressed genes (Ramsköld et al., 2009).

A small proportion of genes were highly expressed across the studied tissues. Two genes, however, were highly expressed in all nine tissues: miRNA-703 and *TPT1*. miRNA-703 has been identified to protect cells from inflammasome-induced pyroptosis following hypoxia in myocardial infarction (Wei et al., 2020). It is possible that *MIR703* expression was upregulated in tissues due to hypoxia after euthanasia. *TPT1* encodes the fortilin protein that demonstrates strong anti-apoptotic effects (Li et al., 2001; Pinkaew et al., 2017). It was previously found to be ubiquitously expressed in healthy tissues (Li et al., 2001). Some highly expressed genes that were found in one of the studied tissues have previously been identified to be tissue-specific in other species. For example, adiponectin (*ADIPOQ*), fatty acid binding protein 4 (FABP4), and adipogenesis regulatory factor (*ADIRF*) are all exclusively expressed or biasedly expressed in the adipose tissue (Fagerberg et al., 2014). -

In addition to uniquely expressed genes, enriched KEGG pathways were also examined. The liver had the most uniquely enriched KEGG pathways. Fourteen of the 23 uniquely enriched pathways participate in the metabolism of various molecules which is a known function of the liver. Of the seven uniquely enriched pathways in the lung, five

were involved in immune function. The epithelial cells in the lungs are known to have a large role in modulating the immune system (reviewed by Hewitt & Lloyd, 2021). Since no other immune tissues were examined in this study, it is unsurprising that pathways such as NF-κB, IL-17, RIG-I-like receptor, and toll-like receptor signaling are only enriched in the lung tissue. Ninety-eight pathways were identified to be enriched in all studied tissues. Many of these enriched pathways are involved in basic cellular function and homeostasis, such as the cell cycle, apoptosis, endocytosis, RNA polymerase, ribosomes, spliceosomes, and protein export. A large number of unexpected pathways were enriched in all tissues, such as Alzheimer disease, Amyotrophic Lateral Sclerosis, and Parkinson disease. It is hypothesized that many of these pathways are considered enriched due to the presence of genes involved in the oxidative phosphorylation pathway including ATP synthases, ATPases, NADH:ubiquinone oxioreductases, and cytochrome c and b subunits. This depicts one of the limitations of KEGG pathway enrichment analyses. Many pathways are named by the diseases they influence and include genes that are broadly involved in basic cellular function. Nonetheless, the pathways that were unique across tissues were informative of that tissue's function and helped verify the identity of the sampled tissue.

Genes expressed across tissues varied not only due to tissue type, but also due to sex. Previous studies in humans have identified many genes exhibiting sex-biased expression across a variety of tissues. Mayne and others (2016) identified over 2,000 genes differentially expressed due to sex across 15 tissue types. Another study identified as much as 37% of genes displaying sex-biased gene expression among 44 tissues (Meritxell et al., 2020). Some of the sex-biased gene expression can be contributed to

regulatory elements with hormone response elements; however, approximately two-thirds of genes displaying sex-biased expression do not contain hormone response elements (Mayne et al., 2016). Even if the chemical processes leading to differential expression across sexes are not fully understood, there is value in identifying these genes. These differentially expressed genes may explain the differences in disease risk or efficacy of drugs across sexes. Furthermore, understanding genes that are differentially expressed due to sex can be used to identify potential bias in differential expression studies where sex is not accounted for across treatment groups.

In our study, 923 genes were differentially expressed due to sex in at least one of the nine studied tissues. Only four genes were differentially expressed across all nine tissues: *HLA-G*, *OFD1*, *LOC111771383,* and *LOC102150010. HLA-G* is a gene encoding part of the major histocompatibility complex. Previous studies have identified sex-biased expression in HLA genes (Stein et al., 2021); however, Stein and others (2021) note that accurate quantification of transcripts in HLA genes can be difficult due to the highly polymorphic nature of these genes. These polymorphisms can impair the mappability of transcripts to HLA genes. Therefore, the differential expression of *HLA-G* may reflect the relatedness of biological replicates to the horse used for the reference genome rather than differential expression due to sex. Both *OFD1* and *LOC102150010* are located on the X chromosome. Although one X chromosome is generally inactivated in females, prior research indicates that as many as 15% of genes on inactivated X chromosome are expressed (Prothero et al., 2009). Therefore, it is possible that the overexpression of *LOC102150010* in females could be due to escape of X chromosome inactivation in the females. Since *LOC102150010* is an uncharacterized ncRNA in the horse, greater

annotation of noncoding genes will be required to understand the impact of sex on this ncRNA. *OFD1* has been described to map to both the X and Y chromosomes in other species (Chang et al., 2011). In cattle, *OFD1Y* was found to be expressed in a variety of tissues (Chang et al., 2011). The overexpression of *OFD1* in the stallions may reflect an unannotated copy of *OFD1* on the Y chromosome which is not currently apart of the EquCab3.0 reference genome. *LOC111771383* represents an uncharacterized pseudogene on equine chromosome 23, and therefore, the potential reasons for differential expression cannot be currently examined.

The degree of differential expression due to sex varied widely across tissues. Adipose had 3,382 genes that were differentially expressed while the lung had only 37. The three tissues with the greatest differential gene expression were adipose, muscle and liver. Both adipose and skeletal muscle were also found to be highly differentially expressed across sexes in a study of human tissues (Gershoni & Pietrokovski, 2017). Understanding sex-biased gene expression is important when analyzing data from studies including just one sex. Although some tissues such as lung and lamina have less sex-bias in gene expression, commonly studied tissues related to athletic performance and metabolic diseases have 900 or more genes that are differentially expressed across the sexes (McGivney et al., 2010; Park et al., 2012; Ertelt et al., 2014; Ropka-Molik et al., 2017). Therefore, it can be difficult to reliably extrapolate results generate in one sex to the other and this result warrants examination of both sexes when considering the impact of various stimuli on gene expression.

The data presented in this study provide a valuable resource to the equine community. The gene expression profiles of nine healthy tissues are characterized which

may be useful in informing researchers which tissues and sexes would be the most beneficial to study regarding their research intents. For example, when examining candidate mutations related to a disease that mainly impacts a single tissue, resources can be focused towards mutations in genes that are expressed in the impacted tissue. While the findings of this study are informative, the small sample size is a limitation of the study. Although a conservative threshold for gene expression was employed to account for this, the genes found to be expressed, and especially those differentially expressed across sexes, may not be representative of the transcriptome present in all healthy adult horses. Furthermore, only nine tissues were compared meaning that many of the genes and pathways identified to be unique in this dataset may also be enriched in other tissues that were not examined in this study. Nonetheless, this study provides a strong foundation for the study of the equine transcriptome and will be valuable for informing the design of applied research in the equine community.

Table 2.1. Tissue Homogenization and RNA Elution Specifications

Tissue	**Homogenization**		**RT Incubation/ Dirty Spin**	**Elution Volume**
	Speed (kRPM)	**Duration**		
Abdominal Adipose (Adipose)	26-28	30 sec (x3)	10 min: Yes	50uL
Parietal Cortex (Brain)	10-12	20 sec (x2)	5 min: No	50uL
Heart Left Ventricle (Heart)	22-24	30 sec (x3)	5 min: No	50uL
Lamina	22-24	30 sec (x3)	5 min: No	100uL
Liver	22-24	30 sec (x3)	5 min: No	80uL
Left Lung (Lung)	22-24	30 sec (x3)	5 min: No	50uL
Longissimus dorsi (Muscle)	18-20	30 sec (x3)	5 min: No	100uL
Left Testis (Testis)	22-24	30 sec (x4)	20 min: Yes	100uL

Table 2.2. RIN Values of RNA Isolated From Stallion Tissues

Tissue	AH3 RIN	AH4 RIN
Adipose	8.4	8.5
Brain	9.5	9.1
Heart	8.3	7.6
Lamina	8.8	8.7
Liver	9.4	9.4
Lung	8.0	8.4
Muscle	8.8	9.2
Testis	9.7	9.7

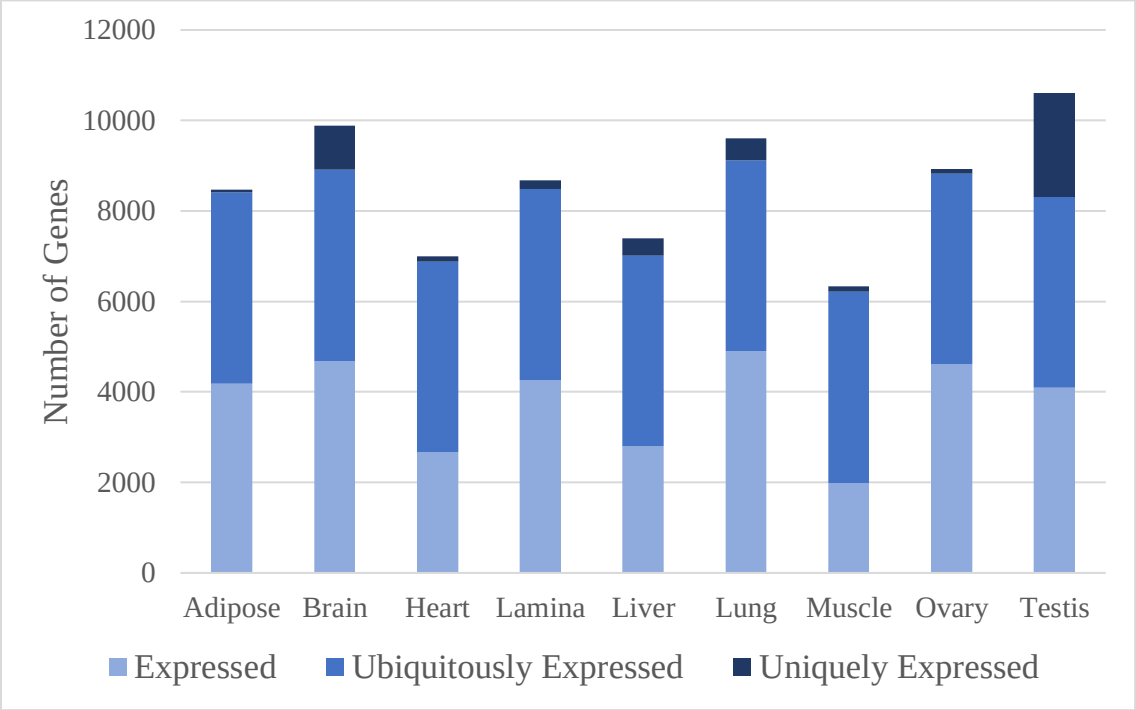


Figure 2.1. Genes Expressed by Tissue
The total number of genes expressed in a tissue across all four replicates (two replicates for ovary and testis) is represented by the total height of the bar. Genes must be expressed in all biological replicates for a gene to be considered expressed in a tissue. In the ovary and testis, genes are considered expressed if both biological replicates of that sex express a given gene. The genes in lightest blue are expressed in that tissue as well as in at least one other tissue in the dataset. Genes in the medium blue are ubiquitously expressed across all nine tissue types. Genes in the darkest blue were expressed at greater than 10 TPM in only that tissue.

Table 2.3. Genes Highly Expressed (>1000 TPM) in All Biological Replicates Across Tissues

Tissue	Highly Expressed Genes	Tissue-Specific High Expression
Adipose	92	11
Brain	35	16
Heart	132	45
Lamina	83	6
Liver	113	93
Lung	80	10
Muscle	122	38
Ovary	90	13
Testis	46	33

Table 2.4. Number of Enriched KEGG Pathways

Tissue	All Pathways	Unique Pathways
Adipose	182	1
Brain	172	18
Heart	154	3
Lamina	162	0
Liver	172	23
Lung	182	7
Muscle	133	0
Ovary	175	3
Testis	130	1

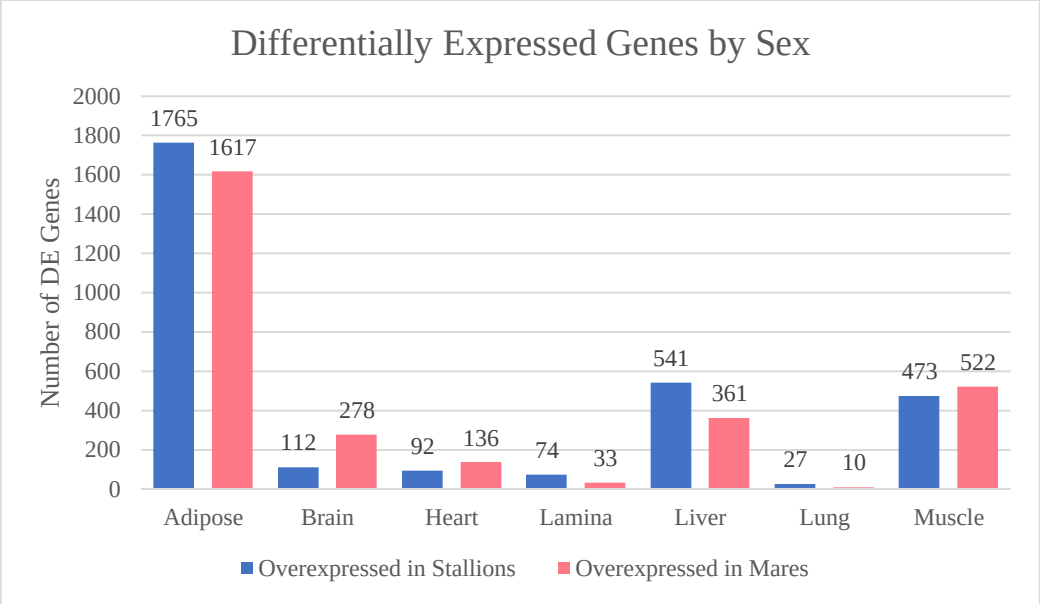


Figure 2.2. Differential Gene Expression Due to Sex
Differentially expressed genes ($P_{adj} < 0.05$) across tissues due sex. Genes found to be upregulated in males are in blue while the genes upregulated in females are in pink. The number of differentially expressed genes is recorded at the top of each bar.

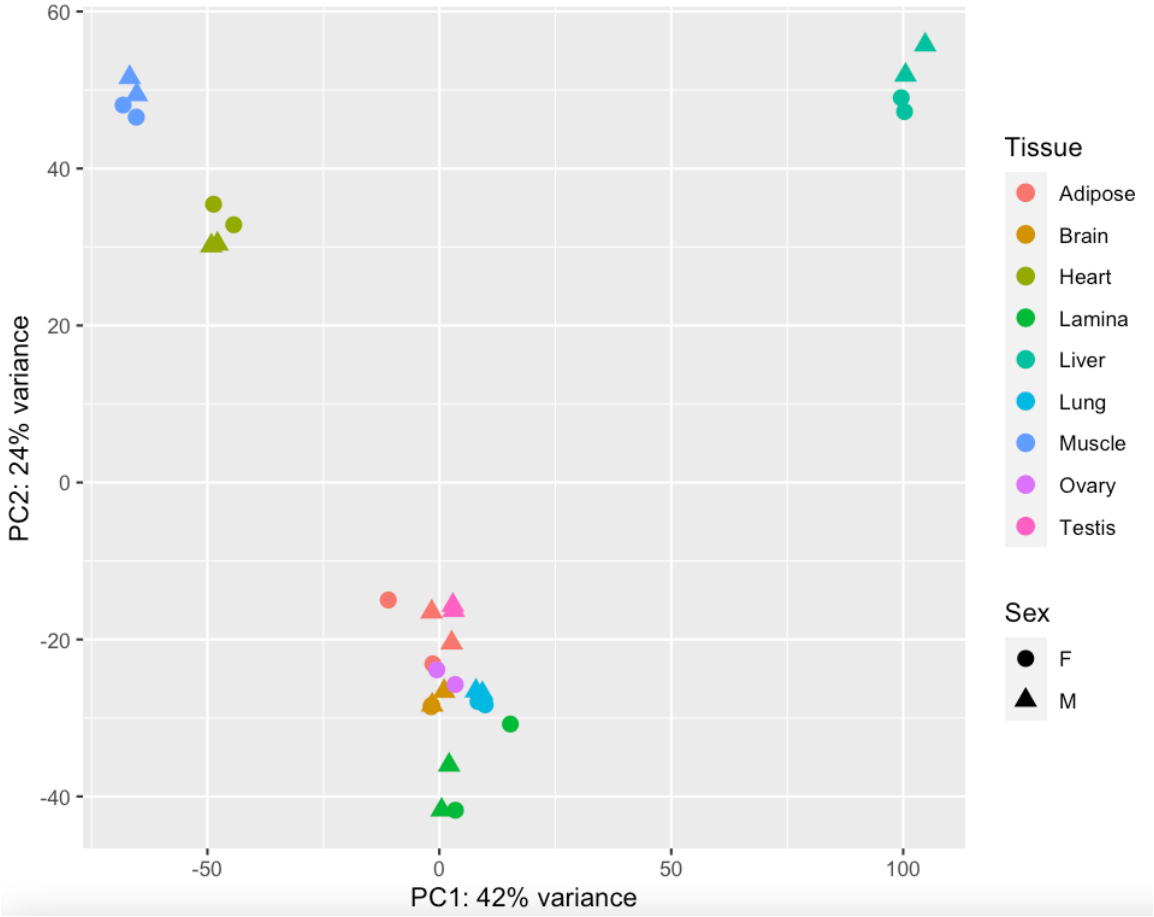


Figure 2.3. Principal Component Analysis
A principal component analysis performed in DESeq2 following differential expression analysis considering both sex and tissue types as factors. Samples cluster strongly by tissue, but not by sex. Together, the first 2 principal components account for 66% of the variability among samples.

Supplementary Table 2.1. Genes Expressed in Only One Sex

	Female Only	Male Only
Adipose	*EDIL3, LOC102150010, LOC111770938, PROKR1*	*EPM2AIP1, LOC100064259, LOC106783151, LOC111771275, LRAT, MIR9184, PRM1, RSPH4A, SLC4A4*
Brain	*LOC100053403, LOC100053499, LOC100053847, LOC102148406, LOC102150010, LOC106782239, LOC111769679, LOC111770252, LOC111770264, LOC111770486, LOC111770496, LOC111770499, LOC111770515, LOC111770525, LOC111770527, LOC111770531, LOC111770545, LOC111770546, LOC111770553, LOC111772707, LOC111773021, LOC111773023, LOC111774090, LOC111774512, LOC111775847, MIR1543*	None
Heart	None	*MIR219-1, MIR8944, PRM1*
Lamina	None	*LOC100051724, LOC111767996, LOC111770981*
Liver	*FZD10, LOC100055856, LOC102148406, LOC102150010, LOC111769679, LOC111769781, LOC111770262, LOC111770497, LOC111770545, LOC111770553, LOC111774220, MIR8985, MIR9041*	*LOC111772972*
Lung	None	*MIR1244*
Muscle	*LOC102148741*	*HBA, LOC111774862, PRM1*

Supplementary Table 2.2. Genes Highly Expressed in a Single Tissue

Adipose	*ADIPOQ, ADIRF, CAVIN1, DCN, G0S2, LGALS1, LOC100049811, LOC100057425, PLIN1, RPS28, S100A1*
Brain	*APP, PLP1, ALDOC, CALM3, CPE, GLUL, MBP, NRGN, PEA15, S100B, SNAP25, SPARCL1, THY1, UCHL1, YWHAE, YWHAH*
Heart	*ACTC1, CHCHD10, ATP2A2, ATP5MG, CYC1, CYCS, ACTN2, ATP5MD, ATP5PB, ATP5PF, ACO2, ANKRD1, ATP5F1C, ATP5IF1, ATP5MF, ATP5PD, BSG, FHL2, GNG5, HSPB7, LOC100055813, LOC100630549, LOC100630564, LOC106781327, LOC106781507, LOC111767815, MYBPC3, MYH7, MYL2, MYL3, MYOZ2, NDUFA4, NDUFB2, NDUFB8, NDUFS2, NDUFS6, PDLIM1, PLN, PRDX2, PSAP, SDHD, SMPX, TNNC1, TNNT2, VDAC3,*
Lamina	*ANXA1, PERP, KRT14, S100A6, SFN, LOC100630872*
Liver	*AFM, AMBP, LOC100073265, LOC106782649, APOA2, CAT, ALDH1L1, CYP2E1, CYP3A97, AGT, APOC2, APOH, ASS1, A1BG, ALDH1A1, APOA1, APOC3, ASGR1, C1S, FGA, FMO3, GC, HRG, ITIH2, ITIH4, LOC100034242, LOC100050685, LOC100053249, LOC100060505, LOC100061234, LOC100061367, LOC100061692, LOC100066603, LOC100067869, LOC100070616, LOC106782650, LOC106782651, METTL7B, ACSL1, ALDOB, C1R, CDO1, CFB, CFI, CYP2A13, AOX1, CPS1, LOC100071061, AHSG, APOB, CCL16, CRP, CYB5A, DHRS7, DPYS, ECHS1, EPHX1, F10, F2, FGB, FGG, FN1, GLUD1, GSTA1, HAAO, HAMP, HPD, HPX, HSD17B13, ITIH1, KNG1, LECT2, LOC100051562, LOC100053468, LOC100056506, LOC100059239, LOC100070400, MAT1A, PAH, PGRMC1, PLG, RARRES2, RGN, RNASE4, SERPINC1, SERPIND1, SERPINF2, SERPING1, SLCO1B3, TAT, TMBIM6, TTR, VTN*
Lung	*HSPA6, LOC100630171, LOC106781303, MARCO, NPC2, RGS2, SCGB1A1, SEC14L3, SFTPA1, SFTPC*
Muscle	*ACTN3, ADSSL1, AK1, ATP2A1, BIN1, CA3, CASQ1, CNBP, CSDE1, EEF2, EIF4A2, GPD1, GPI, KLHL41, LDB3, LDHA, LOC100051065, LOC100058290, MYBPC2, MYH1, MYL1, MYLK2, MYLPF, MYOT, MYOZ1, PDLIM3, PFKM, PGK1, PGM1, PKM, PPP1R1A, PPP1R27, SH3BGR, SLN, TMOD4, TNNC2, TNNI2, TNNT3*
Ovary	*ENO1, GSTO1, HMGCR, HMGCS1, IGFBP7, LOC100065786, LOC111775780, MIR675, MIR9182, MSMO1, PLA2GIB, PRDX1, RPS15, STAR*
Testis	*AKAP4, CABYR, CMTM2, CRISP2, DKKL1, DNAAF1, DYNLL1, GSG1, GSTM3, HAGH, HMGB4, INSL3, LOC100072403, LOC100073295, LOC102147484, LOC102148276, OAZ3, ODF1, ODF2, PABPC1, PRM1, PRM2, PSMD2, RAN, RPL38, SPA17, STMN1, TNP1, TPPP2, TSACC, TUBB4B, WASHC3, YBX2*

Supplementary Table 2.3. Ubiquitously-Enriched and Tissue-Unique KEGG Pathways

Tissue	Enriched KEGG Pathways
All Tissues	Adherens junction, Adipocytokine signaling pathway, AGE-RAGE signaling pathway in diabetic complications, Alzheimer disease, AMPK signaling pathway, Amyotrophic lateral sclerosis, Apoptosis, Autophagy - animal, Autophagy - other, Bacterial invasion of epithelial cells, Basal transcription factors, Biosynthesis of cofactors, Carbon metabolism, Cell cycle, Cellular senescence, Central carbon metabolism in cancer, Chemical carcinogenesis - reactive oxygen species, Choline metabolism in cancer, Chronic myeloid leukemia, Citrate cycle (TCA cycle), Colorectal cancer, Coronavirus disease - COVID-19, Diabetic cardiomyopathy, EGFR tyrosine kinase inhibitor resistance, Endocrine resistance, Endocytosis, Endometrial cancer, Epstein-Barr virus infection, ErbB signaling pathway, Fatty acid degradation, Fatty acid metabolism, Fluid shear stress and atherosclerosis, Focal adhesion, FoxO signaling pathway, Glioma, Glucagon signaling pathway, Glyoxylate and dicarboxylate metabolism, Hepatitis B, Hepatocellular carcinoma, HIF-1 signaling pathway, Human cytomegalovirus infection, Human immunodeficiency virus 1 infection, Human papillomavirus infection, Human T-cell leukemia virus 1 infection, Huntington disease, Insulin resistance, Insulin signaling pathway, Longevity regulating pathway, Longevity regulating pathway - multiple species, Lysine degradation, Lysosome, Metabolic pathways, Mitophagy - animal, mRNA surveillance pathway, mTOR signaling pathway, N-Glycan biosynthesis, Neurotrophin signaling pathway, Non-alcoholic fatty liver disease, Non-small cell lung cancer, Nucleocytoplasmic transport, Nucleotide excision repair, Oocyte meiosis, Oxidative phosphorylation, Pancreatic cancer, Parkinson disease, Pathways of neurodegeneration - multiple diseases, PD-L1 expression and PD-1 checkpoint pathway in cancer, Peroxisome, Prion disease, Propanoate metabolism, Prostate cancer, Proteasome, Protein export, Protein processing in endoplasmic reticulum, Pyruvate metabolism, Regulation of actin cytoskeleton, Renal cell carcinoma, Ribosome, Ribosome biogenesis in eukaryotes, RNA degradation, RNA polymerase, Salmonella infection, Small cell lung cancer, Sphingolipid signaling pathway, Spinocerebellar ataxia, Spliceosome, T cell receptor signaling pathway, Terpenoid backbone biosynthesis, Thermogenesis, Thyroid hormone signaling pathway, Tight junction, Toxoplasmosis, Ubiquitin mediated proteolysis, Valine, leucine and isoleucine degradation, Vasopressin-regulated water reabsorption, VEGF signaling pathway, Viral carcinogenesis, Yersinia infection

Supplementary Table 3 (cont.)

Adipose	Regulation of lipolysis in adipocytes
Brain	Glutamatergic synapse, Synaptic vesicle cycle, GABAergic synapse, Morphine addiction, Cholinergic synapse, Circadian entrainment, Aldosterone synthesis and secretion, Long-term depression, Amphetamine addiction, Glycosaminoglycan biosynthesis - heparan sulfate / heparin, Gastric acid secretion, GnRH secretion, cAMP signaling pathway, Insulin secretion, Aldosterone-regulated sodium reabsorption, Nicotine addiction, Alanine, aspartate and glutamate metabolism, Fatty acid biosynthesis,
Heart	Hypertrophic cardiomyopathy, Dilated cardiomyopathy, Arrhythmogenic right ventricular cardiomyopathy
Lamina	None
Liver	Glycine, serine and threonine metabolism; Complement and coagulation cascades, Tryptophan metabolism, Cholesterol metabolism, Drug metabolism - other enzymes, Pentose and glucuronate interconversions, Porphyrin metabolism, Drug metabolism - cytochrome P450, Ascorbate and aldarate metabolism, beta-Alanine metabolism, Histidine metabolism, Metabolism of xenobiotics by cytochrome P450, Pantothenate and CoA biosynthesis, Riboflavin metabolism, Chemical carcinogenesis - receptor activation, Glycerolipid metabolism, Phenylalanine metabolism, Chemical carcinogenesis - DNA adducts, Nicotinate and nicotinamide metabolism, Tyrosine metabolism, Folate biosynthesis, Biosynthesis of unsaturated fatty acids, Fatty acid elongation
Lung	NF-kappa B signaling pathway, Toll-like receptor signaling pathway, Cell adhesion molecules, Rheumatoid arthritis, Transcriptional misregulation in cancer, IL-17 signaling pathway, RIG-I-like receptor signaling pathway
Muscle	None
Ovary	Glutathione metabolism, Selenocompound metabolism, Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
Testis	Pyrimidine metabolism

Supplementary Table 2.4. The Top 20 Differentially Expressed Genes by Sex

Adipose				Brain			
Gene ID	Mean Female Expression	log2 Fold Change	P adj	Gene ID	Mean Female Expression	log2 Fold Change	P adj
LOC 100064259	36.80	8.74	5.1E-06	LOC 111775630	29.03	8.72	1.64E-05
LOC 100056127	40.67	7.41	2.8E-03	LOC 111774627	29.98	7.33	2.66E-02
IYD	13.36	7.27	6.4E-03	MYH2	24.62	7.03	1.71E-03
LOC 100072708	12.77	7.20	2.7E-03	LOC 111770337	23.46	6.94	2.75E-02
LOC 111769125	12.68	7.19	1.1E-02	LOC 100062546	50.97	5.47	7.40E-07
LOC 111774060	12.53	7.18	2.4E-03	CPLX4	16.58	5.44	3.69E-03
LRAT	383.97	7.11	8.8E-31	LOC 100061551	26.77	3.61	2.51E-03
LOC 111772591	11.68	7.08	3.4E-03	LOC 100630497	27.75	3.51	4.55E-04
LOC 111771275	11.01	6.99	6.1E-03	LOC 100056099	269.58	3.33	4.47E-04
LOC 106783067	83.03	6.89	2.09E-05	LOC 111771597	50.60	3.10	4.81E-05
SIM1	16.38	-7.39	3.2E-03	LOC 111773712	305.23	-7.24	1.11E-24
DIRAS1	49.05	-7.47	3.5E-03	LOC 111770452	19.56	-7.31	1.87E-03
TRHDE	17.84	-7.51	4.8E-03	LOC 100065590	19.66	-7.32	2.34E-03
LOC 106780963	150.85	-8.18	8.4E-11	LOC 102149722	19.65	-7.32	1.83E-03
HTR2A	87.18	-8.25	3.6E-02	LOC 111770904	58.86	-7.45	1.36E-04
NELL1	42.50	-8.77	1.1E-05	LOC 100053403	286.76	-7.74	4.81E-20
LOC 102147462	52.74	-9.08	5.7E-07	LOC 100054521	307.45	-7.84	1.52E-20
KRT3	86.39	-9.79	4.4E-02	MIR1543	29.55	-7.91	1.15E-04
EGFL6	148.08	-10.57	3.1E-09	LOC 102150010	4243.46	-8.76	1.82E-119
LOC 106781302	246.91	-24.65	1.1E-05	LOC 111774090	350.70	-11.48	3.83E-12

*Positive fold changes represent genes with greater expression in the stallions. Negative fold changes represent genes with greater expression in the mares.

Supplementary Table 2.4. (cont.)

Heart				Lamina			
Gene ID	Mean Female Expression	log2 Fold Change	P adj	Gene ID	Mean Female Expression	log2 Fold Change	P adj
HSPA6	531.21	4.25	7.95E-20	SPTSSB	1510.57	10.37	4.87E-02
PRUNE2	131.14	3.65	7.49E-09	PRR9	9581.37	9.62	4.49E-02
BCAT1	301.69	3.46	2.18E-09	LOC 111774465	47.43	9.34	8.97E-05
LOC 111775969	226.31	3.40	8.31E-10	LOC 111775630	32.99	8.82	9.16E-04
LOC 100056099	824.62	3.39	8.31E-10	LOC 111770981	42.06	7.74	1.80E-03
CPXM2	144.65	3.18	7.49E-09	ADAM7	15.02	7.68	2.40E-02
SPON1	137.75	3.01	1.99E-08	LOC 111775638	195.48	7.14	2.56E-04
NR4A3	170.72	2.82	6.42E-04	LOC 111769006	75.05	7.13	4.95E-02
CHRDL1	102.93	2.76	2.45E-06	SPOCK3	70.61	7.03	3.20E-02
SPHKAP	83.17	2.76	3.04E-04	LTF	22.72	6.81	2.09E-02
LOC 100062359	197.13	-2.01	1.34E-04	ACHE	261.30	-3.10	4.43E-02
CISH	357.81	-2.03	1.41E-07	ACP7	44.52	-3.21	1.59E-02
FRZB	1023.51	-2.21	2.18E-09	LOC 100051371	302.81	-3.57	9.16E-04
OST4	100.86	-2.47	8.12E-05	LOC 100069585	31.59	-3.75	4.80E-02
CITED1	152.91	-2.50	5.87E-04	PRKG2	42.14	-5.05	3.17E-02
TSPO	228.75	-2.54	9.19E-06	CYP2E1	121.98	-5.08	4.63E-02
LOC 100146514	105.53	-2.62	1.21E-04	STK32B	36.57	-5.27	2.55E-03
LOC 100050506	271.18	-2.63	5.28E-07	LOC 102150010	4481.73	-6.22	4.82E-38
THRSP	120.03	-3.23	3.02E-06	LOC 102150085	429.63	-7.28	1.67E-06
LOC 102150010	1255.41	-5.22	1.69E-05	P19	66.82	-9.19	8.00E-04

*Positive fold changes represent genes with greater expression in the stallions. Negative fold changes represent genes with greater expression in the mares.

Supplementary Table 2.4. (cont.)

Liver				Lung			
Gene ID	Mean Female Expression	log2 Fold Change	P adj	Gene ID	Mean Female Expression	log2 Fold Change	P adj
LOC 111772972	367.45	11.06	5.25E-11	LOC 111767996	21.40	8.18	6.12E-03
LOC 111770995	68.48	10.08	1.38E-08	LOC 111774465	26.77	6.06	8.20E-03
LOC 100064796	342.63	9.37	1.26E-02	LOC 106781675	23.17	5.85	1.47E-02
LOC 100629895	11.80	7.54	1.69E-03	LOC 111775630	35.52	5.13	4.67E-04
ADAM7	11.37	7.49	1.75E-03	LOC 100056585	35.86	4.62	1.77E-03
LOC 102150542	10.09	7.31	3.34E-03	LOC 111770981	64.03	4.34	2.93E-04
LOC 111775631	9.24	7.18	1.13E-02	LOC 100051724	64.46	4.26	1.30E-03
FGF21	24.07	7.11	2.06E-02	LOC 111768886	43.52	4.10	1.21E-02
PRM1	7.85	6.95	1.44E-02	LOC 111775343	54.54	4.01	2.61E-03
GPX5	6.95	6.77	2.06E-02	LOC 111769025	46.84	4.01	5.16E-04
LOC 111772597	15.17	-6.84	1.54E-02	LOC 111774782	2279.68	-2.14	1.58E-02
LOC 111773712	414.84	-6.86	3.98E-33	LOC 100067178	926.48	-2.21	4.47E-02
LOC 102148948	15.51	-6.87	2.98E-03	CRMP1	277.87	-2.56	1.03E-04
LOC 111770083	405.06	-6.95	1.63E-32	GABRA3	234.70	-2.73	2.05E-04
KCTD4	17.75	-7.06	1.51E-03	LOC 111770338	140.00	-3.12	2.14E-02
LOC 111770262	18.95	-7.16	9.93E-04	CYP1A1	238.25	-3.18	2.05E-04
LOC 111773958	21.37	-7.33	5.32E-04	LOC 100147051	603.15	-3.55	8.86E-08
LOC 111769679	25.84	-7.61	2.00E-04	LOC 100147109	54.24	-3.83	6.23E-03
LOC 102150010	4108.12	-8.05	1.63E-136	LOC 100066745	124.43	-4.31	2.05E-04
LOC 111774090	411.22	-8.58	2.68E-19	LOC 102150010	3441.89	-5.35	9.32E-18

*Positive fold changes represent genes with greater expression in the stallions. Negative fold changes represent genes with greater expression in the mares.

Supplementary Table 2.4. (cont.)

Skeletal Muscle				Ovary/Testis			
Gene ID	Mean Female Expression	log2 Fold Change	P adj	Gene ID	Mean Female Expression	log2 Fold Change	P adj
NEFM	120.23	7.82	4.72E-02	LOC 100072403	10138.70	16.69	3.06E-27
NEFL	104.41	7.66	4.67E-05	GSG1	8381.28	16.42	2.20E-26
HSPA6	476.03	7.57	3.89E-02	CRISP2	8300.23	16.41	2.07E-26
LOC 111771118	43.31	7.40	3.43E-02	PRM1	7439.17	16.25	6.72E-26
NEFH	79.37	7.26	2.00E-04	CMTM2	7080.09	16.18	1.15E-25
LOC 111775630	31.93	6.94	1.22E-03	TNP1	6210.52	15.99	3.84E-25
NR4A3	543.36	6.24	2.71E-02	HMGB4	5506.54	15.81	1.38E-24
HBA	21.23	5.32	2.68E-03	PRM2	5217.67	15.74	2.39E-24
DSP	30.96	5.29	3.36E-03	PIWIL1	4721.97	15.59	6.16E-24
LOC 111769187	52.47	4.80	4.75E-04	ODF1	4449.61	15.51	1.15E-23
KHDRBS3	50.20	-3.19	4.25E-03	SLC5A7	55.55	-9.28	8.54E-07
CCKBR	71.36	-3.34	1.18E-03	GALNT5	55.77	-9.29	9.97E-07
LOC 111775091	27.44	-3.46	1.55E-02	PHLDA2	2157.34	-9.42	1.56E-21
KCTD16	147.45	-3.54	1.83E-10	NTS	426.31	-9.81	1.05E-02
SIX2	45.05	-3.85	1.27E-02	LOC 111768621	478.98	-9.96	2.54E-12
LOC 102148741	34.00	-4.77	3.30E-05	COL11A1	7328.34	-10.07	6.30E-20
LOC 102150010	298.71	-5.05	2.16E-36	LOC 100050034	1150.05	-12.22	1.95E-03
HOXA11	48.15	-6.62	3.72E-03	SERPINB2	613.55	-12.75	1.60E-02
LOC 102147462	30.41	-8.41	8.88E-05	CGA	3905.46	-12.95	4.48E-04
LOC 111771163	31.04	-8.44	1.11E-04	SPINK9	2201.24	-13.12	1.28E-03

*Positive fold changes represent genes with greater expression in the stallions. Negative fold changes represent genes with greater expression in the mares.

CHAPTER 3: FUNCTIONAL ANNOTATION OF PUTATIVE CIS-REGULATORY ELEMENTS IN THE GENOMES OF TWO THORUGHBRED STALLIONS

Introduction

With an annual economic impact of $50 billion in the US alone, the equine industry has long focused on improving the health and performance of horses (American Horse Council, 2018). The role of genomics on traits such as racing ability, fertility, and conformation have been widely studied since the release of the first equine reference genome in 2007 (Wade et al., 2009; Hill et al., 2010; Raudsepp et al., 2012; Singer-Hasler et al., 2012). Varying degrees of success have met these studies as limitations in the annotation of the equine genome have impaired identification of functional candidates for many traits of interest. Similar difficulties have been observed in humans with as much as 88% of trait-associated variants falling outside protein coding sequences (Hindorff et al., 2009). Yet, the annotation of most genomes, including that of the horse, is largely based on transcriptome data which primarily capture protein coding sequence (Coleman et al., 2010; Hestand et al., 2015; Mansour et al., 2017). Therefore, using additional methods to annotate the genome beyond protein-coding sequence may improve the success of genomic studies into equine health and performance.

The human ENCODE project aimed to address similar shortcomings in the human genome annotation. With the ambitious goal of identifying all functional elements in the human genome, the ENCODE project attributed function to as much as 80% of the human genome. Protein coding sequence comprised less than 3% of the human genome indicating a large presence of other functional elements (The ENCODE Project

Consortium, 2012). Despite the small footprint of protein coding genes in the genome, a large portion of the genome is dedicated to regulating their expression. Cis-regulatory elements aid in maintaining transcriptional programming and include promoters, enhancers, and silencers. The ENCODE project identified nearly 400,000 enhancer regions and 70,000 promoter regions in the human genome (The ENCODE Project Consortium, 2012). The functional annotation performed by the ENCODE project and subsequent research efforts have elucidated the impact of non-coding variants on disease traits in humans. For example, an enhancer region variant was associated with coronary artery disease and promoter region variants have been associated with schizophrenia and various forms of cancer (Gupta et al., 2017; Warburton et al., 2016; Vinagre et al., 2013). Together, these studies demonstrate the value of annotating regulatory elements throughout the genome.

The Functional Annotation of Animal Genomes (FAANG) project aims to functionally annotate the genomes of domesticated species to improve the understanding of the genotype-to-phenotype link (The FAANG Consortium, 2015). Part of this effort includes annotating regulatory elements through the use of chromatin immunoprecipitation and sequencing (ChIP-seq). ChIP-seq of four histone modifications associated with enhancers (H3K4me1), promotors (H3K4me3), active genomic regions (H3K27ac), and repressed genomic regions (H3K27me3) has been prioritized as core assays of the FAANG initiative. The equine FAANG project has previously characterized these four histone modifications in a variety of tissues from two Thoroughbred mares (Kingsley et al., 2020; Kingsley et al., 2021); however, regulatory elements have yet to be characterized in the tissues of stallions. Our group identified thousands of genes

differentially expressed between the tissues of mares and stallions indicating differences in gene regulation across sexes. Indeed, differences in histone modifications have been observed between sexes in other species demonstrating the need for annotating regulatory elements in both sexes (Shen et al., 2015; Kfoury et al., 2021). This project aims to characterize histone modifications in the tissues of two Thoroughbred stallions to allow for a comparison of regulatory elements present in the tissues of mares and stallions.

Materials and Methods

Chromatin Extraction and Immunoprecipitation

Tissue samples from abdominal adipose, parietal cortex (brain), left ventricle (heart), lamina, liver, lung, *Longissimus dorsi* (muscle), and testis of two Thoroughbred stallions were prioritized for chromatin immunoprecipitation and sequencing (ChIP-seq) analysis and retrieved from the equine FAANG Biobank (Donnelly et al., 2021). Collected tissues had been flash frozen in liquid nitrogen and stored -80 °C (Donnelly et al., 2021). ChIP preparation and sequencing was performed by Diagenode using their ChIP-seq Profiling Service (Diagenode, Cat# G02010000, Liège, Belgium). Chromatin was extracted and prepared using the iDeal ChIP-seq kit for Histones (Diagenode Cat# C01010059). Tissue samples were homogenized for two minutes using the Tissue Lyser II (Qiagen, Germany) and then fixed in 1% formaldehyde for 9 minutes (10 minutes for adipose) to crosslink histone proteins with DNA. Chromatin was sheared using a Bioruptor Pico (Diagenode, Cat# B01060001, Liège, Belgium) to achieve a targeted fragment size of 200 bp. The Bioruptor water cooler was used to maintain a temperature of 4 °C (10 °C for adipose) during shearing. Shearing occurred in cycles of 30 seconds

where samples were rested for 30 seconds between bursts. Optimization of chromatin extraction, ChIP, and library preparation was previously completed at Diagenode in equine adipose, brain (parietal cortex), left ventricle (heart), lamina, liver, lung, skeletal muscle, and ovary for part of the equine FAANG project published by Kingsley et al. (2020). Experiments to optimize fixation and shearing times of testis were performed by Diagenode prior to the analysis of the prioritized tissue samples. Information regarding the homogenization, fixation, and shearing of each sample is reported in **Table 1**. After crosslink reversal and DNA purification, shearing was assessed using the High Sensitivity NGS Fragment Analysis Kit (DNF-474) on an Agilent Fragment Analyzer™ (Santa Clara, CA, USA).

Immunoprecipitation (IP) of H3K27ac, H3K27me3, H3K4me1, and H3K4me3 histone marks was performed using the IP-Star Compact Automated System (Diagenode, Cat# B03000002, Liège, Belgium) in all samples except muscle which was done manually due to low chromatin retrieval. IP of IgG served as a negative control across samples. Additionally, 1% of chromatin from each sample was set aside for an input sample that serves to correct for background noise in downstream analysis. The amount of antibody used to precipitate each histone mark and IgG differed across tissues and was previously optimized by Diagenode (Kingsley et al., 2020) (Supplementary Table 3.1).

Library Preparation and Sequencing

Libraries for the input and four ChIP samples were prepared on the IP-Star Compact Automated System (Diagenode, Cat# B03000002, Liège, Belgium) using the MicroPlex Library Preparation Kit v3 (Diagenode Cat# C05010001). Seven to thirteen

PCR cycles were used to amplify libraries and achieve appropriate concentrations for sequencing. Libraries were double size-selected for fragments with insert sizes of ~200bp using Agencourt® AMPure® XP (Beckman Coulter, Brea, CA, USA) and quantified with the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific, Q32854, Waltham, MA, USA). Libraries were sequenced as 50bp, paired-end reads on an Illumina HiSeq 4000 platform (San Diego, CA, USA) to a target depth of 100 million raw reads for H3K27me3 (broad mark) and 50 million raw reads for H3K327ac, H3K4me1, and H3K4me3 (narrow marks) and input samples.

Library Mapping and Read Filtering

Adapters were removed and reads trimmed using Trim-Galore (Krueger, 2019). Reads were mapped to EquCab3.0 with BWA-mem (Li, 2013) using the mapping script from the slurm-genotyping pipeline (https://github.com/esrice/slurm-genotyping). PCR duplicates were marked and removed with samtools (Li et al., 2009). Additionally, read pairs that were unmapped, non-primary alignments, or had a mapping alignment quality score (MAPQ) of less than 30 were removed with samtools (Li et al., 2009) to ensure that only high-quality reads remained for peak calling. The target usable fragment count was 45 million for H3K27me3 and 20 million for the remaining marks and input samples as outlined in the ENCODE project (https://www.encodeproject.org/chip-seq/histone/). An H3K27ac adipose sample had less than half of the targeted usable fragments, so an additional library was prepared and sequenced. The filtered reads from both rounds of sequencing were merged for downstream analysis.

Stallion Peak Calling and Tissue-Unique Peak Identification

Peaks, representing regions of read pileup, were identified using the pipeline established by Kingsley et al. (https://faang.org/ebi/ftp.ebi.ac.uk/faang/ftp/protocols/analyses/UCD_SOP_processing_and_analyzing_equine_PE_ChIP_data_20201230.pdf). MACS2 (Zhang et al., 2008) was used to call peaks across all four histone marks, and SICERpy (https://github.com/dariober/SICERpy, a wrapper for SICER from Zang et al., 2009) was used to call peaks for H3K27me3. Paired-end (PE) reads were used for MACS2 peak calling, while only the first reads (R1) of the stallion libraries were used for peak calling in SICERpy (Zang et al., 2009) as this software has yet to be optimized for PE libraries. The effective genome size, or genome fraction for SICERpy (Zang et al., 2009), was determined by merging all input samples and determining the percentage of the genome covered by the merged bam file. The genome size for MACS2 (Zhang et al., 2008) was equivalent to the sum of EquCab3.0 chromosome lengths (chr1-chrX and the mitochondrial chromosome (1660bp)). The bioinformatic parameters used in peak calling for each mark are defined in **Table 2**. Overlapping peaks significant in one biological replicate and at least called (enriched) in the other for a given mark and tissue were determined using the output from MACS2 (Zhang et al., 2008) and bedtools intersect (Quinlan & Hall, 2010) with default settings. Only peaks significant/enriched in both biological replicates were considered in the study. Peaks unique to a single tissue for a given mark were identified with bedtools (Quinlan & Hall, 2010). Microsoft Excel (Microsoft Corporation, 2018) was used to calculate the correlations between average usable reads and total peak number in addition to correlations between total peak number and tissue-unique peaks.

Data Availability

The signal tracks (bigwig) generated for each sample can be download from https://equinegenomics.uky.edu/faangHorses_ChIP-Seq.html. Additionally, the combined peaks and those identified for each biological replicate will be available as BED tracks which can be found at FAANG.org under BioSampleIDs SAMEA9462146 and SAMEA9462145.

Comparison of Peaks in Mares and Stallions

Eight of the 80 tissues in the equine FAANG biobank from two Thoroughbred mares previously underwent ChIP-seq analysis for H3K27ac, H3K27me3, H3K4me1, and H3K4me3 (Burns et al., 2018; Kingsley et al., 2020). The prioritized tissues included adipose, brain parietal cortex (brain), left ventricle (heart), lamina, liver, lung, skeletal muscle, and ovary (Kingsley et al., 2020). The ChIP-seq libraries consisted of 50bp single-end (SE) reads generated by Diagneode's ChIP-seq Profiling Service (Diagenode, Cat# G02010000, Liège, Belgium; Kingsley et al., 2020). Both the raw and analyzed datasets corresponding to the mare ChIP-seq project can be found within the FAANG data portal under BioSample IDs SAMEA104728877 and SAMEA104728862, as well as in the European Nucleotide Archive (ENA) under project accession PRJEB35307.

The raw fastq files from the mare ChIP-seq project were retrieved, and reads were trimmed, mapped, and filtered as previously described. To match the SE libraries of the mares, the first reads (R1) from the stallion libraries were processed as SE libraries and trimmed, mapped, and filtered in the same manner as the mare libraries. The resulting bam files for the mare and stallion ChIP and input samples were converted to bigwig files

using deeptools bamCoverage (Ramírez et al., 2016) with default settings. The S3V2_IDEAS_ESMP pipeline (https://github.com/guanjue/S3V2_IDEAS_ESMP) was employed to normalize libraries for sequencing depth and background noise (Xiang et al., 2021). The S3V2_IDEAS_ESMP pipeline uses the S3norm (Xiang et al., 2020) normalization method to create the signal tracks used for peak calling. The signal tracks are generated in a manner similar to MACS2 (Zhang et al., 2008), but the Poisson model that adjusts for local background in MACS2 is replaced with a negative binomial model to allow for flexibility in the estimates of the mean and variation within the model (Xiang et al., 2020).

The signal tracks generated by the S3V2_IDEAS_ESMP pipeline contain 200 bp bins with *P*-value scores representing the significance of the signal in each bin (Xiang et al., 2020). The signal tracks were converted from bigwig format to bedgraph format using UCSC's bigWigToBedGraph package (Kent et al., 2010). Peaks were called from the bedgraph files using MACS2 bdgpeakcall (Zhang et al., 2008) for H3K27ac, H3K4me1, and H3K4me3 and MACS2 bdgbroadcall (Zhang et al., 2008) for H3K27me3 with a p-value cutoff of 0.05. Peaks identified as significantly enriched in both biological replicates of a given sex for a given mark and tissue were identified using bedtools (Quinlan & Hall, 2010). Only the peaks identified in both of the biological replicates within each sex were considered in the study.

Peaks unique to a tissue for a given mark and sex were identified using bedtools (Quinlan & Hall, 2010). The overlap of peaks called for the mares and stallions were also assessed within a given mark and tissue. Furthermore, the correlations between usable

reads, peak number, and unique peak number were evaluated for both the mares and stallions using Microsoft Excel (Microsoft Corporation, 2018).

Results

Sequencing Depth and Read Filtration of Paired-End Stallion Libraries

On average, each stallion sample had 52 million (M) raw reads for H3K27ac and H3K4me1, 55 M for H3K4me3, and 134 M raw reads for H3K27me3. Filtering removed PCR duplicates, unmapped, and low-quality reads to create a set of reads used for peak calling, termed usable read pairs. The average number of usable reads pairs was 28 M for H3K27ac, 32 M for H3K4me1, 30 M for H3K4me3, and 68 M for H3K27me3. Each stallion sample had an input sample with an average of 34 M reads used to remove background noise during peak calling. Despite generating over 215 M raw read pairs between the two H3K27ac_Adipose_AH3 libraries, less than 12 M usable read pairs were available for peak calling. Additionally, the number of usable read pairs fell below the target of 20 M usable read pairs in the other H3K27ac_Adipose replicate, in one H3K4me3_Lamina replicate, and in the Muscle_AH3 replicates for H3K27ac, H3K4me3, and the input sample. Six of the H3K27me3 samples fell below the 45 M usable read pair target for this mark, yet all samples had at least 43 M usable read pairs (Supplementary Table 3.2).

Quantifying Peaks for Paired-End Stallion Libraries

On average, each tissue had 76,668 H3K27ac peaks, 120,309 H3K4me1 peaks, and 33,969 H3K4me3 peaks (**Figure 1**). Similar peak widths were observed across the narrow marks, with average peak widths of 1,360 bp-, 1,219 bp, and 1,535 bp for

H3K27ac, H3K4me1, and H3K4me3, respectively (**Table 3).** The number of peaks called for H3K27me3 varied considerably based on the software used for peak calling. MACS2 identified an average of 158,480 H3K27me3 peaks while SICERpy called an average of 32,315 H3K27me3 peaks across tissues (**Figure 1**). The average peak width of MACS2 H3K27me3 peaks was 1,650 bp while the average peak width of SICERpy H3K27me3 peaks was 18,466 bp (**Table 3**).

H3K4me3 had the lowest genome coverage across tissues with an average of 2.2% coverage. H3K27ac peaks covered an average of 4.1% of the genome across tissues, and H3K4me1 covered approximately 6.2% of the genome in each tissue. H3K27me3 peaks called by MACS2 covered, on average, 10.4% of the genome, while H3K27me3 peaks called by SICERpy covered ~24.3% of the genome in each tissue (Supplementary Table 3.3).

Tissue-Unique Peaks in Paired-End Stallion Libraries

The brain had the largest proportion of unique peaks for H34me1 and H3K27ac. Twenty-seven percent of H3K4me1 peaks and 33% of H3K27ac peaks identified in the brain were unique to the tissue. Nearly 50% of H3K4me3 peaks in the testis were unique. The liver displayed the greatest proportion of unique peaks for H3K27me3 in peaks called with both MACS2 and SICERpy. Muscle consistently demonstrated low uniqueness across all four marks (**Figure 1**). The majority of peaks called for each histone mark were shared across tissues.

Quantifying Peaks for Single-End Stallion Reads

The first read (R1) of read pairs was used to create mock single-end sequencing in the stallion data. The average number of usable reads for the stallion SE analysis for H3K27ac, H3K4me1, H3K4me3, and H3K27me3 was 26 M, 30 M, 24 M, and 63 M, respectively. Samples that fell below the 20 M usable read target included both H3K27ac adipose samples, one H3K27ac muscle sample, one H3K4me3 muscle sample, and one H3K4me3 lamina sample. Each of these samples had a usable read counts ranging from 11.7 M reads to 18.4 M reads. For H3K27me3, four samples fell below the 45 M usable read target. This includes an H3K27me3 heart sample with 42 M reads, an H3K27me3 muscle sample with 41 M reads, and both H3K27me3 testis samples with 41 M reads each (Supplementary Table 3.4)

Quantifying Peaks for Single-End Mare Reads

The number of raw reads generated from the mare samples was generally less than that generated for the stallions. The mare samples had approximately 42 M, 49 M, 41 M, and 88 M raw reads for H3K27ac, H3K4me1, H3K4me3, and H3K27me3, respectively. This resulted in average usable read counts of 23 M for H3K27ac, 26 M for H3K4me1, 17 M for H3K4me3, and 26 M for H3K27me3. Four H3K27ac, two H3K4me1, and eight H3K4me3 samples fell below their 20 M usable read target. All mare H3K27me3 samples fell below their 45 M usable read target (Supplementary Table 3.5).

Peaks Called from Normalized, Single-End Stallion Reads

An average of 66,586 H3K27ac, 48,605 H3K4me1, 26,273 H3K4me3, and 14,283 H3K27me3 peaks were identified across tissues in the normalized SE stallion analysis (**Figure 2**). The average peak widths of the SE stallion peaks were 711 bp, 455 bp, 1,048 bp, and 2,337 bp for H3K27ac, H3K4me1, H3K4me3, and H3K27me3, respectively (**Table 3**). No more than 2.6% of the genome was covered by a single histone mark in a given tissue. The greatest genome coverage was observed for H3K27ac with an average coverage of 1.97%. H3K4me1 peaks covered the smallest proportion of the genome with an average coverage of 0.99% (Supplementary Table 3.6). The greatest proportion of tissue-specific peaks were identified in the lung (21%) for H3K27me3, the brain (43%) for H3K27ac, the liver (44%) for H3K4me1, and the testis (46%) for H3K4me3 (**Figure 2**).

Peaks Called from Normalized, Single-End Mare Reads

In the mares, an average of 65,643 H3K27ac, 47,490 H3K4me1, 21,055 H3K4me3, and 6,719 H3K27me3 peaks were called across tissues (**Figure 2**). The average peak width was 682 bp for H3K27ac, 442 bp for H3K4me1, 1015 for H3K4me3, and 1,708 bp for H3K27me3 in the mares (**Table 3**). The H3K27ac peaks had the greatest genome coverage for all mare samples with an average coverage of 1.86%. H3K27me3 peaks covered the smallest portion of the genome with an average of 0.48% (Supplementary Table 3.6). In the mares, the liver had the greatest proportion of unique peaks in H3K27ac (34%), H3K4me1 (54%), and H3K4me3 (13%). The adipose had the greatest number of unique peaks for H3K27me3 (**Figure 2**).

Direct Comparison of Mare and Stallion Regulatory Elements

Large discrepancies were observed between the number of peaks unique due to sex for a given tissue and mark. For the broad mark, H3K27me3, the number of peaks identified as unique in the mares never exceeded 24% for a given tissue while in the stallion over 57% of peaks were identified as unique in all tissues. The difference in the percent of unique peaks identified between the mares and stallions resembled the percent difference of total peaks called between the sexes for most tissues. A similar trend was observed across the other marks (Supplementary Table 3.7).

Correlation of Usable Reads, Peak Number, and Unique Peak Number

Positive correlations existed between the number of reads used for peak calling and the number of peaks called across all datasets. Correlation coefficients (r) between usable reads and total peaks called ranged from 0.17 to 0.83 across marks and datasets. The normalization for sequencing depth in the stallion SE dataset resulted in a smaller correlation between usable reads and peaks called for H3K27ac, HK27me3, and H3K4me1, but not for H3K4me3. Correlations ranging between 0.20 and 0.60 remained in the stallion SE dataset even after normalization. Furthermore, out of all three datasets (stallion PE, stallion SE, and mare SE), the mare SE dataset that had also been normalized for sequencing depth had the greatest correlations between usable reads and peaks called for each mark. The number of unique peaks in a tissue was also highly correlated with the total peaks called for that tissue. Correlations between total peaks and unique peaks ranged from 0.52 to 0.96 across datasets and histone marks (**Table 4**).

Discussion

Putative Regulatory Elements in the Stallion Called from Paired-End ChIP-seq

On average, each tissue in the stallions had over 250,000 peaks corresponding to regulatory elements. The most common regulatory elements identified were repressors or silencers represented by H3K27me3. This broad peak covered the greatest percentage of the genome in the PE dataset with some tissues having repressive marks covering as much as 32% of the genome. The enzymes that are involved in trimethylation of H3K27 often follow a positive feedback loop in which the presence of H3K27me3 increases trimethylation of nearby histones (Oksuz et al., 2018; Schmitges et al., 2011). Although H3K27me3 is known to create broad peaks, both the number and width of the H3K27me3 peaks identified in the study varied considerably based on the peak calling software. MACS2 identified nearly five times as many H3K27me3 peaks as SICER, yet the average width of the peaks called by SICER was over ten times larger than those called by MACS2.

Similar observations were made by Steinhauser et al. (2016) in which SICER called considerably wider peaks than software based on MACS2 peak calling. As a gold standard method for peak calling has yet to be established, simulated ChIP-seq datasets are required to examine the sensitivity and specificity of peak calling software. On simulated datasets, SICER outperformed 10 different peak-calling tools for both identifying true peaks and limiting false positives when examining a broad-peaked histone modification (Steinhauser et al., 2016). Since SICER was designed to better capture broad and diffuse peaks, such as those of H3K27me3, peaks called by SICER may better represent the proportion of the genome repressed due to H3K27me3 (Xu et al.,

2014). MACS2 is most often used to identify narrow peaks suggesting that H3K27me3 peaks called by MACS2 may represent regions of the genome with the strongest H3K27me3 signals. Therefore, which H3K27me3 peak set is superior will likely be defined by the application of the dataset. For example, the greater genome coverage of peaks called by SICERpy may be more valuable in determining which genes are inactive or poised while the sharper H3K27me3 peaks called by MACS2 may be more helpful in identifying primary binding sites for the polycomb complexes, PCR1 and PCR2, involved in establishing facultative heterochromatin (Oksuz et al., 2018; Schmitges et al., 2011). Since PCR2 deposits addition trimethylation radiating outward from an initial site of H3K27me3, the sharper peaks may represent positions in the genome that are more often trimethylated or potential initiator loci for facultative heterochromatin (Oksuz et al., 2018; Schmitges et al., 2011).

The number of peaks called for each mark varied by tissue, yet muscle samples consistently had fewer peaks across all histone marks in the stallion data. This could be due to using a smaller amount of chromatin for library preparation; however, the same amount of chromatin was used in the mare analysis published by Kingsely et al. (2020) and no reduction in peak number was observed across muscle samples. The muscle samples from one of the stallions, AH3, failed to produce the targeted number of usable reads for H3K27ac, H3K27me3, H3K4me3, and the input sample, yet in all cases, the number of usable reads was within 20% of the target. Even so, moderate positive correlations between the number of reads used for peak calling and the number of peaks called suggest that additional ChIP and/or sequencing may improve the identification of regulatory elements in the muscle.

Despite the general increase in peaks called from libraries of greater sequencing depth, it is unlikely that the additional peaks called in the more deeply sequenced libraries were false positives. The use of input samples to normalize for background noise and the requirement of peaks to be enriched in both biological replicates reduces the likelihood of insignificant peaks in the final dataset. Rather, the positive correlation between the number of usable reads and peaks called suggests that the ideal sequencing depth has not been reached in many of the samples. Although most of the stallion samples produced enough usable reads to achieve the thresholds established by ENCODE (https://www.encodeproject.org/chip-seq/histone/), 20 M for narrow marks and 45 M for broad marks, our data suggest that this threshold may not be sufficient in all tissue types. Additional research will need to be done to determine at which point additional reads no longer provide additional peak calls, which may differ across tissue types. Even if all regulatory elements in the assayed tissues were not captured, those that were provide valuable information into the genome function of those tissues.

In addition to a moderate correlation between usable reads and peaks called, a strong positive correlation was identified between the number of peaks called in a tissue and those identified as unique to that tissue. This correlation makes it difficult to determine if these uniquely identified peaks represent biological differences in the regulatory elements of tissues. Yet, in the case of the H3K27ac and H3K4me1 brain samples, the highest percentage of uniqueness is observed in a tissue with fewer peaks than the other samples. Many studies, including our own (see Chapter 2), have identified a large number of transcripts that are preferentially expressed in the brain (Ramsköld et al., 2009; Uhlén et al., 2015). Therefore, the unique H3K27ac and H3K4me1 peaks

identified in the brain may represent some of the regulatory elements involved in activating these differentially expressed genes. In H3K4me3, the testis had a strikingly large proportion of promoter peaks identified as unique. Although many of these unique peaks could be due to the larger number of total peaks called, the testis also has a great deal of tissue specific gene expression as demonstrated by us (see Chapter 2) and others (Ramsköld et al., 2009). Examining how the tissue-unique peaks correlate with transcriptome data could help clarify which regulatory elements are likely to be tissue-specific and which are likely shared among tissues.

Comparing Peaks Called in Mares and Stallions from Normalized, Single-End Libraries

Previous work has identified differences in the ability to identify peaks from single-end and paired-end sequencing (Zhang et al., 2015). Similar differences can be observed between the total number of peaks identified in the PE stallion peaks presented in this study and the finding published by Kingsley et al. (2020) corresponding to peaks called from single-end sequencing. Therefore, to compare the regulatory elements identified in mares and stallions, the stallion data was treated as single-end libraries by only examining the first read of each read pair. As correlations were identified between the number of reads used for peak calling and the number of peaks called, the S3V2_IDEAS_ESMP pipeline (Xiang et al., 2021) was employed to normalize for sequencing depth across samples in the mares and stallions. This software employs a normalization method that generates scaling factors for each sample that are based on the signal intensity in both enriched (peak) regions and background regions shared across samples (Xiang et al., 2020). By adjusting the signal enrichment in both background

regions and shared peak regions, the impact of variation in the signal-to-noise ratio, as defined by the input sample, and the sequencing depth across samples is supposed to be minimized. Signal tracks that represent the significance of reads enrichment across 200bp windows is generated using a method similar to MACS2; however, the Poisson model employed by MACS2 is replaced with a negative binomial model that allows for greater flexibility in estimation of the mean and variances used in the model (Xiang et al., 2020). Since the S3V2 pipeline identifies signal enrichment in a similar manner to MACS2, MACS2 was used to call peaks for both the narrow and broad histone marks after normalization.

Normalization of the SE libraries resulted in a smaller number of peaks being called in the mares and stallions compared to the PE stallion data and the original data published by Kingsley et al. (2020). Conversely, the number of unique peaks was greater in nearly all tissues and marks in the SE normalized data than in the PE stallion data and the original mare data. The increase in unique peaks could be attributed to the smaller number of peaks identified across tissues and the resulting reduction in genome coverage. Additionally, the normalization method forces peaks to fall into 200 bp windows requiring peaks to overlap by a minimum of 200 bp to be considered shared. The original pipeline used to assess peaks in the mares (Kingsley et al., 2020) and the stallions (PE dataset) identifies peaks continuously across the genome and considers peaks shared if they display any degree of overlap. Therefore, it is not surprising that the reduced genome coverage and strict classification for overlapping peaks in the SE normalized data sets resulted in more peaks being identified as unique across tissues.

Although the S3V2_IDEAS_ESMP pipeline was chosen to normalize for sequencing depth, a clear reduction in the correlation between usable reads and peaks called was not observed. In half of the marks, the SE normalized stallion data had a smaller correlation between reads and peaks than in the PE stallion data. Interestingly, the highest correlations between usable reads and peaks called was observed in the mare data normalized for sequencing depth. These finding suggests that additional library preparation may be need in the mares' samples, but more so, that the methods used to correct for sequencing depth were not adequate. Due to variation in sequencing depth and total peaks called between the mares and stallions, direct comparisons between peaks called across sexes did not provide meaningful insight into the role of sex on the presence of regulatory elements.

The annotation of regulatory elements has proven beneficial in characterizing the function of the genome and associating genomic variation with disease in humans (The ENCODE Project Consortium, 2012; Gupta et al., 2017; Warburton et al., 2016; Vinagre et al., 2013). In this study, hundreds of thousands of putative regulatory element were identified across tissues in the horse providing valuable information into the function of the equine genome. The data from the previously published ChIP-seq analyses in the mares has already aided in the identification of variants associated with distichiasis and the characterization of centromere sliding in horses (Kingsley et al., 2020; Hisey et al., 2020; Cappelletti et al., 2022). The ChIP-seq analyses in the stallion provide additional support for the annotation of regulatory elements present in the tissues of adult horses. Although peaks unique due to sex or tissue type could not be clearly defined, these analyses demonstrate some of the shortcomings in the current methodology used for

identifying regulatory elements. As previously suggested, sequencing depth, sequencing method, and peak calling software have large impacts on the outcome of ChIP-seq studies (Steinhauser et al., 2016; Zhang et al., 2015; Xiang et al., 2020). These artifacts of data processing impair the ability to identify biological differences across datasets. While much progress has been made in our understanding of genome function, technological advancements will be necessary for comparative studies into genome regulation.

Table 3.1. Parameters for ChIP Extraction and Shearing

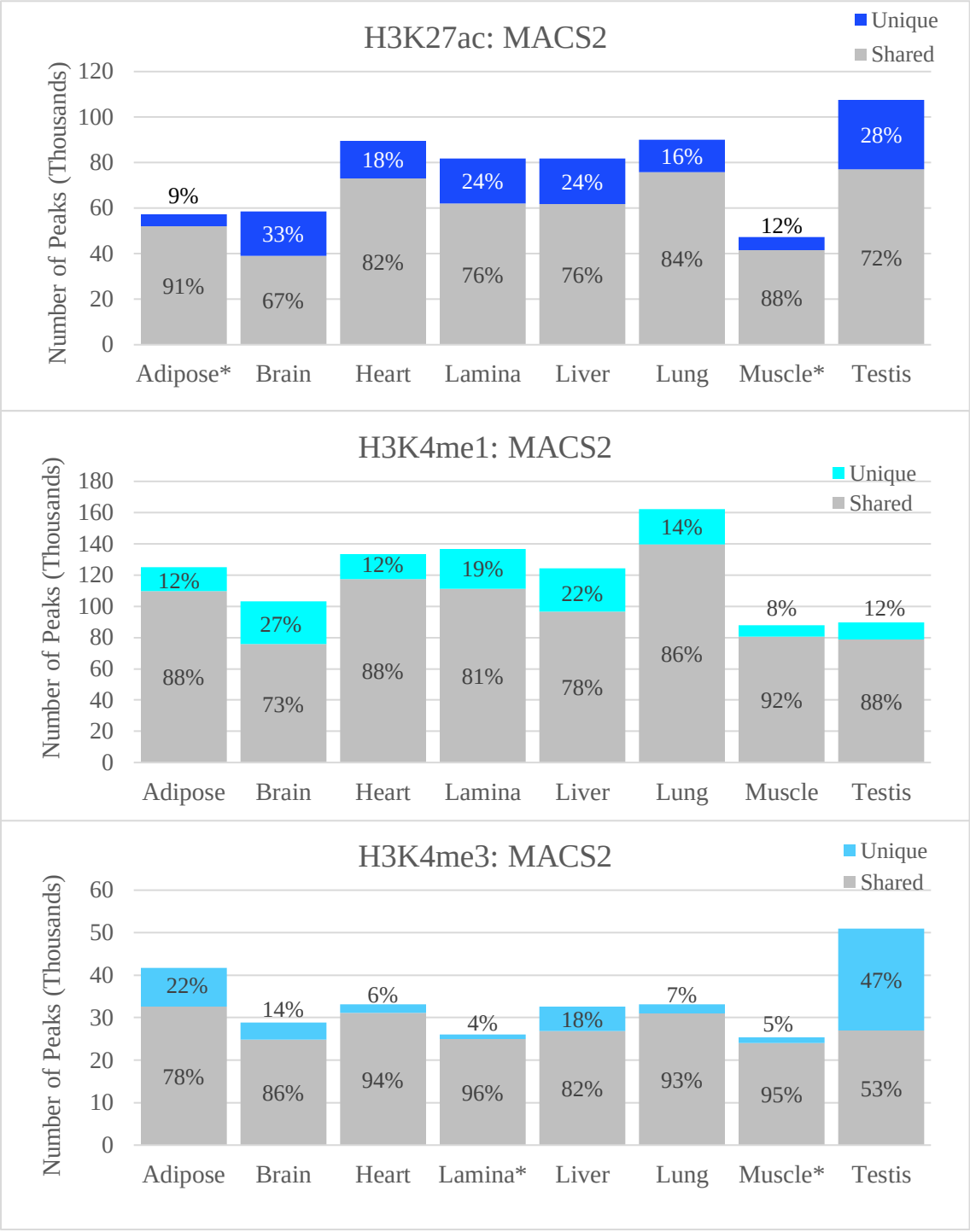
	Adipose	Brain	Heart	Lamina	Liver	Lung	Muscle	Testis
Starting Tissue (mg)	208	109	103	95	55	53	106	105
Homogenization Time (min)	8	5	5	5	5	5	5	5
Fixation Time (min)	10	9	9	9	9	9	9	12
Shearing Volume (uL)	400	1500	1800	1800	1500	1500	1500	1300
Shearing Cycles	5 x 8 cycles[1]	10	10	10	8	10	10	8
Chromatin per IP (ng)	400	700	800	900	1500	1500	280	1500

[1]Chromatin from adipose was sheared for five times for 8 cycles

Table 3.2. Software Parameters for ChIP-seq Peak Calling in MACS2 and SICERpy

Software	Parameter	H3K4me1	H3K4me3	H3K27ac	H3K27me3
MACS2	Size	Narrow/ Intermediate	Narrow	Narrow	Broad
	Size Flag	none	none	none	---broad
	FDR[1]	0.05	0.01	0.01	0.1
	Genome size	2,409,159,894	2,409,159,894	2,409,159,894	2,409,159,894
SICERpy[2]	Gap Size	n/a	n/a	n/a	4
	Window Size	n/a	n/a	n/a	200
	Genome Fraction	n/a	n/a	n/a	0.973

[1]The FDR cutoff dictates both peak number and peak width in MACS2, so histone marks with broader peaks have looser FDR cutoffs (https://github.com/hbctraining/Intro-to-ChIPseq). H3K4me1 was at one point considered to have broad peaks but has since been determined to have an intermediate peak width (Kingsley et al., 2020).
[2]SICERpy was only used to call peaks for the broad mark, H3K27me3


H3K27ac: MACS2
Unique
Shared
Number of Peaks (Thousands)
120
100
80
60
40
20
0
9%
33%
18%
24%
24%
16%
12%
28%
91%
67%
82%
76%
76%
84%
88%
72%
Adipose*
Brain
Heart
Lamina
Liver
Lung
Muscle*
Testis
H3K4me1: MACS2
Unique
Shared
Number of Peaks (Thousands)
180
160
140
120
100
80
60
40
20
0
12%
27%
12%
19%
22%
14%
8%
12%
88%
73%
88%
81%
78%
86%
92%
88%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Testis
H3K4me3: MACS2
Unique
Shared
Number of Peaks (Thousands)
60
50
40
30
20
10
0
22%
14%
6%
4%
18%
7%
5%
47%
78%
86%
94%
96%
82%
93%
95%
53%
Adipose
Brain
Heart
Lamina*
Liver
Lung
Muscle*
Testis

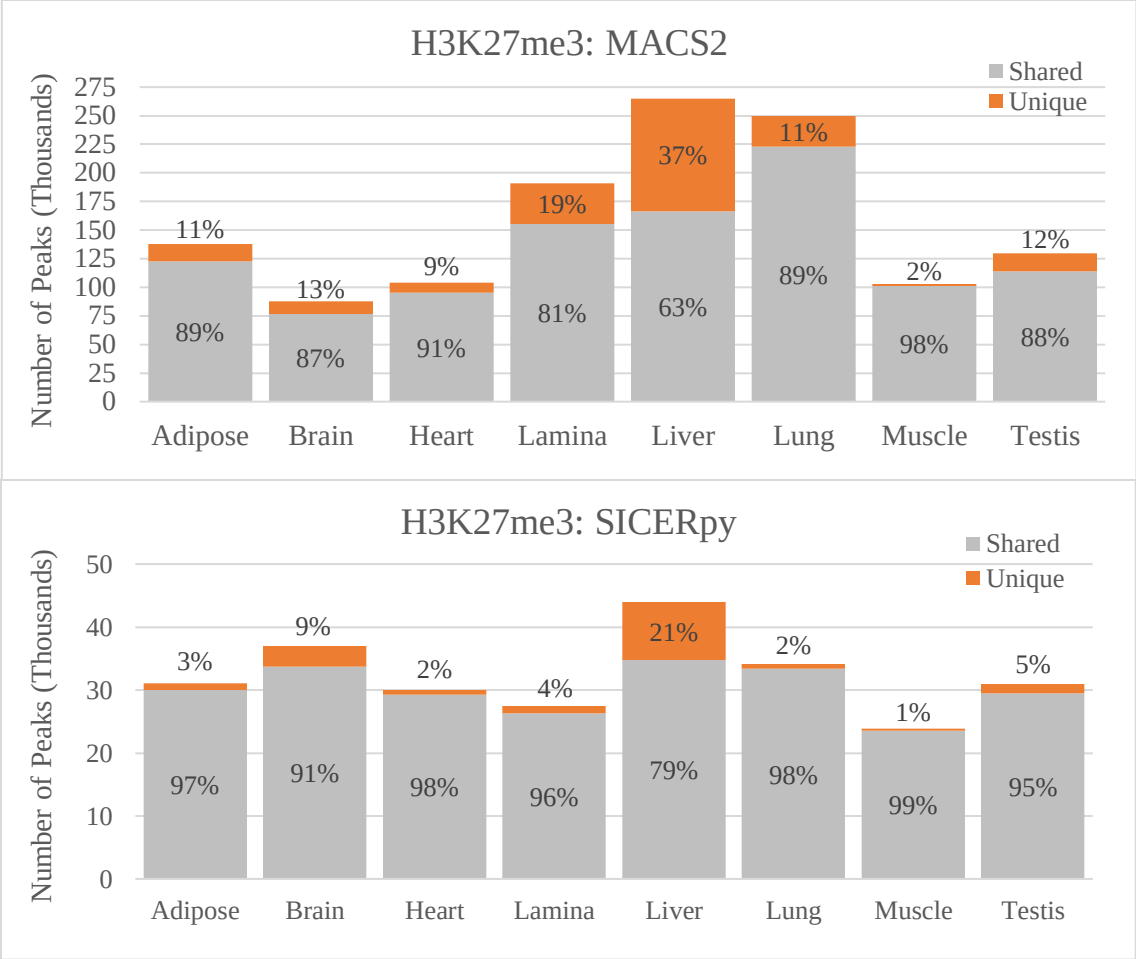
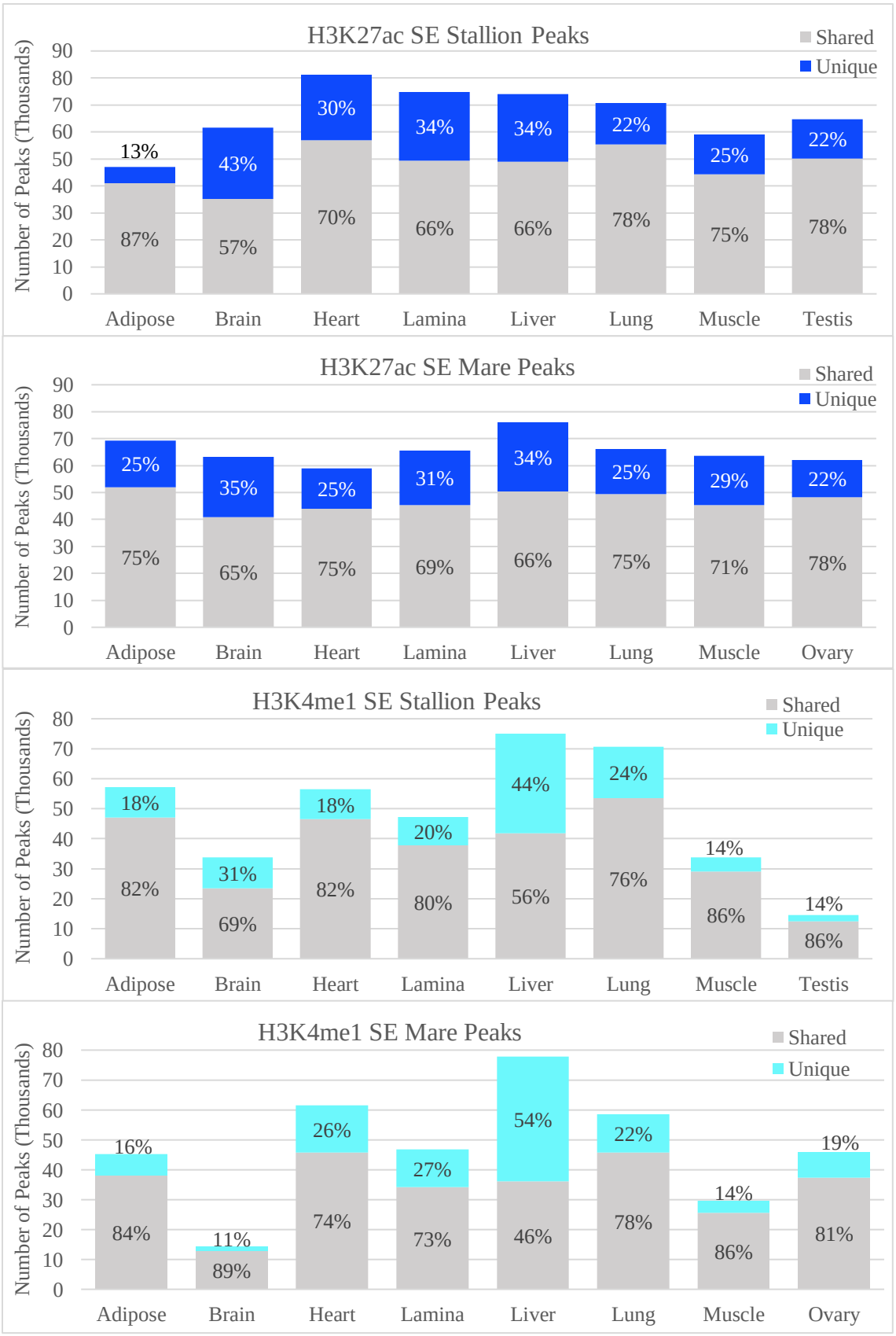


Figure 3.1. Tissue-specific peaks by histone mark. Each graph represents the peaks detected in tissues for the corresponding marks: H3K4me1, H3K4me3, H3K27ac, and H3K27me3. The gray area represents peaks detected in two or more tissues, while the colored area represents peaks unique to the respective tissue. The percentages correspond to the percent of peaks shared and unique, respectively. *Tissue did not meet targeted usable read count

Table 3.3. Average Peak Width of Peaks Called from Paired-End and Single-End Libraries

Average Peak Width (bp)					
Analysis	H3K27ac	H3K4me1	H3K4me3	H3K27me3-M	H3K27me3-S*
Stallion PE	1360	1219	1535	1650	18466
Stallion SE	711	455	1048	2337	-
Mare SE	682	443	1015	1708	-

* H3K27me3-M represents peaks called by MACS2 while H3K27me3-S represent peaks called by SICERpy. SICERpy was only used to call peaks for the PE analysis.


H3K27ac SE Stallion Peaks
Shared
Unique
Number of Peaks (Thousands)
0
10
20
30
40
50
60
70
80
90
13%
43%
30%
34%
34%
22%
25%
22%
87%
57%
70%
66%
66%
78%
75%
78%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Testis
H3K27ac SE Mare Peaks
Shared
Unique
Number of Peaks (Thousands)
25%
35%
25%
31%
34%
25%
29%
22%
75%
65%
75%
69%
66%
75%
71%
78%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Ovary
H3K4me1 SE Stallion Peaks
Shared
Unique
Number of Peaks (Thousands)
0
10
20
30
40
50
60
70
80
18%
31%
18%
20%
44%
24%
14%
14%
82%
69%
82%
80%
56%
76%
86%
86%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Testis
H3K4me1 SE Mare Peaks
Shared
Unique
Number of Peaks (Thousands)
16%
11%
26%
27%
54%
22%
14%
19%
84%
89%
74%
73%
46%
78%
86%
81%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Ovary

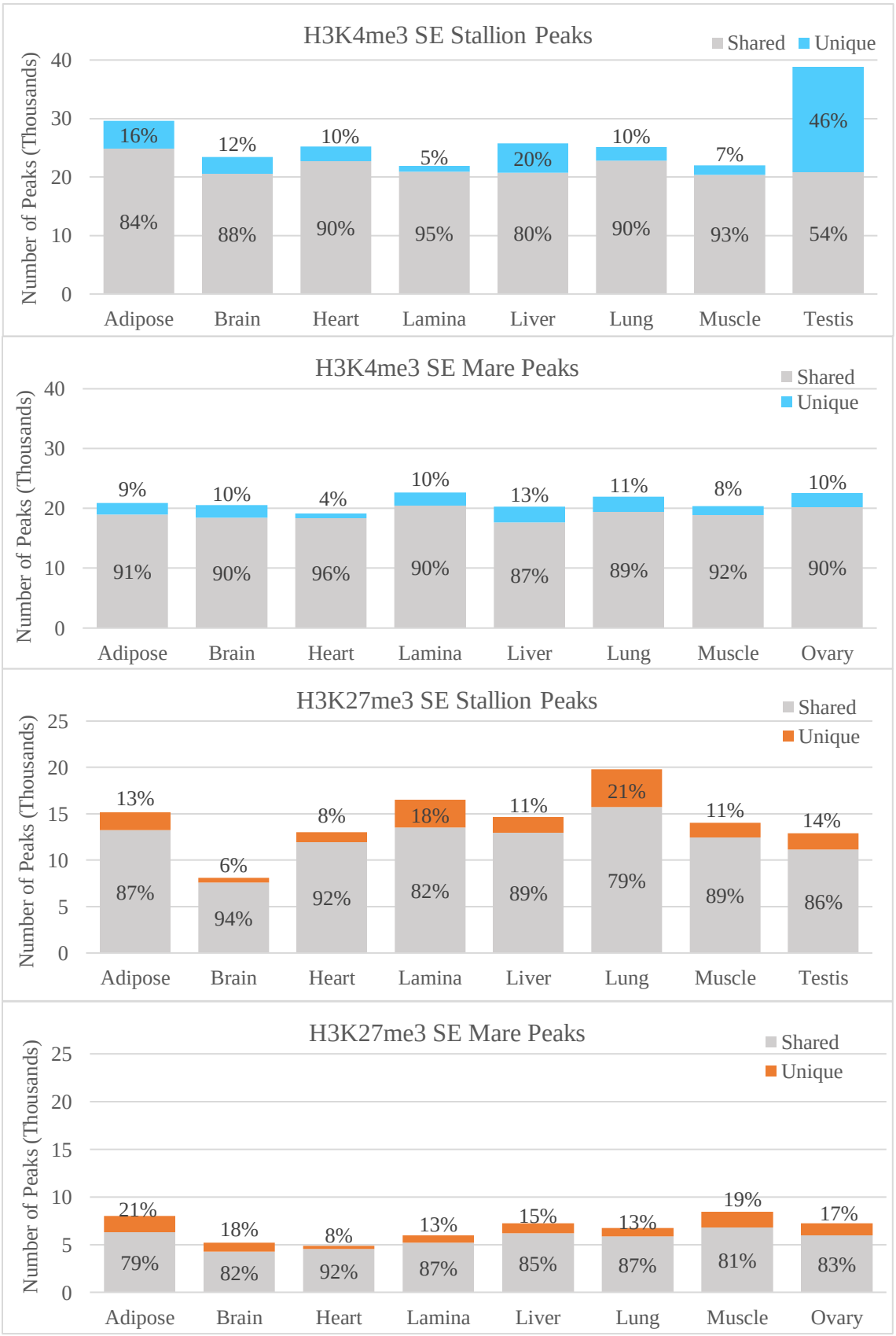

H3K4me3 SE Stallion Peaks
Shared
Unique
Number of Peaks (Thousands)
0
10
20
30
40
16%
12%
10%
5%
20%
10%
7%
46%
84%
88%
90%
95%
80%
90%
93%
54%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Testis
H3K4me3 SE Mare Peaks
Shared
Unique
Number of Peaks (Thousands)
0
10
20
30
40
9%
10%
4%
10%
13%
11%
8%
10%
91%
90%
96%
90%
87%
89%
92%
90%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Ovary
H3K27me3 SE Stallion Peaks
Shared
Unique
Number of Peaks (Thousands)
0
5
10
15
20
25
13%
6%
8%
18%
11%
21%
11%
14%
87%
94%
92%
82%
89%
79%
89%
86%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Testis
H3K27me3 SE Mare Peaks
Shared
Unique
Number of Peaks (Thousands)
0
5
10
15
20
25
21%
18%
8%
13%
15%
13%
19%
17%
79%
82%
92%
87%
85%
87%
81%
83%
Adipose
Brain
Heart
Lamina
Liver
Lung
Muscle
Ovary

Figure 3.2. Single-End, Normalized Peaks in Tissues from Stallions and Mares
The histone mark and sex are in the title of each graph. Bar heights denote the total peaks called for each tissue. The gray portion of the bar represent peaks shared amongst at least two tissues for given mark and sex. The colored portion of the bar represents peaks unique to a tissue for a given mark and sex. The percentages present within the gray portion and within or above the colored portion of the bars denote the percentage of peaks shared and unique, respectively.

Table 3.4 Correlations of Peak Number with Usable Reads and Unique Peak Number

Correlation of Usable Reads[1] and Total Peak Number				
Analysis	H3K27ac	H3K27me3	H3K4me1	H3K4me3
Stallion PE	0.62	0.71	0.17	0.40
Stallion SE	0.49	0.54	0.20	0.60
Mare SE	0.63	0.70	0.83	0.80
Correlation of Total Peak Number and Unique Peak Number				
Analysis	H3K27ac	H3K27me3	H3K4me1	H3K4me3
Stallion PE	0.78	0.80	0.52	0.92
Stallion SE	0.74	0.93	0.81	0.96
Mare SE	0.71	0.90	0.86	0.67

[1]Usable reads were calculated as the average number reads used for peak calling across biological replicates for a given dataset.

Supplementary Table 3.1. Chromatin and Antibody Amounts Used for ChIP of Stallion Tissues

Tissue	Chromatin IP (ng)	Antibody Amount (µg)			
		H3K27ac	H3K4me1	H3K4me3	H3K27me3
Adipose	200	1	0.5	0.5	1
Brain	700	0.5	0.5	0.5	1
Heart	800	0.5	0.5	0.5	1
Lamina	900	1	0.5	0.5	1
Liver	1500	2	1	1	2
Lung	1500	2	1	1	1
Muscle	280	1	0.5	0.5	0.5
Testis	1500	2	1	1	2

Supplementary Table 3.2. The Number of Raw Read Pairs Generated and Usable Read Pairs Remaining After Filtering for Stallion Samples

Tissue_Replicate	H3K27ac		H3K4me1		H3K4me3		H3K27me3		Input	
	Raw Read Pairs	Usable Read Pairs	Raw Read Pairs	Usable Read Pairs	Raw Read Pairs	Usable Read Pairs	Raw Read Pairs	Usable Read Pairs	Raw Read Pairs	Usable Read Pairs
Adipose_AH3*	68,730,027	**9,159,119**	87,399,621	44,861,149	53,728,350	28,666,715	163,649,254	61,348,040	57,287,919	33,439,495
Adipose_AH4	64,121,448	**19,626,387**	73,755,325	22,974,469	71,588,828	30,630,416	172,819,418	77,345,817	56,710,987	31,576,540
Brain_AH3	44,153,723	28,765,325	51,902,716	39,350,197	48,165,691	30,241,426	89,113,219	51,491,049	**43,318,326**	30,755,494
Brain_AH4	53,550,924	37,160,660	45,941,617	33,746,787	44,199,605	31,021,321	123,828,178	70,342,831	45,370,082	29,343,590
Heart_AH3	51,902,076	32,670,198	43,860,168	31,024,746	51,509,237	33,772,055	115,749,332	46,118,523	**43,193,521**	31,249,453
Heart_AH4	45,090,183	28,984,929	64,302,815	36,581,050	44,069,238	32,085,936	141,578,920	65,733,036	59,163,545	41,928,341
Lamina_AH3	40,535,808	23,822,100	51,986,630	23,740,973	48,277,375	24,566,917	91,980,974	58,479,470	**44,456,575**	32,288,689
Lamina_AH4	51,953,800	24,587,647	49,361,910	29,678,316	41,990,329	**17,498,222**	175,850,495	107,498,910	57,425,242	38,115,284
Liver_AH3	55,697,969	35,488,532	45,896,092	30,855,813	69,512,707	39,270,074	126,294,140	68,466,690	66,316,277	44,077,398
Liver_AH4	46,094,286	25,054,182	40,668,291	28,444,945	58,847,906	30,313,903	192,622,865	102,590,156	51,996,032	36,917,707
Lung_AH3	51,223,693	36,116,504	54,431,763	38,103,476	45,017,929	32,393,855	120,694,401	66582349.5	50,253,855	36,330,557
Lung_AH4	38,186,321	26,003,430	46,855,312	32,873,458	48,153,388	33,420,217	174,764,785	85,469,928	**44,634,249**	31,782,488
Muscle_AH3	59,556,353	**16,214,295**	50,633,437	29,979,478	71,090,388	**19,682,061**	80,658,715	43,452,929	**43,227,165**	**18,251,509**
Muscle_AH4	59,556,353	34,272,759	50,633,437	34,358,365	71,090,388	31,284,468	211,037,887	99,609,279	**43,227,165**	30,296,940
Testis_AH3	59,238,170	42,298,232	42,281,966	29,479,553	56,479,066	27,738,020	80,574,901	44,702,585	49,793,842	36,025,563
Testis_AH4	44,952,694	31,629,947	38,900,952	27,585,434	51,773,961	34,075,725	88,305,332	45,980,036	54,636,399	36,614,620
Adipose_AH3_Rq*	146,616,053	**2,679,759**	-	-	-	-	-	-	-	-

*The "Usable Read Pairs" from both the original and resequenced (Rq) H3K27ac_Adipose_AH3 samples were merged prior to peak calling.

** The bolded cells represent samples that did not reach the usable read pair target of 45 M for H3K27me3 and 20 M for the remaining marks and input samples.

Supplementary Table 3.3. Percentage of the Genome Covered by Stallion Peaks Called from Paired-End (PE) Sequencing Reads

Genome Coverage (%) of Stallion PE Peaks					
Tissue	**H3K27ac**	**H3K4me1**	**H3K4me3**	**H3K27me3- M**	**H3K27me3-S**
Adipose	3.5	8.4	2.8	13.9	27.0
Brain	4.2	5.3	1.9	4.5	20.4
Heart	4.9	7.6	2.0	12.0	26.2
Lamina	4.1	5.7	1.6	11.3	24.0
Liver	4.9	8.8	2.1	17.4	31.9
Lung	4.4	7.9	2.0	11.2	24.4
Muscle	3.5	3.8	1.7	5.2	21.9
Testis	3.4	2.3	3.3	7.9	18.8
Average	4.1	6.2	2.2	10.4	24.3

*H3K27me3-M refers to peaks called by MACS2 and H3K27me3-S corresponds to peaks called by SICER

Supplementary Table 3.4. The Number of Raw Reads (SE) Generated and Usable Read Remaining After Filtering for Stallion SE Libraries

Tissue_Replicate	Usable Reads				
	H3K27ac	H3K4me1	H3K4me3	H3K27me3	Input
Adipose_AH3	**11,726,267**	42,727,844	26,161,927	57,544,729	31,599,110
Adipose_AH4	**18,440,913**	22,068,341	26,765,709	71,542,011	29,887,369
Brain_AH3	26,978,372	37,408,812	25,181,921	47,348,607	28,706,086
Brain_AH4	34,781,010	32,090,313	26,573,113	65,584,580	27,875,383
Heart_AH3	29,990,876	29,788,367	26,198,176	**42,563,147**	29,521,680
Heart_AH4	26,751,589	34,778,776	25,018,901	60,914,733	39,119,625
Lamina_AH3	22,446,342	22,657,405	21,539,456	54,036,862	30,634,716
Lamina_AH4	22,885,088	28,326,171	**14,646,913**	98,780,005	35,767,290
Liver_AH3	32,957,464	29,199,086	29,167,280	62,980,452	41,238,530
Liver_AH4	23,030,281	27,017,793	22,634,290	94,798,294	34,684,694
Lung_AH3	33,349,291	36,151,847	25,697,164	61,807,452	33,935,456
Lung_AH4	24,427,976	31,255,694	25,972,450	79,021,225	29,726,014
Muscle_AH3	**15,369,015**	28,618,869	**16,387,282**	**40,970,497**	**17,174,306**
Muscle_AH4	31,519,488	32,612,832	25,991,828	92,652,752	28,486,147
Testis_AH3	39,807,561	28,008,508	24,624,383	**41,072,942**	33,884,571
Testis_AH4	29,894,317	26,283,274	29,574,827	**41,757,310**	34,345,132

* The bolded cells represent samples that did not reach the usable read pair target of 45 M for H3K27me3 and 20 M for the remaining marks and input samples.

Supplementary Table 3.5. The Number of Raw Reads (SE) Generated and Usable Reads Remaining After Filtering for Mare Samples

Tissue_Replicate	H3K27ac		H3K4me1		H3K4me3		H3K27me3		Input	
	Raw Reads	Usable Reads	Raw Reads	Usable Reads	Raw Reads	Usable Reads	Raw Reads	Usable Reads	Raw Reads	Usable Reads
Adipose_AH1	42,824,865	20,285,907	42,474,777	22,976,419	46,573,713	22,382,348	78,893,855	**39,467,535**	43,121,326	26,283,444
Adipose_AH2	48,915,696	29,598,127	42,239,655	25,503,801	44,139,072	20,835,785	67,639,834	**34,455,960**	47,030,891	29,981,696
Brain_AH1	43,974,930	21,565,725	39,837,681	24,123,648	40,856,531	**10,785,248**	58,814,934	**26,689,432**	43,523,581	26,563,224
Brain_AH2	40,688,277	20,720,820	36,538,575	**13,786,304**	34,937,537	**14,697,550**	75,947,873	**27,707,092**	46,263,586	27,285,703
Heart_AH1	36,136,771	**18,483,810**	95,807,379	31,223,234	41,465,195	**11,695,832**	65,993,929	**14,039,244**	44,242,333	25,385,371
Heart_AH2	35,569,728	**15,570,052**	44,745,372	28,549,374	41,419,274	**14,340,384**	78,002,179	**23,580,422**	41,709,767	25,061,680
Lamina_AH1	37,994,974	22,518,048	42,526,642	21,314,248	41,437,394	20,524,291	59,375,211	**19,967,360**	43,576,023	26,395,838
Lamina_AH2	35,912,806	**19,272,229**	44,412,262	24,127,700	45,700,886	20,649,024	64,949,083	**23,166,407**	33,994,964	21,390,454
Liver_AH1	44,216,613	27,972,377	73,403,419	33,542,235	41,146,455	**14,528,367**	72,283,887	**24,509,202**	44,555,911	26,116,877
Liver_AH2	38,117,020	23,236,188	52,222,345	34,116,013	39,119,796	**14,805,249**	73,937,379	**30,853,685**	52,257,642	28,414,218
Lung_AH1	54,431,033	31,436,935	70,386,807	40,515,514	44,451,765	22,092,067	57,340,656	**24,920,879**	44,375,386	26,906,108
Lung_AH2	57,372,074	25,395,989	57,250,724	36,479,723	42,712,275	22,196,353	61,251,734	**30,077,163**	48,412,181	29,106,329
Muscle_AH1	44,360,817	**19,826,000**	37,596,912	**19,018,397**	40,544,487	**19,009,675**	82,238,330	**28,738,934**	40,839,362	23,118,540
Muscle_AH2	46,011,699	24,938,732	32,919,675	20,469,885	35,257,194	**7,056,497**	71,172,395	**28,617,004**	41,413,933	26,760,212
Ovary_AH1	35,313,302	21,954,835	41,748,385	26,186,014	40,900,341	21,501,928	67,738,747	**26,430,617**	38,701,828	24,444,447
Ovary_AH2	39,121,462	26,035,899	35,814,629	22,695,425	36,698,698	20,413,333	57,017,668	**23,487,440**	41,753,846	26,508,982

* The bolded cells represent samples that did not reach the usable read pair target of 45 M for H3K27me3 and 20 M for the remaining marks and input samples.

Supplementary Table 3.6. Percentage of the Genome Covered by Single-End, Normalized Peaks in Stallions and Mares

Genome Coverage (%) of Stallion SE Peaks				
Tissues	H3K27ac	H3K4me1	H3K4me3	H3K27me3-M
Adipose	1.24	1.22	1.24	1.55
Brain	2.31	0.58	1.03	0.78
Heart	2.55	1.17	1.13	1.29
Lamina	2.12	0.81	0.97	1.62
Liver	2.22	1.83	1.19	1.35
Lung	2.01	1.57	1.15	1.90
Muscle	1.63	0.56	0.97	1.40
Testis	1.70	0.17	1.43	1.21
Average	1.97	0.99	1.14	1.39

Genome Coverage (%) of Mare SE Peaks				
Tissues	H3K27ac	H3K4me1	H3K4me3	H3K27me3-M
Adipose	2.01	0.85	0.99	0.68
Brain	1.84	0.16	0.87	0.46
Heart	1.71	1.42	0.76	0.26
Lamina	1.60	0.92	0.92	0.34
Liver	2.13	1.70	0.84	0.46
Lung	1.94	1.17	0.98	0.50
Muscle	1.84	0.44	0.75	0.57
Ovary	1.79	0.80	0.99	0.57
Average	1.86	0.93	0.89	0.48

***** All peaks were called with MACS2

Supplementary Table 3.7. Peaks Unique to Mares and Stallions within Tissues and Differences in the Percentage of Unique Peaks and Total Peaks Between Sexes

H3K27me3					
Tissue	Sex	Combined Peaks	% Unique	Difference in Percent Unique	Difference in Total Peak (%)
Adipose	Mare	8,019	11.5	53.1	47.1
	Stallion	15,172	64.6		
Brain	Mare	5,195	23.5	34.3	36.1
	Stallion	8,127	57.8		
Heart	Mare	4,892	6.5	69.3	62.3
	Stallion	12,997	75.8		
Lamina	Mare	5,962	6.1	72.2	64.0
	Stallion	16,539	78.2		
Liver	Mare	7,254	8.1	58.4	50.4
	Stallion	14,638	66.5		
Lung	Mare	6,758	4.4	71.8	65.9
	Stallion	19,818	76.1		
Muscle	Mare	8,429	13.2	48.0	40.0
	Stallion	14,041	61.2		
Ovary	Mare	7,245	21.8	41.6	44.0
Testis	Stallion	12,932	63.4		
H3K27ac					
Tissue	Sex	Combined Peaks	% Unique	Difference in Percent Unique	Difference in Total Peak (%)
Adipose	Mare	69,314	58.4	31.7	32.2
	Stallion	46,994	26.8		
Brain	Mare	63,270	40.0	10.3	2.6
	Stallion	61,631	50.3		
Heart	Mare	58,909	19.6	32.5	27.3
	Stallion	81,127	52.2		
Lamina	Mare	65,521	28.6	17.8	12.3
	Stallion	74,724	46.4		
Liver	Mare	76,104	36.3	1.6	2.8
	Stallion	73,938	38.0		
Lung	Mare	66,259	38.4	4.0	6.2
	Stallion	70,646	42.4		
Muscle	Mare	63,693	42.6	9.4	7.4
	Stallion	59,008	33.3		
Ovary	Mare	62,078	52.8	2.6	3.9
Testis	Stallion	64,616	50.2		

H3K4me1					
Tissue	Sex	Combined Peaks	% Unique	Difference in Percent Unique	Difference in Total Peak (%)
Adipose	Mare	45,184	38.1	19.2	21.1
	Stallion	57,244	57.3		
Brain	Mare	14,417	50.7	32.2	57.4
	Stallion	33,856	82.9		
Heart	Mare	61,475	47.2	11.4	8.2
	Stallion	56,432	35.8		
Lamina	Mare	46,763	49.1	4.5	1.0
	Stallion	47,219	44.6		
Liver	Mare	77,898	36.2	1.5	3.7
	Stallion	75,012	37.7		
Lung	Mare	58,595	32.8	17.1	17.0
	Stallion	70,620	49.8		
Muscle	Mare	29,602	50.5	9.3	12.6
	Stallion	33,876	59.8		
Ovary	Mare	45,988	87.6	32.5	68.3
Testis	Stallion	14,583	55.1		
H3K4me3					
Tissue	Sex	Combined Peaks	% Unique	Difference in Percent Unique	Difference in Total Peak (%)
Adipose	Mare	20,925	9.5	27.3	29.3
	Stallion	29,614	36.8		
Brain	Mare	20,545	12.4	13.1	12.2
	Stallion	23,399	25.5		
Heart	Mare	19,114	2.7	31.2	24.1
	Stallion	25,174	33.9		
Lamina	Mare	22,677	14.5	1.4	0.0
	Stallion	21,902	15.9		
Liver	Mare	20,255	4.3	30.6	21.4
	Stallion	25,771	34.9		
Lung	Mare	21,958	9.9	15.8	12.7
	Stallion	25,149	25.7		
Muscle	Mare	20,427	7.8	18.7	7.0
	Stallion	21,964	26.5		
Ovary	Mare	22,540	21.0	33.7	41.2
Testis	Stallion	38,810	54.7		

* Difference in Total Peak (%) was calculated as the difference in combined peaks divided by the larger of the two combined peak numbers.

CHAPTER 4: WHOLE-GENOME SEQUENCING TO INVESTIGATE A POSSIBLE GENETIC BASIS OF PEROSOMUS ELUMBIS IN A CALF RESULTING FROM A CONSANGUINEOUS MATING

Barber, A**.,** Helms, A., Thompson, R., Whitlock, B., Steffen, D., & Petersen, J. (2021) Whole-genome sequencing to investigate a possible genetic basis of perosomus elumbis in a calf resulting from a consanguineous mating. *Translational Animal Science*, 5(Supplement_S1), S1–S5.

Introduction

Perosomus elumbis (PE) is a lethal, congenital defect marked by aplasia of the lumbar and sacral spine and spinal cord. Contracture of the hind limbs is also commonly observed in affected individuals. PE has been reported in many domestic species, with numerous case reports in Holstein cattle in the past two decades (Jones, 1999; Karakaya et al., 2013; Agerholm et al., 2014) The etiology of PE remains unknown. In one instance a stillborn Holstein calf with PE was found to be infected with Bovine Viral Diarrhea Virus (BVDV) (Karakaya et al., 2013), and thus it is possible PE may be due to genetic and/or environmental factors. Recently, a stillborn Angus calf was diagnosed with PE following an accidental mother-son mating (Helms et al., 2020). BVDV was not detected in the affected Angus calf, dam, nor sire. Due to the relationship between the sire and dam it was hypothesized that a novel, recessive genetic variant may be responsible for the development of PE in this Angus calf. The objective of this study was to use whole-genome sequencing to address this hypothesis and identify candidate variants for PE in this calf.

Materials and Methods

IACUC Statement

All procedures and protocols were performed following the University of Nebraska-Lincoln's Institutional Animal Care and Use Committee guidelines.

Sample Collection and DNA Isolation

Case presentation and diagnosis are reported in Helms et al. (2020). Tissue samples were collected from the affected calf following necropsy at the University of Tennessee Veterinary Medical Center. Blood samples were also taken from the dam, sire, and ten paternal half-siblings; tissue and blood were sent to the University of Nebraska-Lincoln. DNA was isolated from tissue and blood utilizing Qiagen Gentra Puregene Kits (Gentra Systems, Minneapolis, MN). Paternity was verified for all calves using the commercially available SeekSire™ parentage assay at Neogen GeneSeek (Lincoln, NE).

Whole Genome Sequencing and Variant Filtering

DNA collected from the affected calf, the dam, the sire, and three paternal half-siblings was sent to Admera Health (South Plainfield, NJ) for KAPA library prep and 150bp paired-end sequencing on an Illumina NovaSeq to a targeted sequencing depth of 12X. After trimming adapter sequences and poor quality bases (TrimGalore; Wu et al., 2011) sequence reads from the calf, dam, sire, and half siblings along with 27 other Angus and Angus-cross animals were mapped to the UOA_Angus_1 reference genome with BWA-MEM (Li, 2013).

Variants were called using Freebayes (Garrison and Marth, 2012) and annotated using SnpEff (Cingolani et al., 2012). SnpSift was also used to filter variants in which the affected calf was homozygous and both the dam and sire were heterozygous. With the assumption that PE is rare in Angus cattle, variants were further pruned using VCFtools (Danecek et al., 2011) to select only variants in which the alternative allele count was between four and seven to account for a homozygous calf, two heterozygous parents, and allow for the half-siblings to be heterozygous. Variants were further reduced to include only those predicted to have a moderate to high impact. Variants fitting the criteria were further investigated. Variants were remapped to the ARS-UCD1.2 reference genome using NCBI's Remap tool to determine if the variants had been previously reported.

PCR and Sanger Sequencing

Primers for regions of interest were developed using sequence from the UOA_Angus_1 reference genome. Oligonucleotides were designed using IDT's PrimerQuest Tool. PCR products were amplified using an annealing temperature ranging between 54-58 C and visualized on 1.2% agarose gels. PCR products were sent to ACGT Inc. (Wheeling, IL) for Sanger sequencing. Sequence results were visualized using Gene Code Corporation's Sequencher.

Sequence Read Archive Search

A search of NCBI's Sequence Read Archive (SRA) was conducted using a variant Search pipeline (https://github.com/SichongP/SRA_variant_search); NCBI's Remap function was used to identify coordinates across genome assemblies. The frameshift

variant was not assessed in the SRA search due to difficulty interpreting indels using this method.

Results

Candidate Variant Filtering

Variant calling across the 31 Angus and Angus cross individuals, including the affected calf, the dam, the sire, and three half siblings, identified 21,223,927 variants across the genome. Using SnpSift to filter for variants in which the calf was homozygous for the variant and the dam and sire were heterozygous yielded 506,813 variants. Removing variants at high frequency in the data set reduced candidate variants to 14,011.

Filtering by predicted impact as annotated in SnpEff resulted in 77 variants with a predicted moderate impact and 5 with predicted high impact. Predicted high impact variants were excluded from further analysis if they were previously annotated and a carrier was found in the original 39 screened animals or if the variant was found in the homozygous state in any individual(s) other than the affected calf. After removing variants fitting those criteria, the final candidate variant list consisted of 18 missense variants and one frameshift resulting from a one base pair deletion. Three of the 19 candidate variants were not previously annotated on Ensembl (Table 4.1). The frameshift variant was in exon 4 of protein tyrosine kinase 7 (*PTK7)* and is predicted to result in a premature stop codon prior to the end exon. Due to its putative deleterious impact on gene function, this variant was further studied as a candidate causal variant.

Sanger Sequencing Verification of Frameshift Mutation in PTK7 and SRA results

Sanger sequencing confirmed the presence of a homozygous, one base pair deletion in the affected calf. Additionally, six of 10 half-siblings were heterozygous for the deletion (Figure 4.1).

The search of the Sequence Read Archive (SRA) resulted in genotypes of 883 additional cattle including 96 Angus and Angus cross. Through this analysis, individuals homozygous for variants were identified at 15 of the 18 missense loci; the indel in *PTK7* was not able to be queried.

The three remaining missense variants in *KDM1A, C2H2orf66,* and *ZSCAN26,* and one frameshift variant in *PTK7* remained as candidate causal variants (Table 4.1). From the SRA data, 1 Holstein was heterozygous for the *KDM1A* variant; 2 Tyrolean Grey cattle, 1 Chianina, and 1 Romagnola were heterozygous for the *C2H2orf66* variant; and 2 Angus, 1 Chi-Angus cross, and 1 Holstein were heterozygous for the *ZSCAN26* variant.

Discussion

In this study, missense mutations in *KDM1A, C2H2orf66,* and *ZSCAN26,* as well as a frameshift mutation in *PTK7* could not be ruled out as causative of PE in this Angus calf. PE is a lethal congenital defect that results in aplasia of the lumbar spine and frequent contracture of the hind limbs. Although relatively rare in Angus cattle, numerous cases of PE have been reported in Holstein cattle. The cause of PE has yet to be determined with both environment and genetics suspected to play a role. In this case, the

affected Angus calf was the result of a consanguineous mating suggesting that a recessive mutation may be the cause.

Of the four variants remaining after filtering out those that did not fit the hypothesized mode of inheritance, and those at high frequency in other cattle, the missense mutation in *KDM1A* and the frameshift mutation in *PTK7* are strong functional candidates due to their roles in early development. *KDM1A* is involved in epigenetic regulation of embryonic gene expression (Ancelin et al., 2016), while *PTK7* functions in the planar cell polarity (PCP) pathway that regulates cell movement and migration (Berger et al., 2017).

KDM1A is an histone 3 lysine 4 (H3K4) lysine demethylase that functions to remove enhancer marks from histones. These epigenetic marks influence early development in part by regulating the spaciotemporal activation of genes which orchestrates proper embryonic development (Ancelin et al., 2016). Dysregulation of *KDM1A* can result in developmental arrest and altered patterns of gene expression in the developing embryos (Ancelin et al., 2016).

PTK7, a member of the tyrosine kinase family, plays a role in the planar cell polarity (PCP) pathway. This pathway establishes polarity in cells and regulates cell movement and migration in embryonic development (Berger et al., 2017). This gene is of particular interest as it has been implicated in congenital scoliosis in zebrafish (Hayes et al., 2014) demonstrating a clear role in the development of the fetal spine. Additionally, another gene with a paralog in this pathway, *VANLG1*, has been implicated in an analogous human disorder called caudal regression syndrome (CRS) (Kibar et al., 2007; Porsch et al., 2016). Furthermore, *VANGL2*, which directly interacts with *PTK7* in the PCP

pathway has also been implicated in neural tube defects (Kibar et al, 2011). These studies demonstrate a clear role of *PTK7* and the PCP pathway in spinal development making a frameshift mutation in *PTK7* a strong functional candidate for PE in cattle.

Although *PTK7* provides a strong functional candidate for PE, this study is limited due to the availability of a single affected calf. This study should be supplemented with additional affected calves as cases are reported. Furthermore, as new sequence reads become available in the SRA database, additional animals can be screened for the associated variants found in this study. Due to the rarity of this condition, this study could be extended to consider affected calves from other breeds.

Implications

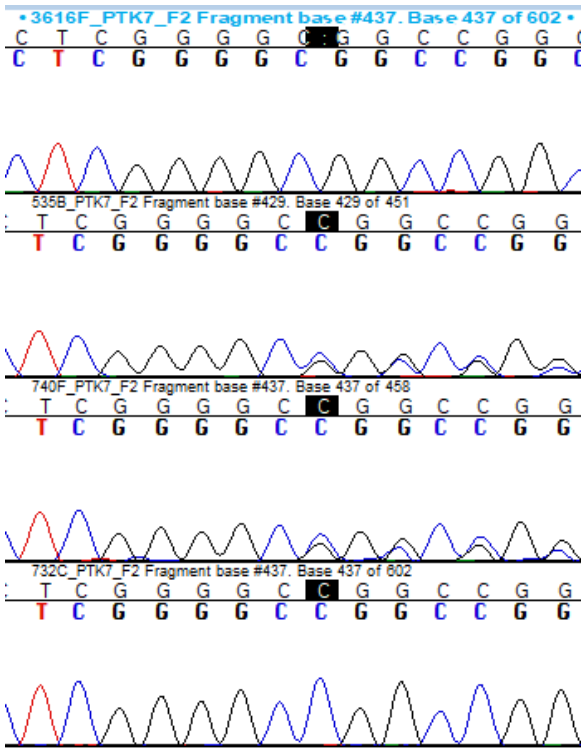
The accumulation of lethal recessive variation within breeds negatively impacts production and breed health. With the growing use of artificial insemination (AI), prolific carrier bulls can rapidly increase the allele frequency of recessive disorders within the breed. Using whole-genome sequencing, disease-associated and disease-causing variation can be identified. Although a causative variant was not validated in this study, in the case that would occur, genetic testing could allow for informed matings to eliminate the production of affected individuals.

Table 4.1. Candidate variants for perosomus elumbis. Italicized rows indicate that no individuals were homozygous for the variant in the Sequence Reads Archive (SRA) search. Bolded rows indicate variants with a predicted high impact on gene function from SnpEff (Cingolani et al., 2012). Positions labelled UOAcorrespon d to the UOA_Angus_1 reference genome, and positions labelled ARS correspond to the ARS_UCD1.2 reference genome. Previously annotated variants are noted under Variant ID. Type represents the predicted position/outcome observed in the UOA_Angus_1 reference genome (top) and the ARS_UCD1.2 reference genome (bottom).

Chr	Position	Gene	Reference	Variant	Type	Variant ID
2	*UOA: 6567555*	*KDM1A*	*C*	*T*	*Missense*	*Novel*
	ARS: 129835952				*Missense*	*Novel*
2	*UOA: 50768895*	*C2H2orf66*	*T*	*C*	*Missense*	-
	ARS: 85400473				*Intergenic*	*rs719944515*
4	UOA: 6313462	ASIC3	C	T	Missense	-
	ARS: 113625394				Missense	rs466455595
15	UOA: 63709344	QSER1	A	G	Missense	-
	ARS: 63626227				Intronic	rs380723979
15	UOA: 63709425	QSER1	CA	GG	Missense	-
	ARS: 63626308				Intronic	rs799405617
17	UOA: 51162578	NCOR2	C	T	Missense	-
	ARS: 51449160				Missense	rs472931263
17	UOA: 51555593	DNAH10	T	C	Missense	-
	ARS: 51850428				Missense	rs136088999

22	UOA: 59081586	EFCC1	G	T	Missense	Novel
	ARS: 58980413				Missense	Novel
23	UOA: 22015758	ZNF165	G	A	Missense	-
	ARS: 30587728				Missense	rs526649482
23	UOA: 22140420	ZSCAN9	C	T	Missense	-
	ARS: 30463098				Downstream	rs463835998
23	*UOA: 22181166*	*ZSCAN26*	*C*	*A*	*Missense*	-
	ARS: 30422277				*Missense*	*rs521986257*
23	UOA: 22192722	PGBD1	G	A	Missense	-
	ARS: 30410722				Missense	rs432139616
23	UOA: 22192789	PGBD1	C	T	Missense	-
	ARS: 30410655				Missense	rs449832006
23	UOA: 23246829	OR1O9	C	T	Missense	-
	ARS: 29295898				CNV	nsv835503
23	UOA: 23452894	OR2H1D	A	C	Missense	-
	ARS: 29099147				Downstream/CNV	rs800181923
23	**UOA: 35361113**	**PTK7**	**CG**	**G**	**Frameshift**	**Novel**
	ARS: 16744942				**Frameshift**	**Novel**
28	UOA: 25626059	TSPAN15	A	G	Missense	-
	ARS: 25846885				Missense	rs469369204
28	UOA: 26696544	ADAMTS14	C	T	Missense	-
	ARS: 26918561				Missense	rs135381293
28	UOA: 30658181	DUSP13	C	T	Missense	-
	ARS: 30876012				Missense	rs379594626

A)



B)

Individual	Genotype
Affected Calf	-/-
Dam	C/-
Sire	C/-
Half Sibling 1	C/-
Half Sibling 2	C/C
Half Sibling 3	C/-
Half Sibling 4	C/-
Half Sibling 5	C/-
Half Sibling 6	C/-
Half Sibling 7	C/C
Half Sibling 8	C/C
Half Sibling 9	C/C
Half Sibling 10	C/-

Figure 4.1. Sanger sequencing confirms the presence of a one base pair deletion in *PTK7*. A) Sequence data of exon 4 of *PTK7* from the affected calf, the dam, and two half siblings depicts the presence of a deletion for which the affected calf was homozygous, the dam and half-sibling heterozygous, and second half-sibling wildtype. **B)** Genotypes of the affected calf, the dam, the sire, and ten half-siblings at the candidate locus in *PTK7*. A dash (-) indicates the 1bp deletion.

REFERENCES

Agerholm, J.S., Holm, W., Schmidt, M. *et al.* Perosomus elumbis in Danish Holstein cattle. *BMC Vet Res* **10,** 227 (2014). https://doi.org/10.1186/s12917-014-0227-2

Allfrey, V.G., Faulkner, R., & Mirksy, A.E. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 51(5), 786-794. https://doi.org/10.1073/pnas.51.5.786

American Horse Council (2018) *Economic impact of the U.S. horse industry.* https://www.horsecouncil.org/resources/economics/

Ancelin, K., Syx, L., Borensztein, M., Ranisavljevic, N., Vassilev, I., Briseño-Roa, L., Liu, T., Metzger, E., Servant, N., Barillot, E., Chen, C. J., Schüle, R., & Heard, E. (2016). Maternal LSD1/KDM1A is an essential regulator of chromatin and transcription landscapes during zygotic genome activation. *eLife*, *5*, e08851. https://doi.org/10.7554/eLife.08851

Andrews, Simon. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E., & Moudrianakis, E. N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proceedings of the National Academy of Sciences of the United States of America*, 88(22), 10148–10152. https://doi.org/10.1073/pnas.88.22.10148

Arya, G., & Schlick, T. (2009). A tale of tails: how histone tails mediate chromatin compaction in different salt and linker histone environments. *The Journal of Physical Chemistry A*, 113(16), 4045–4059. https://doi.org/10.1021/jp810375d

Bae, S. & Lesch, B.J. (2020) H3K4me1 distribution predicts transcription state and poising at promoters. Frontiers Cell and Developmental Biology, 8, 289. https://doi.org/10.3389/fcell.2020.00289

Bannister, A. & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21, 381-395. https://doi.org/10.1038/cr.2011.22

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bauer, W.R., Hayes, J.J., White, J.H., & Wolffe, A.P. (1994). Nucleosome structural changes due to acetylation. *Journal of Molecular Biology*, 236(3), 685-690. https://doi.org/10.1006/jmbi.1994.1180

Bauer, A., Hiemesch, T., Jagannathan, V., Neuditschko, M., Bachmann, I., Rieder, S., Mikko, S., Penedo, M.C., Tarasova, N., Vitková, M., Sirtori, N., Roccabianca, P., Leeb, T., & Welle, M.M. (2017). A nonsense variant in the *ST14* gene in Akhal-Teke horses with Naked Foal Syndrome. *G3 Genes|Genomes|Genetics*, 7(4), 1315-1321. https://doi.org/10.1534/g3.117.039511

Berger, H., Wodarz, A., and Borchers, A. (2017). PTK7 faces the Wnt in development and disease. *Front. Cell. Dev. Biol.* https://doi.org/10.3389/fcell.2017.00031

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., Jaenisch, R., Wagschal, A., Feil, R., Schreiber, S.L., & Lander, E.S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2), 315-326. https://doi.org/10.1016/j.cell.2006.02.041

Bierne, H., Tham, T. N., Batsche, E., Dumay, A., Leguillou, M., Kernéis-Golsteyn, S., Regnault, B., Seeler, J. S., Muchardt, C., Feunteun, J., & Cossart, P. (2009). Human BAHD1 promotes heterochromatic gene silencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(33), 13826–13831. https://doi.org/10.1073/pnas.0901259106

Binns, M.M, Boehler, D.A., & Lamber, D.H. (2010). Identification of the myostatin locus (*MSTN*) as having a major effect on optimum racing distance in the Thoroughbred horse in the USA. *Animal Genetics*, 41(s2), 154-158. https://doi.org/10.1111/j.1365-2052.2010.02126.x

Bogliotti, Y. S., & Ross, P. J. (2012). Mechanisms of histone H3 lysine 27 trimethylation remodeling during early mammalian development. Epigenetics, 7(9), 976–981. https://doi.org/10.4161/epi.21615

Bordoli, L., Hüsser, S., Lüthi, U., Netsch, M., Osmani, H., & Eckner, R. (2001). Functional analysis of the p300 acetyltransferase domain: The PHD finger of p300 but not of CBP is dispensable for enzymatic activity. *Nucleic Acids Research*, 29(21), 4462-4471. https://doi.org/10.1093/nar/29.21.4462

Boyer, LA., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., Bell, G.W., Otte, A.P., Vidal, M., Gifford,

D.K., Young, R.A., & Jaenisch, R. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. https://doi.org/10.1038/nature04733

Brooks, S.A., Gabreski, N., Miller, D., Brisbin, A., Brown, H.E., Streeter, C., Mezey, J., Cook, D., & Antczak, D.F. (2010). Whole-genome SNP association in the horse: Identification of a deletion in Myosin Va responsible for Lavender Foal Syndrome. *PLoS Genetics*, 6(4), e1000909. https://doi.org/10.1371/journal.pgen.1000909

Burns, E.N., Bordbari, M.H., Mienaltowski, M.J., Affolter, V.K., Barro, M.V., Gianino, F., Gianino, G., Giulotto, E., Kalbfleisch, T.S., Katzman, S.A., Lassaline, M., Leeb, T., Mack, M., Müller, E.J., MacLeod, J.N., Ming-Whitfield, B., Alanis, C.R., Raudsepp, T., Scott, E., … Finno, C.J. (2018). Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Animal Genetics*, 49(6), 564-570. https://doi.org/10.1111/age.12717

Capomaccio, S., Vitulo, N., Verini-Supplizi, A., Barcaccia, G., Albiero, A., D'Angelo, M., Campagna, D., Valle, G., Felicetti, Silverstrelli, M., & Cappelli, K. (2013). RNA sequencing of the exercise transcriptome in equine athletes. *PLos ONE,* 8(12), e83504. https://doi.org/10.1371/journal.pone.0083504

Carninci, P., Kvam, C., Kitamura, A., Ohsumi, T., Okazaki, Y., Itoh, M., Kamiya, M., Shibata, K., Sasaki, N., Izawa, M., Muramatsu, M., Hayashizaki, Y., & Schneider, C. (1996). High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, 37(3), 327-336. https://doi.org/10.1006/geno.1996.0567

Chang, T.C., Klabnik, J.L., & Liu, W.S. (2011). Regional selection acting on the *OFD1* gene family. *PLos ONE,* 6(10), e26195. https://doi.org/10.1371/journal.pone.0026195

Chassier, M., Barrey, E., Robert, C., Duluard, A., Danvy, S., & Ricard, A. (2018). Genotype imputation accuracy in multiple equine breeds from medium- to high-density genotypes. *Journal of Animal Breeding and Genetics*, 135(6), 420-431. https://doi.org/10.1111/jbg.12358

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D.K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D.S., & Gingeras, T.R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science,* 308(5725), 1149-1154. https://doi.org/10.1126/science.1108625

Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S., & Dynlacht, B.D. (2014). A role of H3K4 monomethylation in gene repression and partitioning of

chromatin readers. *Molecular Cell,* 53(6), 979-992. https://doi.org/10.1016/j.molcel.2014.02.032

Cingolani, P., Platts, A., Wang, l., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6(2), 80–92. https://doi.org/10.4161/fly.19695

Christensen, J., Agger, K., Cloos, P.A.C., Pasini, D., Rose, S., Sennels, L., RAppsilber, J., Hansen, K.H., Salcini, A.E., & Helin, K. (2007). RBP2 belongs to a family of demethylases, specific for tri- and demethylated lysine 4 on histone 3. *Cell*, 128(6), 1063-1076. https://doi.org/10.1016/j.cell.2007.02.003

Coleman, S.J., Zeng, Z., Wang, K., Luo, S., Khrebtukova, I., Mienaltowski, M.J., Liu, J., & MacLeod, J.N. (2010). Structural annotation of equine protein-coding genes determined by mRNA sequencing. *Animal Genetics*, 41(s2), 121-130. https://doi.org/10.1111/j.1365-2052.2010.02118.x

Collins, F. S., & Fink, L. (1995). The Human Genome Project. *Alcohol Health and Research World,* 19(3), 190–195.

Corbin, L.J., Blott, S.C., Swinburne, J.E., Sibbons, C., Fox-Clipsham, L.Y., Helwegen, M., Parkin, T.D.H., Newton, J.R., Bramlage, L.R., McIlwraith, C.W., Bishop, S.C., Woolliams, J.A., & Vaudin M. (2012). A genome-wide association study of osteochondrosis dissecans in the Thoroughbred. *Mammalian Genome,* 23, 294-303. https://doi.org/10.1007/s00335-011-9363-1

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS*, 107(50), 21931-21936. www.pnas.org/cgi/doi/10.1073/pnas.1016071107

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R., Lunter, G., Marth, G., Sherry, S.T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Das, P.J., McCarthy, F., Vishnoi, M., Paria, N., Gresham, C., Li, G., Kachroo, P., Sudderth, A.K., Teague, S., Love, C.C., Varner, D.D., Chowdhary, B.P., & Raudsepp, T. (2013). Stallion sperm transcriptome comprises functionally coherent coding and

regulatory RNAs as revealed by microarray analysis and RNA-seq. *PLoS ONE*, 8(2), e56535. https://doi.org/10.1371/journal.pone.0056535

Davenport, K.M., Massa, A.T., Bhattarai, S., McKay, S.D., Mousel, M.R., Herndon, M.K., White, S.N., Cockett, N.E., Smith, T.P.L., Murdoch, B.M., & The Ovine FAANG Project Consortium (2021). Characterizing genetic regulatory elements in ovine tissues. https://doi.org/10.3389/fgene.2021.628849

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

De Santa, F., Totaro, M.G., Prosperini, E., Notarbartolo, S., Testa, C., & Natoli, G. (2007). The histone H3 lysine-27 demethylase JMJD3 links inflammation to inhibition of polycomb-mediated gene silencing. *Cell,* 130(6), 1083-1094. https://doi.org/10.1016/j.cell.2007.08.019

Dierks, C., Löhring, K., Lampe, V., Wittwer, C., Drögemüller, C., & Distl, O. (2007). Genome-wide search for markers associated with osteochondrosis in Hanoverian warmblood horses. *Mammalian Genome,* 18(10), 739–747. https://doi.org/10.1007/s00335-007-9058-9

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 489(7414), 101–108. https://doi.org/10.1038/nature11233

Doblin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T.R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Donnelly, C. G., Bellone, R. R., Hales, E. N., Nguyen, A., Katzman, S. A., Dujovne, G. A., Knickelbein, K. E., Avila, F., Kalbfleisch, T. S., Giulotto, E., Kingsley, N. B., Tanaka, J., Esdaile, E., Peng, S., Dahlgren, A., Fuller, A., Mienaltowski, M. J., Raudsepp, T., Affolter, V. K., Petersen, J. L., … Finno, C. J. (2021). Generation of a biobank from two adult Thoroughbred stallions for the Functional Annotation of Animal Genomes Initiative. *Frontiers in Genetics*, 12, 650305. https://doi.org/10.3389/fgene.2021.650305

Dong, X., Greven, M. C., Kundaje, A., Djebali, S., Brown, J. B., Cheng, C., Gingeras, T. R., Gerstein, M., Guigó, R., Birney, E., & Weng, Z. (2012). Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology*, 13(9), R53. https://doi.org/10.1186/gb-2012-13-9-r53

Donnelly, C. G., Bellone, R. R., Hales, E. N., Nguyen, A., Katzman, S. A., Dujovne, G. A., Knickelbein, K. E., Avila, F., Kalbfleisch, T. S., Giulotto, E., Kingsley, N. B., Tanaka, J., Esdaile, E., Peng, S., Dahlgren, A., Fuller, A., Mienaltowski, M. J., Raudsepp, T., Affolter, V. K., Petersen, J. L., … Finno, C. J. (2021). Generation of a biobank from two adult Thoroughbred stallions for the Functional Annotation of Animal Genomes Initiative. *Frontiers in Genetics*, 12, 650305. https://doi.org/10.3389/fgene.2021.650305

Drögemüller, M., Jagannathan, V., Welle, M.M., Graubner, C., Straub, R., Gerber, V., Burger, D., Signer-Hasler, H., Poncet, P.A., Klopfenstein, S., von Niederhäusern, R., Tetens, J., Thaller, G., Rieder, S., Drögemüller, C., & Leeb, T. (2014). Congenital hepatic fibrosis in the Franches-Montagnes horse is associated with the polycystic kidney and hepatic disease 1 (*PKHD1*) gene. *PLoS One*, 9(10), e110125. https://doi.org/10.1371/journal.pone.0110125

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047-3048. https://doi.org/10.1093/bioinformatics/btw354

Ertelt, A., Barton, A. K., Schmitz, R. R., & Gehlen, H. (2014). Metabolic syndrome: is equine disease comparable to what we know in humans? *Endocrine Connections*, 3(3), R81–R93. https://doi.org/10.1530/EC-14-0038

Fagerberg, L., Hallström, B. M., Oksvold, P., Kampf, C., Djureinovic, D., Odeberg, J., Habuka, M., Tahmasebpoor, S., Danielsson, A., Edlund, K., Asplund, A., Sjöstedt, E., Lundberg, E., Szigyarto, C. A., Skogs, M., Takanen, J. O., Berling, H., Tegel, H., Mulder, J., Nilsson, P., … Uhlén, M. (2014). Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & Cellular Proteomics : MCP*, 13(2), 397–406. https://doi.org/10.1074/mcp.M113.035600

Fang, L., Liu, S., Liu, M., Kang, X., Lin, S., Li, B., Connor, E.E., Baldwin VI, R.L., Tenesa, A., Ma., L., Liu, G.E., & Li, C.J. (2019). Functional annotation of the cattle genome through the systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biology*, 17, 68. https://doi.org/10.1186/s12915-019-0687-8

Finnin, M.S., Donigian, J.R., Cohen, A., Richon, V.M., Rifkind, R.A., Marks, P.A., Breslow, R., & Pavletich, N.P. (1999). Structures of histone deacetylase homologue

bound to the TSA and SAHA inhibitors. *Nature,* 401, 188-193. https://doi.org/10.1038/43710

Flanagan, J.F., Mi, L.Z., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor, W., Rastinejad, F., & Khorasanizadeh, S. (2005). Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature,* 438, 1181-1185. https://doi.org/10.1038/nature04290

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, 49(D1), D916–D923. https://doi.org/10.1093/nar/gkaa1087

Frischknecht, M., Jagannathan, V., Plattet, P., Neuditschko, M., Singer-Hasler, H., Bachmann, I., Pacholewska, A., Drögemüller, C., Dietschi, E., Flury, C., Rieder, S., & Leeb, T. (2015). A non-synonymous *HMGA2* variant decreases height in Shetland ponies and other small horses. *PLoS One,* 10(10), e0140749. https://doi.org/10.1371/journal.pone.0140749

Garcia-Ramirez, M., Rocchini, C., & Ausio, J. 1995. Modulation of chromatin folding by histone acetylation. *Journal of Biological Chemistry,* 270, 17923–17928. https://doi.org/10.1074/jbc.270.30.17923

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [qbio.GN] https://doi.org/10.48550/arXiv.1207.3907

Gershoni, M. & Pietrokovski, S. (2017). The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biology*, 15, 7. https://doi.org/10.1186/s12915-017-0352-z

Gerstel, B., Tuite, M. F., & McCarthy, J. E. (1992). The effects of 5'-capping, 3'-polyadenylation and leader composition upon the translation and stability of mRNA in a cell-free extract derived from the yeast Saccharomyces cerevisiae. *Molecular Microbiology*, 6(16), 2339–2348. https://doi.org/10.1111/j.1365-2958.1992.tb01409.x

Gordon, F., Luger, K., & Hansen, J. C. (2005). The core histone N-terminal tail domains function independently and additively during salt-dependent oligomerization of nucleosomal arrays. *The Journal of Biological Chemistry*, 280(40), 33701–33706. https://doi.org/10.1074/jbc.M507048200

Gottschalk, M., Metsger, J., Martinsson, G., Sieme, H., & Distl, O. (2016). Genome-wide association study for semen quality traits in German warmblood stallions. *Animal Reproductive Science,* 171, 81-86. http://dx.doi.org/10.1016/j.anireprosci.2016.06.002

Grant P. A. (2001). A tale of histone modifications. *Genome Biology*, 2(4), reviews0003.1-reviews0003.6. https://doi.org/10.1186/gb-2001-2-4-reviews0003

Grau, D. J., Chapman, B. A., Garlick, J. D., Borowsky, M., Francis, N. J., & Kingston, R. E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. *Genes & Development*, 25(20), 2210–2221. https://doi.org/10.1101/gad.17288211

Gujral, P., Mahajan, V., Lissaman, A.C., & Ponnampalam, A.P. (2020). Histone acetylation and the role of histone deacetylases in normal cyclic endometrium. *Reproductive Biology and Endocrinology*, 18, 84. https://doi.org/10.1186/s12958-020-00637-5

Gupta, R. M., Hadaya, J., Trehan, A., Zekavat, S. M., Roselli, C., Klarin, D., Emdin, C. A., Hilvering, C. R., Bianchi, V., Mueller, C., Khera, A. V., Ryan, R. J., Engreitz, J. M., Issner, R., Shoresh, N., Epstein, C. B., de Laat, W., Brown, J. D., Schnabel, R. B., . . . Kathiresan, S. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell*, *170*(3), 522–533.e15. https://doi.org/10.1016/j.cell.2017.06.049

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., & Lander, E.S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458, 223-227. https://doi.org/10.1038/nature07672

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Hayes, M., Gao, X., Yu, L.X., Paria, N., Henkelman, R.M., Wise, C.A., and Ciruna, B. (2014). *PTK7* mutant zebrafish models of congenital and idiopathic scoliosis implicate dysregulated Wnt signaling in disease. *Nature Communications*, 5, 4777. https://doi.org/10.1038/ncomms5777

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Calcar, S.V., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., & Ren, B. (2007). Distinct and predictive signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics,* 39, 311-318. https://doi.org/10.1038/ng1966

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., Antoziewicz-Bourget, J.E., Liu, H., Zhang, X., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459, 108-112. https://doi.org/10.1038/nature07829

Helms, A.B., Thompson, R.E., Lawton, S., Petersen, J.L., Watson, A., Sula, M.J., Steffen, D., and Whitlock, B.K. (2020) Uterine torsion dystocia complicated by Perosomus Elumbis in an Angus calf associated with a consanguineous mating. *Case Reports in Veterinary Medicine,* 20. https://doi.org/10.1155/2020/6543037

Hestand, M.S., Kalbfleisch, T.S., Coleman, S.J., Zeng, Z., Liu, J., Orlando, L., & MacLeod, J.N. (2015). Annotation of the protein coding regions of the equine genome. *PLoS One*, 10(6), e0124375. https://doi.org/10.1371/journal.pone.0124375

Hewitt, R.J. & Lloyd, C.M. (2021). Regulation of immune responses by the airway epithelial cell landscape. *Nature Reviews Immunology*, 21, 347-362. https://doi.org/10.1038/s41577-020-00477-9

Hill, E.W., Gu, J., Eivers, S.S., Fonseca, R.G., McGivney, B.A., Govindarajan, P., Orr, N., Katz, L.M., & MacHugh, D. (2010). A sequencing polymorphism in *MSTN* predicts sprinting ability and racing stamina in Thoroughbred horses. *PLoS ONE,* 5(1), e8645. https://doi.org/10.1371/journal.pone.0008645

Hisey, E. A., Hermans, H., Lounsberry, Z. T., Avila, F., Grahn, R. A., Knickelbein, K. E., Duward-Akhurst, S. A., McCue, M. E., Kalbfleisch, T., Lassaline, M. E., Back, W., & Bellone, R. R. (2020). Whole genome sequencing identified a 16 kilobase deletion on ECA13 associated with distichiasis in Friesian horses. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-07265-8

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., & Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS,* 106(23), 9362-9367. https://doi.org/10.1073/pnas.0903103106

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., Young, R.A., & Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. *Cell*, 155(7), 934-947. https://doi.org/10.1016/j.cell.2013.09.053

Ho, J. W., Bishop, E., Karchenko, P. V., Nègre, N., White, K. P., & Park, P. J. (2011). ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC Genomics*, 12, 134. https://doi.org/10.1186/1471-2164-12-134

Hood, L. & Rowen, L. (2013). The Human Genome Project: Big science transforms biology and medicine. *Genome Medicine*, 5, 79. https://doi.org/10.1186/gm483

Hosogane, M., Funayama, R., Shirota, M., & Nakayama, K. (2016). Lack of Transcription Triggers H3K27me3 Accumulation in the Gene Body. *Cell Reports*, 16(3), 696–706. https://doi.org/10.1016/j.celrep.2016.06.034

Hu, Y., Kireev, I., Plutz, M., Ashourian, N., & Belmont, A. S. (2009). Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template. *The Journal of Cell* Biology, 185(1), 87–100. https://doi.org/10.1083/jcb.200809196

Huang, d., Sherman, B. T., & Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Huang, d., Sherman, B. T., & Lempicki, R. A. (2009b). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1), 1–13. https://doi.org/10.1093/nar/gkn923

Husmann, D. & Gozani, O. (2019). Histone lysine methyltransferases in biology and disease. *Nature Structural & Molecular Biology*, 26, 880-889. https://doi.org/10.1038/s41594-019-0298-7

Inoue, A., & Fujimoto, D. (1969). Enzymatic deacetylation of histone. *Biochemical and Biophysical Research Communications*, 36(1), 146–150. https://doi.org/10.1016/0006-291x(69)90661-5

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921. https://doi.org/10.1038/35057062

Imai, S., Armstrong, C. M., Kaeberlein, M., & Guarente, L. (2000). Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature*, 403(6771), 795–800. https://doi.org/10.1038/35001622

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431, 931-945. https://doi.org/10.1038/nature03001

Iqbal, K., Chitwood, J.L., Meyers-Brown, G.A., Roser, J.F., & Ross, P.J. (2014). RNA-seq transcriptome profiling of equine inner cell mass and trophectoderm. *Biology of Reproduction*, 90(3), 1-9. https://doi.org/10.1095/biolreprod.113.113928

Isono, K., Endo, T. A., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B. E., & Koseki, H. (2013). SAM domain polymerization links subnuclear clustering of PRC1 to gene silencing. *Developmental Cell*, 26(6), 565–577. https://doi.org/10.1016/j.devcel.2013.08.016

Jones, C. J. (1999) Perosomus Elumbis (vertebral agenesis and arthrogryposis) in a stillborn Holstein calf. *Veterinary Pathology,* 36(1), 64-70.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orland, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., Taatjes, D.J., Dekker, J., & Young, R.A. (2010). Mediator and cohesion connect gene expression and chromatin architecture. *Nature*, 467, 430-435. https://doi.org/10.1038/nature09380

Kalbfleisch, T.S., Rice, E.S., DePrist, M.S., Walenz, B.P., Hestand, M.S., Vermeesch, J.R., O'Connell, B.L., Fiddes, I.T., Vershinina, A.O., Saremi, N.F., Petersen, J.L., Finno, C.J., Bellone, R.R., McCue, M.E., Brooks, S.A., Bailey, E., Orlando, L., Green, R.E., Miller, D.C., … MacLeod, J.N. (2018). *Communications Biology*, 1, 197. https://doi.org/10.1038/s42003-018-0199-z

Karakaya, E., Alpay, G., Yilmazbas-Mecitoglu, G., Alasonyalilar-Demirer, A., Akgül, B., Inan-Ozturkoglu, S., Ozyigit, M. O., Seyrek-Intas, D., Seyrek-Intas, K., Yesilbag, K., Gumen, A., & Keskin, A. (2013). Perosomus elumbis in a Holstein calf infected with bovine viral diarrhea virus. *Tierarztliche Praxis. Ausgabe G, Grosstiere/Nutztiere*, *41*(6), 387–391.

Karlić, R., Chung, H. R., Lasserre, J., Vlahovicek, K., & Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7), 2926–2931. https://doi.org/10.1073/pnas.0909344107

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17), 2204–2207. https://doi.org/10.1093/bioinformatics/btq351

Kern, C., Wang, Y., Xu, X., Pan, Z., Halstead, M., Chanthaviaxy, G., Saelao, P., Waters, S., Xiang, R., Chamberlain, A., Korf, I., Delany, M.E., Cheng, H.H., Medrano, J.F., Van Eenennaam, A.L., Tuggle, C.K., Ernst, C., Flicek, P., Quon, G., ... Zhou, H. (2021). Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nature Communications*, 12, 1821. https://doi.org/10.1038/s41467-021-22100-8

Kfoury, N., Qi, Z., Prager, B.C., & Rubin, J.B. (2021). Brd4-bound enhancers drive cell-intrinsic sex differences in glioblastoma. *PNAS*, 118(16), e2017148118. https://doi.org/10.1073/pnas.2017148118

Kibar, Z., Torban, E., McDearmid, J.R., Reynolds, A., Berghout, J., Mathieu, M., Kirillova, I., De Marco, P., Merello, E., Hayes, J.M., Wallingford, J.B., & Drapeau, P. (2007). Mutations in *VANLG1* associated with neural-tube defects. *New England Journal of Medicine,* 356, 1432-1437. https://doi.org/10.1056/NEJMoa060651

Kibar, Z., Salem, S., Bosoi, C.M., Pauwels, E., De Marco, P., Merello, E., Bassuk, A.G., Capra, V., and Gros, P. (2011). Contribution of *VANGL2* mutations to isolated nerual tube defects. *Clinical Genetics,*, 80(1), 76-82. https://doi.org/10.1111/j.1399-0004.2010.01515.x

Kingsley, N.B., Kern, C., Creppe, C., Hales, E.N., Zhou, H., Kalbfleisch, T.S., MacLeod, J.N., Petersen, J.L., Finno, C.J., Bellone, R.R. (2020). Functionally annotating regulatory elements in the equine genome using histone mark ChIP-seq. *Genes*, 11(1), 3. https://doi.org/10.3390/genes11010003

Kingsley, N. B., Hamilton, N. A., Lindgren, G., Orlando, L., Bailey, E., Brooks, S., McCue, M., Kalbfleisch, T. S., MacLeod, J. N., Petersen, J. L., Finno, C. J., & Bellone, R. R. (2021). "Adopt-a-Tissue" Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.649959

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karöz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovery, O.M., Ellise, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., … Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Research*, 17, 691-707. https://doi.org/10.1101/gr.5704207

Kooistra, S.M. & Helin, K. (2012). Molecular mechanisms and potential functions of histone demethylases. *Nature Reviews Molecular Cell Biology*, 13, 297-311. https://doi.org/10.1038/nrm3327

Koppens, M. A., Bounova, G., Gargiulo, G., Tanger, E., Janssen, H., Cornelissen-Steijger, P., Blom, M., Song, J. Y., Wessels, L. F., & van Lohuizen, M. (2016). Deletion of Polycomb Repressive Complex 2 From Mouse Intestine Causes Loss of Stem Cells. *Gastroenterology*, 151(4), 684–697.e12. https://doi.org/10.1053/j.gastro.2016.06.020

Krueger, Felix. 2019. Trim Galore! (version 0.6.5) https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Kwon, T., Chang, J. H., Kwak, E., Lee, C. W., Joachimiak, A., Kim, Y. C., Lee, J., & Cho, Y. (2003). Mechanism of histone lysine methyl transfer revealed by the structure of SET7/9-AdoMet. *The EMBO Journal*, 22(2), 292–303. https://doi.org/10.1093/emboj/cdg025

Lai, B., Lee, J.E., Jang, Y., Wang, L., Peng, W., & Ge, K. (2017). MLL3/MLL4 are required for CBP/p300 binding on enhancers and super-enhancer formation in brown adipogenesis. *Nucleic Acids Research*, 45(11), 6388-6403. https://doi.org/10.1093/nar/gkx234

Lampe, V., Dierks, C., & Distl, O. (2009). Refinement of a quantitative trait loci on equine chromosome 5 responsible for fetlock osteochondrosis in Hanoverian warmblood horses. *Animal Genetics*, 40(4), 553-555. https://doi.org/10.1111/j.1365-2052.2009.01865.x

Lan, F., Bayliss, P.E., Rinn, J.L., Whetstine, J.R., Wang, J.K., Chen, S., Iwase, S., Alpatov, R., Issaeva, I., Canaani, E., Roberts, T.M., Chang, H.Y., & Shi, Y. (2007). A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature,* 449, 689-694. https://doi.org/10.1038/nature06192

Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., & Lieb, J.D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 36, 900-905. https://doi.org/10.1038/ng1400

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isonon, K., Koseki, H., Fuchikami, K., Abe, K., Murray, H.L., Zucker, J.P., Yuan, B., Bell, G.W., Herbolsheimer, E., Hannet, N.M., … Young, R.A. (2006). Control of developmental regulators by polycomb in human embryonic stem cells. *Cell,* 125(2), 301-313. https://doi.org/10.1038/nature04733

Li, F., Zhang, D., & Fujise, K. (2001). Characterization of fortilin, a novel antiapoptotic protein. *Journal of Biological Chemistry*, 276(50), P47542-P47549. https://doi.org/10.1074/jbc.M108954200

Li, H., Ilin, S., Wang, W., Duncan, E.M., Wysocka, J., Allis, C.D., & Patel, D.J. (2006). Molecular basis for site-specifc read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, 442, 91-95. https://doi.org/10.1038/nature04802

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics*. https://doi.org/10.6084/M9.FIGSHARE.963153.V1

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

List, K., Szabo, R., Wertz, P.W., Segre, J., Haudenschild, C.C., Kim, S.Y., & Bugge T.H. (2003). Loss of proteolytically processed filaggrin caused by epidermal deletion of Matriptase/MT-SP1. *Journal of Cell Biology*, 163(4), 901-910. https://doi.org/10.1083/jcb.200304161

Lloret-Villas, A., Bhati, M., Kadri, N.K., Fries, R., & Pausch, H. (2021). Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. *BMC Genomics*, 22, 363. https://doi.org/10.1186/s12864-021-07554-w

Local, A., Huang, H., Albuquerque, C.P., Singh, N., Lee, A.Y., Wang, W., Wang, C., Hsia, J.E., Shiau, A.K., Ge, K., Corbett, K., Wang, D., Zhou, H., & Ren, B. (2018). Identification of H3K4me1-associated proteins at mammalian enhancers. *Nature Genetics*, 50(1), 73-82. https://doi.org/10.1038/s41588-017-0015-6

Love, M.I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Low, W.Y., Tearle, R., Liu, R., Koren, S., Rhie, A., Bickhart, D.M., Rosen, B.D., Kronenberg, Z.N., Kingar, S.B., Tseng, E., Thibaud-Nissen, F., Martin, F.J., Billis, K., Ghurye, J., Hastie, A.R., Lee, J., Pang, A.W.C., Heaton, M.P., Phillippy, A.M., … Williams, J.L. (2020). Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nature Communications*, 11, 2071. https://doi.org/10.1038/s41467-020-15848-y

Lykkjen, S., Dolvik, N.I., McCue, M.E., Mickelson, J.R., & Roed, K.H. (2010). Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters. *Animal Genetics,* 41(s2), 111-120. https://doi.org/10.1111/j.1365-2052.2010.02117.x

Mansour, T.A., Scott, E.Y., Finno, C.J., Bellone, R.R., Mienaltowski, M.J., Penedo, M.C., Ross, P.J., Valberg, S.J., Murray, S.D., & Brown, C.T. (2017). Tissue resolved, gene structure refined equine transcriptome. *BMC Genomics*, 18, 103. https://doi.org/10.1186/s12864-016-3451-2

Mayne, B.T., Bianco-Miotto, T., Buckberry, S., Breen, J., Clifton, V., Shoubridge, C., & Roberts, C.T. (2016). Large scale gene expression meta-analysis reveals tissue-specific, sex-biased gene expression in humans. *Frontiers in Genetics*, 7. https://doi.org/10.3389/fgene.2016.00183

McCoy, A.M., Beeson, S.K., Splan, R.K., Lykkjen, S., Ralston, S.L., Mickelson, J.R., & McCue, M.E. (2016). Identification and validation of risk loci for osteochondrosis in standardbreds. *BMC Genomics*, 17, 41. https://doi.org/10.1186/s12864-016-2385-z

McCue, M.E., Bannasch, D.L., Petersen, J.L., Gurr, J., Bailey, E., Binns, M.M., Distl, O., Guérin, G., Hasegawa, T. Hill, E.W., Leeb, T., Lindgren, G., Penedo, C.T., Røed, K.H., Ryder, O.A., Swinburne, J.E., Tozaki, T., Valberg, S.J., Lindblad-Toh, K., … Mickelson, J.R. (2012). A high density SNP array for the domestic horse and extant Perissodactyla: Utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genetics,* 8(1), e1002451. https://doi.org/10.1371/journal.pgen.1002451

McGinty, R. K., & Tan, S. (2015). Nucleosome structure and function. *Chemical Reviews*, 115(6), 2255–2273. https://doi.org/10.1021/cr500373h

McGivney, B.A., McGettigan, P.A., Browne, J.A., Evans, A.C.O., Fonseca, R.G., Loftus, B.J., Lohan, A., MacHugh, D.E., Murphy, B.A., Katz, L.M., & Hill, E.W. (2010). Characterization of the equine skeletal muscle transcriptome identifies novel functional response to exercise training. *BMC Genomics,* 11, 398. https://doi.org/10.1186/1471-2164-11-398

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., GTEx Consortium, Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., … Guigó, R. (2015). Human genomics. The human transcriptome across tissues and individuals. *Science (New York, N.Y.)*, 348(6235), 660–665. https://doi.org/10.1126/science.aaa0355

Mellor, J. (2005). The dynamics of chromatin remodeling at promoters. *Molecular Cell,* 19(2), 147-157. https://doi.org/10.1016/j.molcel.2005.06.023

Meritxell, O., Muñoz-Aguirre, M., Kim-Hellmuth, S., Wucher, V., Gerwirtz, A.D.H., Cotter, D.J., Parsana, P., Kasela, S., Balliu, B., Viñuela, A., Castel, S.E., Mohammadi, P., Aguet, F., Zou, Y., Khramtsova, E.A., Skol, A.D., Garrido-Martín, D., Reverter, F., Brown, A., … Stranger, B.E. (2020). The impact of sex on gene expression across human tissues. *Science*, 369(6509). https://doi.org/10.1126/science.aba3066

Metzger, J., Ohnesorge, B., & Distl, O. (2012). Genome-wide linkage and association analysis identifies major gene loci for guttural pouch tympany in Arabian and German Warmblood horses. *PLoS One,* 7(7), e41640. https://doi.org/10.1371/journal.pone.0041640

Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Moreton, J., Malla, S., Aboobaker, A.A., Tarlinton, R.E., & Emes, R.D. (2014). Characterisation of the horse transcriptome from immunologically active tissues. *PeerJ,* 2, e382. https://doi.org/10.7717/peerj.382

Morrison, O. & Thakur, J. (2021). Molecular complexes at euchromatin, heterochromatin and centromeric chromatin. *International Journal of Molecular Sciences,* 22(13), 6922. https://doi.org/10.3390/ijms22136922

Murray, K. (1964). The occurance of ε-*N*-methyl lysine in histones. *Biochemistry*, 3(1), 10-15. https://doi.org/10.1021/bi00889a003

Narita, T., Ito, S., Higashijima, Y., Chu, W. K., Neumann, K., Walter, J., Satpathy, S., Liebner, T., Hamilton, W. B., Maskey, E., Prus, G., Shibata, M., Iesmantavicius, V., Brickman, J. M., Anastassiadis, K., Koseki, H., & Choudhary, C. (2021). Enhancers are activated by p300/CBP activity-dependent PIC assembly, RNAPII recruitment, and pause release. *Molecular Cell*, 81(10), 2166–2182.e6. https://doi.org/10.1016/j.molcel.2021.03.008

Nurse, N. P., Jimenez-Useche, I., Smith, I. T., & Yuan, C. (2013). Clipping of flexible tails of histones H3 and H4 affects the structure and dynamics of the nucleosome. *Biophysical Journal*, 104(5), 1081–1088. https://doi.org/10.1016/j.bpj.2013.01.019

Nurse, N. P., Jimenez-Useche, I., Smith, I. T., & Yuan, C. (2013). Clipping of flexible tails of histones H3 and H4 affects the structure and dynamics of the nucleosome. *Biophysical Journal*, 104(5), 1081–1088. https://doi.org/10.1016/j.bpj.2013.01.019

Oksuz, O., Narendra, V., Lee, C. H., Descostes, N., LeRoy, G., Raviram, R., Blumenberg, L., Karch, K., Rocha, P. P., Garcia, B. A., Skok, J. A., & Reinberg, D. (2018). Capturing the Onset of PRC2-Mediated Repressive Domain Formation. *Molecular Cell*, 70(6), 1149–1162.e5. https://doi.org/10.1016/j.molcel.2018.05.023

Pacholewska, A., Drögmüller, M., Klukowska-Rötzler, J., Lanz, S., Hamza, E., Dermitzakis, E.T., Marti, E., Gerber, V., Leeb, T., & Jagannathan, V. (2015). The transcriptome of equine peripheral blood mononuclear cells. *PLoS ONE*, 10(3): e0122011. https://doi.org/10.1371/journal.pone.0122011

Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L., Kern, C., Halstead, M., Chanthaviaxy, G., Trakooljul, N., Wimmers, K., Sahana, G., Su, G., Lund, M.S., Fredholm, M., Karlskov-Mortensen, P., Ernst, C.W., Ross, P., Tuggle, C.K., Fang, L., & Zhou, H. (2021). Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nature Communications*, 12, 5848. https://doi.org/10.1038/s41467-021-26153-7

Park, K.D., Park, J., Ko, J., Kim, B.C., Kim, H.S., Ahn, K., Do, K.T., Choi, H., Kim, H.M., Song, S., Lee, S., Jho, S., Kong, H.S., Yang, Y.M., Jhun, B.H., Kim, C., Kim, T.H., Hwang, S., Bhak, J., … , Cho, B.W. (2012). Whole transcriptome analyses of six Thoroughbred horses before and after exercise using RNA-seq. *BMC Genomics*, 13, 473. https://doi.org/10.1186/1471-2164-13-473

Parvi, R., Zhu, B., Li, G., Trojer, P., Mandal, S., Shilatifard, A., & Reinberg, D. (2006). Histone H2B monoubiquination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell*, 125(4), 703-717. https://doi.org/10.1016/j.cell.2006.04.029

Paterson, Y.Z., Cribbs, A., Espenel, M., Smith, E.J., Henson, F.M.D., & Guest, D.J. (2020). Genome-wide transcriptome analysis reveals equine embryonic stem cell-derived tenocytes resemble fetal, not adult tenocytes. *Stem Cell Research & Therapy*, 11, 184. https://doi.org/10.1186/s13287-020-01692-w

Petersen, J.L., Mickelson, J.R., Rendahl, A.K., Valberg, S.J., Andersson, L.S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M.M., Borges, A.S., Brama, P. Machado, A.D.C., Capomaccio, S., Cappelli, K., Cothran, E.G., Distl, O., Fox-Clipsham, L., Graves, K.T.,

Guérin, G., … McCue, M.E. (2013). Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genetics*, 9(1), e1003211. https://doi.org/10.1371/journal.pgen.1003211

Pinkaew, D., Chattopadhyay, A., King, M.D., Chunhacha, P., Liu, Z., Stevenson, H.L., Chen, Y., Sinthujaroen, P., McDougal, O.M., & Fujise, K. (2017). Fortilin binds IRE1α and prevents ER stress from signaling apoptotic cell death. *Nature Communications*, 8, 18. https://doi.org/10.1038/s41467-017-00029-1

Porsch, R.M., Merello, M., De Marco, P., Cheng, G., Rodriguez, L., So, M., Sham, P.C., Tam, P.K., Capra, V., Cherny, S.S., Garcia-Barcelo, M.M., and Campbell, D.D. (2016). Sacral agenesis: a pilot whole exome sequencing and copy number study. *BMC Medical Genetics*, 17, 98. https://doi.org/10.1186/s12881-016-0359-2

Prothero, K.E., Stahl, J.M. & Carrel, L. (2009). Dosage compensation and gene expression on the mammalian X chromosome: One plus one does not always equal two. *Chromosome Research*, 17, 637–648. https://doi.org/10.1007/s10577-009-9063-9

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1), W160-W165. https://doi.org/10.1093/nar/gkw257

Ramsköld, D., Wang, E. T., Burge, C. B., & Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Computational Biology*, 5(12), e1000598. https://doi.org/10.1371/journal.pcbi.1000598

Raudsepp, T., McCue, M.E., Das, P.J., Dobson, L., Vishnoi, M., Fritz, K.L., Schaefer, R., Rendahl, A.K., Derr, J.N., Love, C.C., Varner, D.D., & Chowdhary, B.P. (2012). Genome-wide association study implicates testis-sperm specific *FKBP6* as a susceptibility locus for impaired acrosome reaction in stallions. *PLoS Genetics*, 8(12), e1003139. https://doi.org/10.1371/journal.pgen.1003139

Rebolledo-Mendez, J., Hestand, M.S., Coleman, S.J., Zeng, Z., Orlando, L., MacLeod, J.N., & Kalbfleisch, T. (2015). Comparison of the equine reference sequence with its Sanger source data and new Illumina reads. *PLoS One*, 10(6), e0126852. https://doi.org/10.1371/journal.pone.0126852

Ropka-Molik, K., Stefaniuk-Szmukier, M., Z'ukowski, K., Piórkowska, K., & Bugno-Poniewierska, M. (2017). Exercise-induced modification of the skeletal muscle

transcriptome in Arabian horses. *Physiological Genomics,* 49(6), 318-326. https://doi.org/10.1152/physiolgenomics.00130.2016

Rosen, B.D., Bickhart, D.M., Schnabel, R.D., Koren, S., Elsik, C.G., Tseng, E., Rowan, T.N., Low, W.Y., Zimin, A., Couldrey, C., Hall, R., Li, W., Rhie, A., Ghurye, J., McKay, S.D., Thibaud-Nissen, F., Hoffman, J., Murdoch, B.M., Snelling, W.M., … Medrano, J.F. (2020). *GigaScience*, 9, 1-9. https://doi.org/10.1093/gigascience/giaa021

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C.T., Schreiber, S.L., Mellor, J., & Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature*, 419, 407-411. https://doi.org/10.1038/nature01080

Schaefer, R.J., Schubert, M., Bailey, E., Bannasch, D.L., Barrey, E., Bar-Gal, G.K., Brem, G., Brooks, S.A., Distl, O., Fries, R., Finno, C.J., Gerber, V., Haase, B., Jagannathan, V., Kalbfleisch, T., Leeb, T., Lindgren, G., Lopes, M.S., Mach, N., … McCue, M.E. (2017). Developing a 670K genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics*, 18, 565. https://doi.org/10.1186/s12864-017-3943-8

Schaefer, R.J. & McCue, M.E. (2020). Equine Genotyping Arrays. *Veterinary Clinics of North America: Equine Practice,* 36(2), 183-193. https://doi.org/10.1016/j.cveq.2020.03.001

Schmitges, F. W., Prusty, A. B., Faty, M., Stützer, A., Lingaraju, G. M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., Bunker, R. D., Wirth, U., Bouwmeester, T., Bauer, A., Ly-Hartig, N., Zhao, K., Chan, H., Gu, J., Gut, H., Fischle, W., … Thomä, N. H. (2011). Histone methylation by PRC2 is inhibited by active chromatin marks. *Molecular Cell*, 42(3), 330–341. https://doi.org/10.1016/j.molcel.2011.03.025

Schneider, R., Bannister, A.J., Myers F.A., Thorne, A.W., Robinson, C.C., & Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biology*, 6, 73-77. https://doi.org/10.1038/ncb1076

Schubert, M., Jónsson, H., Chang, D., Sarkissian, C.D., Ermini, L., Ginolhac, A., Albrechsten, A., Dupanloup, I., Foucal, A., Petersen, B., Fumagalli, M., Raghavan, M., Seguin-Orlando, A., Korneliussen, T.S., Velazquez, A.M.V., Stenderup, J., Hoover, C.A., Rubin, C.J., Alfarhan, A.H., … Orlando, L. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *PNAS*, 111(52), E5661-E5669. https://doi.org/10.1073/pnas.1416991111

Schuettengruber, B., Ganapathi, M., Leblanc, B., Portoso, M., Jaschek, R., Tolhuis, B., van Lohuizen, M., Tanay, A., & Cavalli, G. (2009). Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS Biology*, 7(1), e1000013. https://doi.org/10.1371/journal.pbio.1000013

Scott, E. Y., Mansour, T., Bellone, R. R., Brown, C. T., Mienaltowski, M. J., Penedo, M. C., Ross, P. J., Valberg, S. J., Murray, J. D., & Finno, C. J. (2017). Identification of long non-coding RNA in the horse transcriptome. *BMC Genomics*, 18(1), 511. https://doi.org/10.1186/s12864-017-3884-2

Shen, E. Y., Ahern, T. H., Cheung, I., Straubhaar, J., Dincer, A., Houston, I., de Vries, G. J., Akbarian, S., & Forger, N. G. (2015). Epigenetics and sex differences in the brain: A genome-wide comparison of histone-3 lysine-4 trimethylation (H3K4me3) in male and female mice. *Experimental Neurology*, 268, 21–29. https://doi.org/10.1016/j.expneurol.2014.08.006

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J. R., Cole, P. A., Casero, R.A., & Shi, Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell, 119(7), 941-953. https://doi.org/10.1016/j.cell.2004.12.012

Sims, R.J., 3rd, Chen, C.F., Santos-Rosa, H., Kouzarides, T., Patel, S.S., & Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *Journal of Biological Chemistry*, 280(51), 41789-41792. https://doi.org/10.1074/jbc.C500395200

Sims, R. J., 3rd, Millhouse, S., Chen, C. F., Lewis, B. A., Erdjument-Bromage, H., Tempst, P., Manley, J. L., & Reinberg, D. (2007). Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Molecular Cell*, 28(4), 665–676. https://doi.org/10.1016/j.molcel.2007.11.010

Singer-Hasler, H., Flury, C., Hasse, B., Burger, D., Simianer, H., Leeb, T., & Rieder, S. (2012). A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLoS One,* 7(5), e37282. https://doi.org/10.1371/journal.pone.0037282

Spencer, G. (2001, February 12). *International Human Genome Sequencing Consortium publishes sequence and analysis of the human genome.* National Human Genome Research Institute. https://www.genome.gov/10002192/2001-release-first-analysis-of-human-genome

Srikanth, K., Kim, N.Y., Park, W, Kim, J.M., Kim, K.D., Lee, K.T., Son, J.H., Chai, H.H., Choi, J.W., Jang, G.W., Kim, H., Ryu, Y.C., Nam, J.W., Park, J.E., Kim, J.M., & Lim, D. (2019). Comprehensive genome and transcriptome analyses reveal genetic relationship, selection signature, and transcriptome landscape of small-sized Korean native Jeju horse. *Scientific Reports,* 9, 16672. https://doi.org/10.1038/s41598-019-53102-8

Stein, M.M., Conery, M., Magnaye, K.M., Clay, S.M., Billstrand, C., Nicolae, R., Naughton, K., Ober, C., & Thompson, E.E. (2021). Sex-specific differences in peripheral blood leukocyte transcriptional response to LPS are enriched for HLA region and X chromosome genes. *Scientific Reports*, 11, 1107. https://doi.org/10.1038/s41598-020-80145-z

Steinhauser, S., Kurzawa, N., Eils, R., & Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in Bioinformatics*, *17*(6), 953–966. https://doi.org/10.1093/bib/bbv110

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, 31(12), 2032-2034. https://doi.org/10.1093/bioinformatics/btv098

Taunton, J., Hassig, C. A., & Schreiber, S. L. (1996). A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science (New York, N.Y.),* 272(5260), 408–411. https://doi.org/10.1126/science.272.5260.408

The ENCODE Project Consortium (2004). The ENCODE (encyclopedia of DNA elements) project. *Science,* 306, 636-640. https://doi.org/10.1126/science.1105136

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* 447, 799-816. https://doi.org/10.1038/nature05874

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489(7414), 57-74. https://doi.org/10.1038/nature11247

The FAANG Consortium, Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., Dalrymple, B.P., Elski, C.G., Foissac, S., Giuffra, E., Groenen, M.A., Hayes, B.J., Huang, L.S., Khatib, H., … Zhou, H. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biology*, 16, 57. https://doi.org/10.1186/s13059-015-0622-4

Tie, F., Banerjee, R., Stratton, C. A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M. O., Scacheri, P. C., & Harte, P. J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development (Cambridge, England)*, 136(18), 3131–3141. https://doi.org/10.1242/dev.037127

Tse, C., & Hansen, J. C. (1997). Hybrid trypsinized nucleosomal arrays: identification of multiple functional roles of the H2A/H2B and H3/H4 N-termini in chromatin fiber compaction. *Biochemistry*, 36(38), 11381–11388. https://doi.org/10.1021/bi970801n

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, S., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., . . . Pontén, F. (2015). Tissue-based map of the human proteome. *Science*, *347*(6220). https://doi.org/10.1126/science.1260419

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L., Melo, M., Rocha, A. G. D., Preto, A., Castro, P., Castro, L., Pardal, F., Lopes, J. M., Santos, L. L., Reis, R. M., . . . Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. *Nature Communications*, *4*(1). https://doi.org/10.1038/ncomms3185

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blöcker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N. Penedo, M.C.T., Raison, J.M., … Lindbald-Toh, K. (2009). Genome sequence, comparative analysis and population genetics of the domestic horse (*Equus caballus*). *Science*, 326(5954), 865–867. https://doi.org10.1126/science.1178158

Wang, X., He, C., Moore, S.C., & Ausió, J. (2001). Effects of histone acetylation on the solubility and folding of the chromatin fiber. *Journal of Biological Chemistry*, 276(16), 12764-12768. https://doi.org/10.1074/jbc.M100501200

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapha, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., & Zhao, K. (2008). Combinatorial patterns of histone aceylations and methylations in the human genome. *Nature Genetics*, 40, 897-903. https://doi.org/10.1038/ng.154

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57-63. https://doi.org/10.1038/nrg2484

Warburton, A., Breen, G., Bubb, V. J., & Quinn, J. P. (2015). A GWAS SNP for Schizophrenia Is Linked to the Internal MIR137 Promoter and Supports Differential Allele-Specific Expression. *Schizophrenia Bulletin*, *42*(4), 1003–1008. https://doi.org/10.1093/schbul/sbv144

Wei, X., Peng, H., Deng, M., Feng, Z., Peng, C., & Yang, D. (2020). MiR-703 protects against hypoxia/reoxygenation induced cardiomyocyte injury via inhibiting the NLRP3/caspase-1-mediated pyroptosis. *Journal of Bioenergetics and Biomembranes*, 52, 155-164. https://doi.org/10.1007/s10863-020-09832-w

Wu, Z., Wang, X., and Zhang, X. (2011). Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics.* 27(4), 502-508. https://doi.org/10.1093/bioinformatics/btq696

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., Wu, C., & Allis, C.D. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442, 86-90. https://doi.org/10.1038/nature04815

Xiang, G., Keller, C. A., Giardine, B., An, L., Li, Q., Zhang, Y., & Hardison, R. C. (2020). S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. *Nucleic Acids Research*, *48*(8), e43. https://doi.org/10.1093/nar/gkaa105

Xiang, G., Giardine, B. M., Mahony, S., Zhang, Y., & Hardison, R. C. (2021). S3V2-IDEAS: a package for normalizing, denoising and integrating epigenomic datasets across different cell types. *Bioinformatics (Oxford, England)*, 37(18), 3011–3013. https://doi.org/10.1093/bioinformatics/btab148

Xu, S., Grullon, S., Ge, K., & Peng, W. (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods in Molecular Biology*, 1150, 97–111. https://doi.org/10.1007/978-1-4939-0512-6_5

Xue-Franzén, Y., Henriksson, J., Bürglin, T.R., & Wright, A.P.H. (2013). Distinct roles of the Gnc5 histone acetyltransferase revealed during transient stress-induced reprogramming of the genome. *BMC Genomics*, 14, 479. https://doi.org/10.1186/1471-2164-14-479

Yan, J., Chen, S.A.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., River, C.M., Wang, C., Ye, Z., Ge, K., Hu, M., Wysocka, J., & Ren, B. (2018). Histone H3 lysine 4

monomethylation modulates long-range chromatin interactions at enhancers. *Cell Research*, 28, 204-220. https://doi.org/10.1038/cr.2018.1

Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G., & Chen, L.L. (2011). Genomewide characterization of non-polyadenylated RNAs. *Genome Biology*, 12, R16. https://doi.org/10.1186/gb-2011-12-2-r16

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., & Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, *25*(15), 1952–1958. https://doi.org/10.1093/bioinformatics/btp340

Zeitz, A., Spötter, A., Blazyczek, I., Diesterbeck, U., Ohnesorge, B., Deegen, E., & Distl, O. (2009). Whole-genome scan for guttural pouch tympany in Arabian and German warmblood horses. *Animal Genetics,* 40(6), 917-924. https://doi.org/10.1111/j.1365-2052.2009.01942.x

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology, 9*(9). https://doi.org/10.1186/gb-2008-9-9-r137

Zhang, Q., Zeng, X., Younkin, S., Kawli, T., Snyder, M. P., & Keleş, S. (2016). Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-0957-1

Zhang, T., Zhang, Z., Dong, Q., Xiong, J., & Zhu, B. (2020). Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biology*, 21, 45. https://doi.org/10.1186/s13059-020-01957-w

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., & von Schack, D. (2018). Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: PolyA+ selection versus rRNA depletion. *Scientific Reports,* 8, 4781. https://doi.org/10.1038/s41598-018-23226-4

Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120

Zhao, Y., Li, M. C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq

Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1). https://doi.org/10.1186/s12967-021-02936-w

Zhou, J., Fan, J.Y., Rangasamy, D., & Tremethick, D.J. (2007). The nucleosome surface regulates chromatin compaction and couples it with transcriptional repression. *Nature Structural & Molecular Biology*, 14, 1070-1076. https://doi.org/10.1038/nsmb1323

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., & Salzberg, S.L. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology*, 10, R42. https://doi.org/10.1186/gb-2009-10-4-r42