

University of Kentucky

UKnowledge

---

Theses and Dissertations--Veterinary Science

Veterinary Science

---


2023

## USE OF MOLECULAR GENETICS TO INVESTIGATE POPULATION STRUCTURE AND SWAYBACK IN HORSES

Navid YousefiMashouf

*University of Kentucky*, [navidyousefimashouf@gmail.com](mailto:navidyousefimashouf@gmail.com)

Author ORCID Identifier:

 <https://orcid.org/0000-0003-3853-2697>

Digital Object Identifier: <https://doi.org/10.13023/etd.2023.302>

[Right click to open a feedback form in a new tab to let us know how this document benefits you.](#)

### Recommended Citation

YousefiMashouf, Navid, "USE OF MOLECULAR GENETICS TO INVESTIGATE POPULATION STRUCTURE AND SWAYBACK IN HORSES" (2023). *Theses and Dissertations--Veterinary Science*. 62.

[https://uknowledge.uky.edu/gluck\\_etds/62](https://uknowledge.uky.edu/gluck_etds/62)

This Doctoral Dissertation is brought to you for free and open access by the Veterinary Science at UKnowledge. It has been accepted for inclusion in Theses and Dissertations--Veterinary Science by an authorized administrator of UKnowledge. For more information, please contact [UKnowledge@lsv.uky.edu](mailto:UKnowledge@lsv.uky.edu).

## **STUDENT AGREEMENT:**

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

## **REVIEW, APPROVAL AND ACCEPTANCE**

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Navid YousefiMashouf, Student

Dr. Ernest F. Bailey, Major Professor

Dr. Martin K. Nielsen, Director of Graduate Studies

USE OF MOLECULAR GENETICS TO INVESTIGATE POPULATION STRUCTURE  
AND SWAYBACK IN HORSES

---

DISSERTATION

---

A dissertation submitted in partial fulfillment of the  
requirements for the degree of Doctor of Philosophy  
in the College of Agriculture, Food and Environment at the University of Kentucky

By  
Navid YousefiMashouf  
Lexington, Kentucky  
Director: Dr. Ernest F. Bailey, Professor of Veterinary Science  
Lexington, Kentucky  
2023

Copyright © Navid YousefiMashouf 2023  
<https://orcid.org/0000-0003-3853-2697>

## ABSTRACT OF DISSERTATION

### USE OF MOLECULAR GENETICS TO INVESTIGATE POPULATION STRUCTURE AND SWAYBACK IN HORSES

The present research incorporated molecular genetic methods to 1) investigate the genetic basis of Juvenile Onset Lordosis or Swayback in the American Saddlebred horses; and 2) conduct a population genetic study to compare the Persian Kurdish, Persian Arabian and American Thoroughbred horse populations.

Juvenile-onset lordosis, or swayback, is a condition in horses where the conformational topline back curvature drops significantly within the first two years of life. The trait has a higher prevalence in Saddlebreds (5%). Prior research on them quantified the trait using a Measurement of Back Contour (MBC), defining an MBC of >7.0 centimeters as swayback, and <7.0 as normal. A genome-wide association study comparing low (<5.0) versus high (>8.0) MBC horses suggested a single recessive variant on chromosome20 to be associated with the trait. The present research aimed to find the causal mutation on chr20 using Whole-Genome Sequencing, testing a hypothesis that a single recessive variant on chr20 causes high-MBC. Eleven Saddlebreds were Whole-Genome Sequenced in two experiments. Experiment1 involved 3 high-MBC horses and 3 low-MBC horses with various haplotype structures on chr20. No variants were found on chr20 to support the hypothesis, suggesting more than one major variant to be involved in swayback. Re-evaluation of the association on chr20 was performed via genotyping for tag markers on a chr20 haplotype in 34 high-MBC versus 75 low-MBC Saddlebreds, where a chi-square comparison confirmed that chr20 has a significant impact on high-MBC and that the earlier GWAS association was not a statistical artifact.

We then evaluated all the genomic variation in the target region of chr20:41,000,000-44,000,000 to identify the best candidate to influence high-MBC in Saddlebreds. A total of 9,691 variant loci were detected that make 21,463 transcript variations. Of these, 599 made coding sequence variations, including 315 synonymous, 250 missense, 14 frameshift, 9 in-frame deletion, 7 in-frame insertion, and 2 splice-donor and 2 start-loss variants. The strongest candidate seems to be a frameshift deletion of 7bp in the exon 1 of the *MDFI* gene, at 20:41873061-41873068. *MDFI*-knockout mice show defects in the formation of thoracic vertebrae and ribs, which restrains fusion of the spinous processes.

The last part of the dissertation research is about a study that aimed to characterize the Persian Kurdish horse population relative to the Persian Arabian and American Thoroughbred populations using genome-wide SNP data. Fifty-eight Kurdish, 38 Persian Arabian and 83 Thoroughbred horses were genotyped across 670,796 markers. The Kurdish horses were generally distinguished from the Persian Arabian and Thoroughbred samples by all analyses including Principal Component Analyses, cluster analyses and calculation of pairwise  $F_{ST}$ . These results together identify the Kurdish horse population as a unique, uniform genetic structure.

**KEYWORDS:** Juvenile Onset Lordosis, Whole-Genome Sequencing, SNP, Population Genomics.

Navid YousefiMashouf

---

07/18/2023

---

Date

USE OF MOLECULAR GENETICS TO INVESTIGATE  
POPULATION STRUCTURE AND SWAYBACK IN HORSES

By  
Navid YousefiMashouf

Ernest F. Bailey

---

Director of Dissertation

Martin K. Nielsen

---

Director of Graduate Studies

07/18/2023

---

Date

## DEDICATION

To my mother who sacrificed her life to see me thrive.

## ACKNOWLEDGMENTS

The following dissertation, while an individual work, benefited from the insights and direction of several people. First, my Dissertation Chair, Dr Ernest Bailey, exemplifies the high-quality scholarship to which I aspire. Next, I wish to thank the complete Dissertation Committee, and outside reader, respectively: Dr. Kathryn Graves, Dr. Theodore Kalbfleisch, Dr. James MacLeod and Dr. Brett Spear. Each provided insights that guided and challenged my thinking, substantially improving the finished product.

I would also like to acknowledge the American Saddlebred Horse Association, that generously provided the funding for my research. Especially, I would like to thank Mr. Fred Sarver, whose support and advice played a key role in conducting this research. All horse owners who were kind to contribute samples to this research are greatly appreciated. Also, Geoffrey Hughes foundation is hereby acknowledged and sincerely appreciated for providing funding to support my graduate studies.

In addition to the technical and instrumental assistance above, I received equally important assistance from friends. Dr John Eberth and James Norris are of special note.



# TABLE OF CONTENTS

ACKNOWLEDGMENTS.....	iii
TABLE OF CONTENTS .....	iv
LIST OF TABLES .....	viii
LIST OF FIGURES.....	ix
LIST OF ADDITIONAL FILES .....	xi
<b>CHAPTER 1. INTRODUCTION AND BACKGROUND INFORMATION.....</b>	<b>1</b>
<i>1.1 Juvenile Onset Lordosis in American Saddlebred Horses .....</i>	<i>1</i>
1.1.1 Genetic Studies of Extreme Lordosis Among Saddlebred Horses .....	4
1.1.2 Genetic Models.....	8
1.1.3 Example: Genetics of height in horses .....	8
<i>1.2 Use of Molecular Genetics to Infer Relationships Among Horse Populations .....</i>	<i>10</i>
1.2.1 Persian Breeds of Horse .....	10
1.2.1.1 Turkoman.....	10
1.2.1.2 Caspian.....	12
1.2.1.3 Persian Arabian.....	13
1.2.1.4 Kurdish.....	14
1.2.2 Molecular Genetic Studies comparing horse populations .....	16
1.2.2.1 Studies with Blood Groups and Biochemical Markers.....	16
1.2.2.2 Genetic Characterization Studies Using Microsatellite Markers.....	17
1.2.2.3 Population Genetic Studies Using Single Nucleotide Polymorphism Markers	19
1.2.2.4 Population Genetics Studies on Persian Horse Breeds.....	20
<b>CHAPTER 2. FINE MAPPING STUDIES OF HORSE CHROMOSOME 20 REGION ASSOCIATED WITH JUVENILE ONSET LORDOSIS IN AMERICAN SADDLEBRED HORSES.....</b>	<b>23</b>
2.1 Summary .....	23
2.2 Introduction.....	24
2.3 Materials and Methods .....	25
2.3.1 Experimental Animals for Whole-Genome Sequencing.....	25
2.3.2 Experimental Animals for Population Genotyping .....	26
2.3.3 Phenotyping.....	26
2.3.4 DNA Isolation .....	27
2.3.5 Whole Genome Sequencing .....	27
2.3.6 Analysis of Whole Genome Sequence .....	28
2.3.7 Genomic Region under Study.....	28

2.3.8	Experimental Design .....	28
2.3.8.1	Experiment 1: Whole-genome comparison of unrelated horses of low versus high MBC .....	29
2.3.8.1.1	Horses in Experiment 1 .....	29
2.3.8.1.2	Variant filtering criteria for Experiment 1 .....	30
2.3.8.2	Experiment 2: Whole-genome study of the core family .....	30
2.3.8.2.1	Horses in Experiment 2 .....	30
2.3.8.2.2	Variant Filtering Criteria for Experiment 2 .....	31
2.3.9	Genetic Marker Genotyping for Re-evaluation of the chr20 Association .....	32
2.3.9.1	ERE1 deletion at chr20:42,222,093 .....	32
2.3.9.2	SNP chr20:42,247,262G>A .....	33
2.3.9.3	215bp Deletion at chr20:42,399,504 .....	33
2.3.10	Analysis of Genotype Distribution and Association Tests .....	33
2.4	<i>Results</i> .....	33
2.4.1	Overall WGS Results .....	33
2.4.2	WGS results for the target region: chr20:41M-44M .....	34
2.4.3	Observations about the MBC phenotypic measurement .....	39
2.4.4	Population Genotyping of Genetic Markers to Test Significance of Association of chr20 Region 40	
2.5	<i>Discussion</i> .....	41
2.5.1	Results from filtering VCF files using criteria for a recessive mode of inheritance .....	41
2.5.2	Construction of Haplotypes Across chr20:41M-44M .....	41
2.5.3	Population Genotyping to Consider possibility of Statistical Artifact .....	42
2.5.4	Insights about the Measurement of Back Contour .....	42
2.5.5	Approaches to Identifying Epistatic Genetic Factors .....	43

## CHAPTER 3. INVESTIGATION OF THE GENETIC EFFECT OF A CHROMOSOME 20 FACTOR ON DEVELOPMENT OF THE JUVENILE ONSET LORDOSIS IN AMERICAN SADDLEBRED HORSES

44

3.1	<i>Summary</i> .....	44
3.2	<i>Introduction</i> .....	45
3.3	<i>Materials and Methods</i> .....	46
3.3.1	Phenotype .....	46
3.3.2	Bioinformatic analysis of the variants identified from the Whole Genome Sequence Data ..	48
3.4	<i>Results</i> .....	49
3.4.1	Investigation of Frameshift Variants .....	50
3.4.2	Investigation of Missense Variants .....	52
3.4.3	Investigation of in-frame insertions/deletions .....	56
3.4.4	Visual Investigation of the Structural Variation .....	57
3.5	<i>Discussion</i> .....	58
3.5.1	<i>MDFI</i> gene variant .....	59
3.5.2	ERE1 deletion at chr20:42,222,093 around 3'UTR of the <i>TAF8</i> gene .....	64
3.5.3	Missense SNP chr20:42,247,262G>A within the <i>C6orf132</i> gene .....	65
3.5.4	215bp Deletion at chr20:42,399,504-42,399,718 within <i>TRERF1</i> gene .....	65
3.5.5	Future Directions .....	67

CHAPTER 4. GENOMIC COMPARISONS OF PERSIAN KURDISH, PERSIAN ARABIAN AND AMERICAN THOROUGHBRED HORSE POPULATIONS.....	70
4.1 Summary .....	70
4.2 Introduction.....	71
4.3 Materials and Methods .....	72
4.3.1 Sampled Individuals .....	72
4.3.2 Blood collection and DNA extraction .....	74
4.3.3 Ethical Statement.....	74
4.3.4 Genotyping .....	74
4.3.5 Data Analysis .....	74
4.4 Results.....	76
4.4.1 Data Pruning.....	76
4.4.2 Principal Component Analysis (PCA).....	76
4.4.3 $F_{ST}$ .....	77
4.4.4 Population Specific Inbreeding .....	78
4.4.5 Analysis of Molecular Variance (AMOVA) .....	78
4.4.6 Runs Of Homozygosity Analysis .....	78
4.4.7 Expected Heterozygosity ( $H_E$ ).....	81
4.4.8 Cluster Analysis.....	81
4.5 Discussion.....	84
4.5.1 Analyses of Population Structure .....	85
4.5.2 Analyses of Individual Diversity .....	87
4.6 Conclusions.....	88
REFERENCES.....	89
VITA .....	94

## LIST OF TABLES

Table 2.1 MBC phenotype and zygosity of the TGTG haplotype in the horses included in the association validation study. “other” represents non-TGTG haplotypes at these sites, e.g., CACT, CATT, CACG, etc.....	26
Table 2.2 Horses with available whole genome sequence data. The red highlight marks the haplotype with highest frequency in the swayback horses in the study by Cook et al. Red color in the last column marks the alleles that were found associated with high-MBC in Cook et al. ....	29
Table 2.3 Specifications of the members of the core family. Same descriptions in the caption of table 2.2 apply to this table. ....	30
Table 2.4 Summary statistics of the Whole Genome Sequence data generated for each of the horses in the study. Depth of coverage has been calculated by dividing the Total read bases by the total sequence length of the EquCab3.0 reference genome assembly reported by NCBI to be 2,506,966,135 base pairs. ....	34
Table 2.5 Distribution of genotypes for each of the horses in the target region of chr20:41M-44M.....	35
Table 2.6 Alternate-homozygous variants shared among high-MBC horses and their comparison to individual low-MBC horses in Experiment 1. ....	35
Table 2.7 Alternate-homozygous variants shared among high-MBC horses and their comparison to individual low-MBC horses in Experiment 2. ....	36
Table 2.8 Distribution of genotypes for ERE1 deletion at 42,222,093, Missense SNP of <i>C6orf132</i> variant at 42,247,262, and 215bp deletion at 42,399,504, and comparison of between high-MBC to low-MBC (Chi-Square test) for each marker. For the ERE1 and 215bp deletion, D corresponds to the deleted (alternate) allele, whereas N signifies the intact (reference) allele. For <i>C6orf132</i> SNP, A is the alternate and G is the reference allele. Numbers in parentheses are horses from that group that were not used in the previous GWAS study (Cook et al., 2010), e.g., data is shown that 28 horses possessed the DD genotype and 9 of those were newly sampled for this study. ....	40
Table 3.1 Summary statistics of the variants included in the VCF file at the target region of chr20:41M-44M.....	49
Table 3.2 Frameshift variants identified within the target region of chr20:41M-44M.....	50
Table 3.3 Genotype distributions of the frameshifts in the target region Chr20:41M-44M called by GATK Haplotype Caller in the cohort of 11 horses whole-genome sequenced in this study. In the genotype cells, 0 codes for the reference allele and 1, 2 and 3 code for the alternate non-reference alleles, and dot “.” stands for unknown basecalls. For ease of visualization, homozygous reference genotypes have been colored in light green, heterozygous 0_1 in light orange, and homozygous for the non-ref in light red. Also at the top row, high-MBC horses have been color shaded as dark red, control low-MBC horses in dark green, and the horses with low-MBC but questionable back structure in dark orange. Positions highlighted in blue identify the variants whose genotype distribution is concordant with haplotype pattern associated with the high-MBC. ....	51
Table 3.4 Missense variants with a putative deleterious SIFT scores in the target region of chr20:41M-44M.....	52

Table 3.5 Genotype distributions of the missense variants in the Table 3.4 called by GATK Haplotype Caller in the cohort of 11 horses whole-genome sequenced in this study. In the genotype cells, 0 codes for the reference allele and 1 code for the alternate non-reference alleles, and dot “.” stands for unknown basecalls. For ease of visualization, homozygous reference genotypes have been colored in light green, heterozygous 0\_1 in light orange, and homozygous for the non-ref in light red. Also at the top row, high-MBC horses have been color shaded as dark red, control low-MBC horses in dark green, and the horses with low-MBC but questionable back structure in dark orange. Positions highlighted in blue identify the variants whose genotype distribution is concordant with haplotype pattern associated with the high-MBC..... 54

Table 3.6 In-frame indel variants called by GATK Haplotype Caller within the genomic region of chr20:41M-44M among the cohort of 11 whole-genome sequenced horses. ... 56

Table 3.7 Genotype distributions of the in-frame indel variants in the Table 6 called by GATK Haplotype Caller in the cohort of 11 horses whole-genome sequenced in this study. In the genotype cells, 0 codes for the reference allele and 1, 2 and 3 code for the alternate non-reference alleles. For ease of visualization, homozygous reference genotypes have been colored in light green, heterozygous 0\_1 in light orange, and homozygous for the non-ref in light red. Also at the top row, high-MBC horses have been color shaded as dark red, control low-MBC horses in dark green, and the horses with low-MBC but questionable back structure in dark orange. Positions highlighted in blue identify the variants whose genotype distribution is concordant with haplotype pattern associated with the high-MBC. .... 57

Table 4.1 Pairwise  $F_{ST}$  values between breed groups. All of the P values were significant ( $P < 0.05$ ). ..... 77

Table 4.2 Population-specific  $F_{IS}$  values. .... 78

Table 4.3 Summary of the total number of runs of homozygosity by each size class. Due to unequal sample size, three iterations, each of 20 randomly selected individuals was evaluated. Given is the mean and standard deviation (in parenthesis) of the replicates.. 79

Table 4.4 Average Expected Heterozygosity values for each breed group. .... 81

Table 4.5 Results on the comparisons of  $K=1$  to 5 tested by STRUCTURE Harvester. The highlighted row belongs to the  $K$  value ( $=2$ ) that maximizes Delta  $K$  per the Evanno method of determining the best fit for the data..... 82

Table 4.6 Average proportion of membership of each pre-defined population in each of the 2 clusters at  $K=2$ . .... 82

Table 4.7 Average proportion of membership of each pre-defined population in each of the 3 clusters at  $K=3$ . .... 83

Table 4.8 Average proportion of membership of each pre-defined population (excluding the Thoroughbred samples from the dataset) in each of the 2 clusters at  $K=2$ . .... 84

## LIST OF FIGURES

Figure 1.1 Comparison of a horse affected with extreme lordosis (left) to a normal-back horse (right).....	2
Figure 1.2 Images taken by Rooney and Prickett comparing the articular processes in thoracic vertebrae in lordotic versus normal horses (Rooney & Prickett, 1967).....	3
Figure 1.3 Visual representation of MBC measurement method. ....	5
Figure 1.4 Distribution of MBC measurements in the population of 305 ASB horses (black bars). White bars represent Arabian horses and grey bars designate Saddlebred-Arabian crosses (Gallagher et al., 2003).....	6
Figure 1.5 Manhattan plot of association p-values versus genomic location in the study by Cook et al. The peak of association on chromosome 20 has been encircled with a red-dotted rectangle (D. Cook et al., 2010).....	7
Figure 1.6 Haplotype by high-MBC/low-MBC status in the study by Cook et al (D. Cook et al., 2010). ....	8
Figure 1.7 Akhal-Teke sub-breed of the Turkoman horse.....	11
Figure 1.8 Yamut sub-breed of the Turkoman horse.....	12
Figure 1.9 Caspian horse.....	13
Figure 1.10 Persian Arabian horse.....	14
Figure 1.11 Kurdish horse.....	15
Figure 1.12 Kurdish horse.....	15
Figure 1.13 Kurdish horse.....	16
Figure 2.1 Side photos showing back conformation of the horses in the core family study. ....	31
Figure 2.2 Haplotype structure of the target genomic region chr20:41.0-44.0M on the horses studied in both experiments. Red color identifies the haplotype containing the TGTG alleles that are most commonly found in swayback horses. White color marks the reference haplotype found in the reference genome. Other colors identify the haplotypes unique to the horses in our study, which were minor modifications of the reference haplotype. Relative distances are proportional. Body colors or photo outlines of the horses in the diagram of the Experiment 1: Red) high-MBC swayback, Green) low-MBC control, Yellow) low-MBC with abnormal back structure. ....	38
Figure 2.3 Back conformation of the horses 3542 (left) with MBC of 4.5 and 3520 (right) with MBC of 5.5. ....	39
Figure 3.1 Straight Back Length and Contour Back Length in Measurement of Back Contour (= Contour Back Length – Straight Back Length).....	47
Figure 3.2 Biological consequence of the 9,691 variants identified in the target region of chr20:41M-44M.....	50
Figure 3.3 Courtesy picture from Kraut et al. (1998), showing defects in formation of ribs and thoracic vertebral bones in mice knocked out for the <i>MDFI</i> gene (-/- being homozygous knockout, +/- being heterozygous). Most notable are the sections D and E of the figure, which show that in <i>MDFI</i> mutant newborns, there are abnormal fusions of spinous processes. The blue staining of spinous processes in the mutant, which is shown on the right (-/-), do not merge together medially as indicated by the white arrow. However, the spinous processes can still fuse in a cranial-caudal direction, as demonstrated by the black arrow. ....	60

Figure 3.4 An IGV screenshot of the local realignment of sequencing reads by GATK Haplotype Caller around the 7bp frameshift deletion of the *MDFI* gene in the horse 3519. .... 62

Figure 3.5 IGV screenshot of FAANG RNA-Seq data (bottom panel) and DNA sequence data around the frameshift deletion from the horse 3603 (top panel). .... 64

Figure 3.6 RNA-Seq on the equine Sesamoid bones and Muscle Tissue in the IGV around the 215 bp deletion of the *TRERFI* gene. The bottom track marks the location of the deletion in the horse 3519 who is homozygous for the deletion. .... 66

Figure 4.1 Sampling locations for Iranian populations. Green and red points signify sampling locations for Kurdish and Persian Arabian horses, respectively. The regions circled in green and red identify the original homeland of Kurdish and Persian Arabian horses, respectively. All the point locations outside the ovals represent horses that were descendants of, or were themselves imported from the original homelands. Map imported from the USGS National Map open resources. .... 73

Figure 4.2 Plot of principal components 1 versus 2 for the 134 horse representing 3 breeds. .... 77

Figure 4.3 Violin plot of mean, quartiles and the frequency (the width of the plot) of the ROH-based inbreeding coefficient ( $F_{ROH}$ ) for each breed group. .... 80

Figure 4.4 Bar plot of the K=2 results. The green color designates cluster 1 (which mostly harbored Kurdish and Persian Arabian horses) and the blue color signifies cluster 2 (which mostly contained Thoroughbred horses). Each individual is represented by a single vertical line broken into K colored segments, with lengths proportional to each of the K inferred clusters. .... 82

Figure 4.5 Bar plot of the K=3 results. The blue color designates cluster 3 (which mostly harbored Thoroughbred horses), the green color signifies cluster 2 (which mostly contained Kurdish horses and covered a part of Persian Arabian’s genome), and the red color represents cluster 1, which is attributable to Persian Arabian. Each individual is represented by a single vertical line broken into K colored segments, with lengths proportional to each of the K inferred clusters. .... 83

LIST OF ADDITIONAL FILES

Supplemental Table: Variant genotypes of the whole genome sequenced animals for the target region chr20:41M-44M ..... EXCEL 584 KB



## **CHAPTER 1. INTRODUCTION AND BACKGROUND INFORMATION**

Genomic tools have become available for horses during the last two decades and have been applied to identify genetic variation among individuals and even to compare variation between populations (Raudsepp, Finno, Bellone, & Petersen, 2019). In this dissertation, genomic tools are applied to characterize genetic variation for juvenile onset lordosis among individuals in a population (American Saddlebreds) and to compare relationships among populations (Comparison of Persian horse populations).

### **1.1 Juvenile Onset Lordosis in American Saddlebred Horses**

Swayback, also called lowback, softback and extreme lordosis, is particularly prevalent among Saddlebred horses. Among Saddlebred horses, the condition may appear within the first two years following foaling and has been described as Juvenile Onset Lordosis (JOL) (Rooney & Prickett, 1967). Lordosis is defined as the concave curvature of the spines, which is a normal feature in most mammalian vertebrates (D. G. Cook, 2014). However, when the spine extends in curvature and deviates from the normal structure, it can become a condition known as extreme lordosis. In horses, this phenomenon is manifest as a drop in topline behind the withers and has been identified as a pathologic condition in some cases (Rooney & Prickett, 1967). Figure 1.1 illustrates differences between the back curvature of a JOL horse and a normal horse.



Figure 1.1 Comparison of a horse affected with extreme lordosis (left) to a normal-back horse (right).

There are two types of extreme lordosis with regards to when in life it happens; the first is geriatric lordosis, which is observed in senior horses, and most likely occurs as the result of the aging process (Rooney & Prickett, 1967). The other type is Juvenile Onset Lordosis which usually happens in the first two years of life and is tied to a hereditary condition (D. G. Cook, 2014).

JOL in horses has some phenotypic similarity to Congenital Scoliosis or Kyphosis in humans, which is a deformity that arises from failure of the formation of vertebral bodies (Williams, McCall, O'Brien, & Park, 1982). McMaster and Singh identified three types of Kyphosis to be most common in humans and also appearing in the childhood ages, around 6 years old (McMaster & Singh, 1999). In their study, most of the cases were attributable to failure in formation of vertebral bodies. Shahcheraghi and Hobbi found hemivertebrae to be the most severe and progressive pattern of deformity in development of scoliosis (Shahcheraghi & Hobbi, 1999). Although these studies provide insights about the malformation of vertebral bodies underlying development of kyphosis in humans, comparability of lordosis in humans with horses remains questionable. This is particularly true because of the key difference in the quadruped posture of horses (horizontal) as

compared to biped posture in humans (vertical), and corresponding differences in both the anatomical relationships and biomechanical forces of axial skeletal elements and affiliated structures.

Rooney and Prickett were first to investigate the pathologic anatomy of congenital lordosis in horses, however, their study did not include any Saddlebred horses, but two yearling fillies, a Thoroughbred and a mix-bred. They observed hypoplasia (underdevelopment) of articular processes in thoracic vertebrae at T5-T10 bones (Figure 1.2) (Rooney & Prickett, 1967). They did not notice any other malformations in the vertebrae or any other organs. Also, their microscopic evaluations of the hypoplastic articular processes of the affected bone found no difference with the normal one.

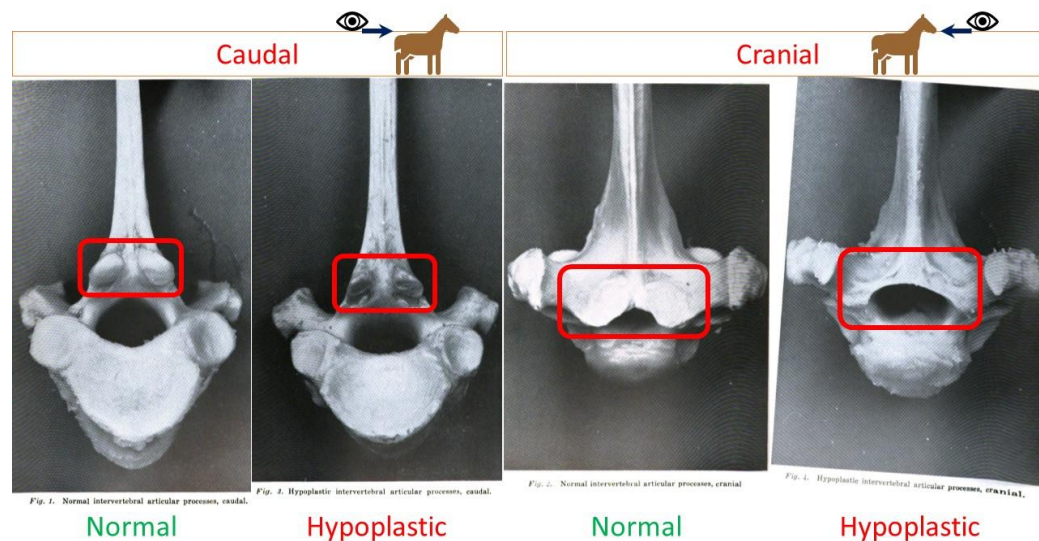


Figure 1.2 Images taken by Rooney and Prickett comparing the articular processes in thoracic vertebrae in lordotic versus normal horses (Rooney & Prickett, 1967).

In agreement with Rooney and Prickett, Coates and McFee also observed hypoplastic articular processes of the vertebral bones in Haflinger foals affected with congenital lordosis (Coates & McFee, 1993).

Later on, an X-ray imaging analysis reported by Gallagher et al. on a 12-year-old lordotic horse found “vertebral bodies of the T5-7 bones to be wedge-shaped and shorter on the ventral side than on the dorsal side. The dorsal spinous processes showed some evidence of impingement in the region T13 to T16 (Gallagher, Morrison, Bernoco, & Bailey, 2003). However, since this was observed in a 12 year old gelding they noted that it was not possible to determine whether the anomaly was cause or effect.

### **1.1.1 Genetic Studies of Extreme Lordosis Among Saddlebred Horses**

Gallagher et al. (2003) found a higher prevalence of extreme lordosis in American Saddlebred horses (five percent of 394 horses) (Gallagher et al., 2003) as compared to other equine breeds (1%) (Rooney & Prickett, 1967). For their study, they devised a method for objective measurement of the trait, Measurement of Back Contour (MBC). It is based on the distance between two landmarks on the horse’s back: point of the withers and point of the hip. The distance between these two landmarks is measured once contour (letting the measuring tape follow the actual contour of the horse’s topline) and is called “contour back length” and then the direct straight line length between the two points is measured as “straight back length”. The MBC is then calculated by simple subtraction of the straight back length from the contour back length:

$$\text{MBC} = \text{Contour Back Length} - \text{Straight Back Length}$$

Figure 1.3 depicts the reference points in measurement of contour and straight back lengths.

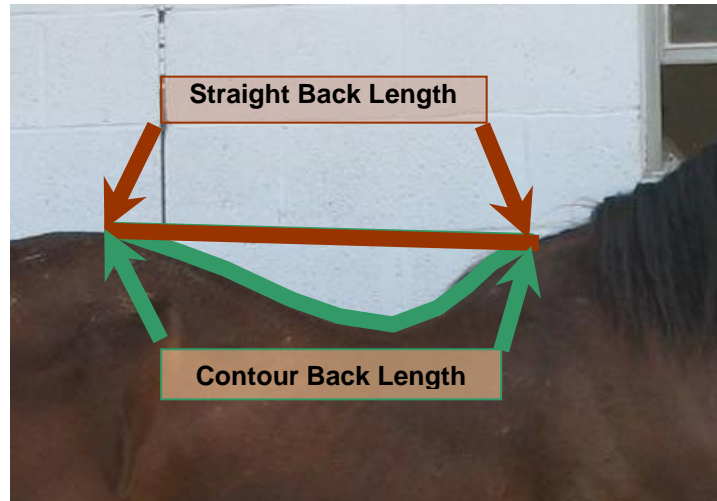


Figure 1.3 Visual representation of MBC measurement method.

The MBC measurement on 305 American Saddlebred (ASB) horses found a range of variation of 1 to 14 cm, and observed a bimodal distribution of the trait, interpretable as the sign of segregation of normal-backed horses from swaybacks. While normal horses had an average of 4 cm MBC, the lordotic horses averaged at 10 cm. Based on these observations, an MBC of 7 cm was designated as the cutoff for a trait designated High-MBC, representing horses with extreme lordosis (Figure 1.4).

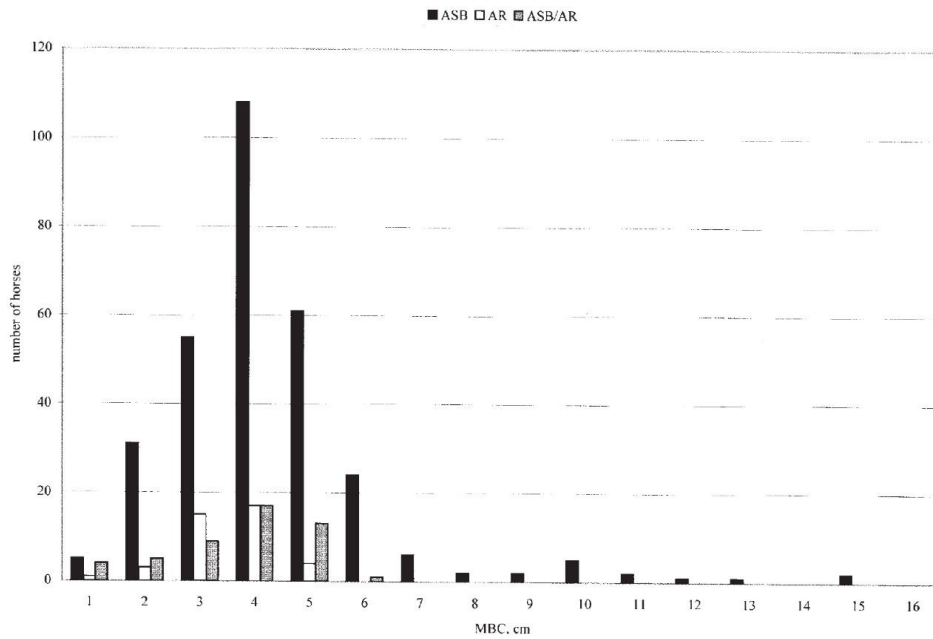


Figure 1.4 Distribution of MBC measurements in the population of 305 ASB horses (black bars). White bars represent Arabian horses and grey bars designate Saddlebred-Arabian crosses (Gallagher et al., 2003).

Gallagher et al. (2003) inferred a hereditary component for high-MBC in ASB horses, with a mode of inheritance to be recessive. Although their study did not prove the recessive mode of inheritance, their suggestion was based on the proportion of the affected horses in the ASB population, 5% with a calculated carrier rate of 25% as well as the observation that crosses of ASB and Arabian horses never produced any affected foals. This conclusion was also consistent with the observations of breeders that horses without JOL could produce offspring with JOL.

Following Gallagher’s study, Cook et al conducted a Genome-Wide Association Study (GWAS) to find the hereditary aspect of JOL in ASB horses on the genome (D. Cook, Gallagher, & Bailey, 2010; D. G. Cook, 2014). Equine 50K SNP genotyping array was used to compare the genome of 20 horses with an MBC of 5 cm or lower versus 20 horses of 7 cm or higher in MBC as the affected group. Their study found a peak of

association with high-MBC on horse chromosome 20. Figure 5 shows the Manhattan plot of association p-values versus genomic location in the study by Cook et al.

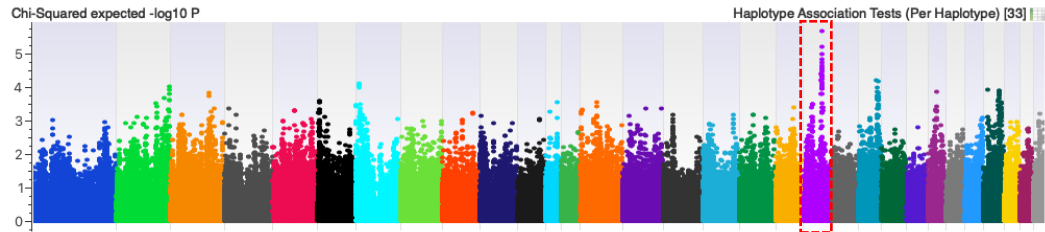


Figure 1.5 Manhattan plot of association p-values versus genomic location in the study by Cook et al. The peak of association on chromosome 20 has been encircled with a red-dotted rectangle (D. Cook et al., 2010).

Next, they designed a denser array of 35 custom SNPs to verify their GWAS results (by testing a larger number of horses, adding 13 affected and 166 unaffected horses). This subsequent study found SNPs spanning a 1.1 Mega-base region on horse chromosome 20 to be associated with lordosis. This haplotype was based on four SNPs at the genomic locations of 42430946, 42504894, 42962032 and 43503716, where coincidence of the nucleotide variants TGTG in those four loci respectively was associated with high-MBC in ASB horses. Exons from several candidate genes in the region, namely TRERF1, RUNX2 and CNPY3 were sequenced, however, a causal variant could not be found.

These results were consistent with Gallagher's earlier model of a recessive mode of inheritance, although not all swayback horses were homozygous for this haplotype and some horses with low MBC possessed the TGTG haplotype (Figure 1.6).

Classification	Number of horses	TGTG homozygotes	TGTG heterozygotes	No TGTG
Swayback	33	23 (70%)	7 (21%)	3 (9%)
Non-affected	287	44 (15%)	135 (47%)	108 (38%)
Combined	320	67 (21%)	142 (44%)	111 (35%)

\*Chi-square for swayback vs. combined = 47.08,  $P < 0.00001$ .

Figure 1.6 Haplotype by high-MBC/low-MBC status in the study by Cook et al (D. Cook et al., 2010).

The conclusion of this study was creation of a hypothesis for the existence of a recessive variant causing juvenile onset lordosis within this region on chromosome 20. Specifically, while the TGTG haplotype seemed to harbor the causal variant, many horses with the TGTG haplotype would not harbor the high-MBC causing variant. Therefore, comparing Whole-Genome Sequencing (WGS) in the region for TGTG-homozygous horses of low versus high MBC could distinguish TGTG haplotypes with and without the causal variant.

### 1.1.2 Genetic Models

Both Gallagher et al (Gallagher et al., 2003) and Cook et al. (D. Cook et al., 2010) proposed an autosomal recessive mode of inheritance for JOL. In this model, which fits their data, horses would be affected if they possessed two alleles for the trait. Horses with one copy would be unaffected. Thus, unaffected parents could have affected offspring. This model is clearly established for many hereditary diseases in horses. However, other traits in horses are thought to be complex, involving multiple loci, management factors and, possibly, incomplete penetrance.

### 1.1.3 Example: Genetics of height in horses

A classic example of a complex quantitative trait with multigenic nature but a major effect from a single locus is the height at withers in horses. The trait has a variation



spectrum of 0.74 to 2 meters in different horse breeds, with a median of 1.6m (Brooks et al., 2010). The first genome-wide association study to map the quantitative trait loci implicated with size traits was conducted by Makvandi-Nejad et al, where they found four loci to capture 83% of the body size variation in the horse (Makvandi-Nejad et al., 2012). These loci involved four genes, namely *LCORL*, *HMGA2*, *ZFAT* and *LASPI*. Of these, *LCORL*, *HMGA2* and *ZFAT* had been known to be involved in controlling human height (Gudbjartsson et al., 2008; Kim et al., 2010; Weedon et al., 2008). Plus, *LCORL* and its neighboring gene *NCAPG* are implicated in cattle growth (Eberlein et al., 2009) and *HMGA2* has a known association with dog size (Boyko et al., 2010; Jones et al., 2008). Soon after this study, subsequent studies found the strongest association of height at the withers with a SNP variant upstream of the *LCORL* (SNP ID: BIEC2-808543, EquCab3.0 chr3:107374136T>C) in several horse breeds, including Thoroughbred (Boyko et al., 2014; Tozaki et al., 2016), Hanoverian (Metzger, Schrimpf, Philipp, & Distl, 2013), Franches-Montagnes (Signer-Hasler et al., 2012), German Warmblood (Tetens, Widmann, Kühn, & Thaller, 2013), Yili horse (He, Zhang, Li, & Liu, 2015) and Persian horse breeds (Mostafavi et al., 2019). Although all studies agree on a strong major effect from the *LCORL* locus, it is obvious that this gene alone cannot explain the entire variation of the height at withers in the horse. For example, a study by Tozaki et al (Tozaki et al., 2016) on trained Thoroughbred horses showed that the C/T genotype will confer a maximum of 1.8-2.1 cm increase in withers height as compared to the T/T. Furthermore, the rare incidence of the C/C genotype in the Thoroughbred population prohibited them to draw any conclusions about the phenotypic effect of the C/C genotype. Taken together, the above-described studies indicate that even a strong major effect from a single locus on a

complex quantitative trait will require co-action of other modifier loci elsewhere in the genome to explain the entire spectrum of variation in the trait, even if the effect by residual loci is infinitesimal, being masked by the major locus. Similar genetic relationships could be applicable to JOL in Saddlebreds, where a major effect from the chr20 factor is in place, but still residual effect by other loci elsewhere in the genome make the phenotypic outcome a complex trait. As a result, cases of contradiction between the genotype and expected phenotype may occur in the population, which could be explained by the putative modifier loci in play.

The presence of major-impact loci in complex quantitative traits in horses could be due to the influence of selective breeding in contrast to randomly-mating populations like humans, loci with major effect in similar traits are less abundant. As noted above, four genes were identified explaining 83% of the variation in height of horses, while it has been estimated that 697 genes, if found, would explain only 15.7% of human height variation (Lango Allen et al., 2010).

## **1.2 Use of Molecular Genetics to Infer Relationships Among Horse Populations**

### **1.2.1 Persian Breeds of Horse**

There are 4 distinct populations of horses in Iran: Turkoman, Caspian, Persian Arabian horse and Kurdish horse.

#### **1.2.1.1 Turkoman**

The homeland and scatter zone of this breed is Northeast of Iran. Turkoman horse is also divided into two subpopulations: Persian Akhal-Teke (Figure 1.7) and Yamut

(Figure 1.8). Note that there is another breed of horse, geographically neighboring to Turkoman, which is considered as Turkmenistani Akhal-Teke. There is scientific evidence that the Turkoman horse has contributed ancestry to modern Thoroughbred horses (Petersen et al., 2013; Wallner et al., 2017).



Figure 1.7 Akhal-Teke sub-breed of the Turkoman horse.



Figure 1.8 Yamut sub-breed of the Turkoman horse.

#### **1.2.1.2 Caspian**

Caspian horse (Figure 1.9) is native to the north of Iran, around the south coast of the Caspian Sea. One of the rarest miniature horse populations, the breed has similar proportionate body structures as large horses. This unique morphology as well as their cooperative temperament makes them an ideal horse for training children with horse riding. A world registry exists for this breed, International Caspian Society.



Figure 1.9 Caspian horse.

### **1.2.1.3 Persian Arabian**

This breed (Figure 1.10) originates from Khuzestan province in the Southwest of Iran. Subpopulations of Persian Arabian horse have been listed as Koheilan, Obayyan, Hamdani, Saglawi and Edban. It has been suggested in the scientific literature that Persian Arabians are a part of the Middle Eastern population of the Arabian horse, is considered as the origin of the global population of Arabian horses (Cosgrove et al., 2020). The World Arabian Horse Organization (WAHO) officially recognizes the Persian Arabian horse as one of the strains of the Arabian breed, registering them under the name of Asil horse of Khuzestan.



Figure 1.10 Persian Arabian horse.

#### **1.2.1.4 Kurdish**

Kurdish horses (Figures 1.11, 1.12 and 1.13) are native to the west of Iran. As this region of the country is mountainous with a cold climate, natural selection has carved the physical characteristics of this breed to be compatible with the environment of their homeland. A standardized description of the physical characteristics of this breed has been developed in a research study by YousefiMashouf et al. (2020) (Yousefi Mashouf, Mehrabani Yeganeh, Nejati Javaremi, Maloufi, & Technology, 2020).



Figure 1.11 Kurdish horse



Figure 1.12 Kurdish horse.



Figure 1.13 Kurdish horse

Kurdish horses are believed to be descendant of an extinct ancient horse, namely Neisayee, which occupied the current geographical range of the Persian Kurdish horse population. Fages et al. (2019) conducted the most comprehensive study on horse domestication by whole genome sequencing of ancient DNA obtained from equine fossils (Fages et al., 2019). Their study identified the Sassanid horse as the origin of the modern horse populations across the world. Interestingly, Sassanid horse remains were discovered in west of Iran, the original homeland of Kurdish horses.

## **1.2.2 Molecular Genetic Studies comparing horse populations**

### **1.2.2.1 Studies with Blood Groups and Biochemical Markers**

In the mid-1900s genetic variation in horses was studied using blood groups and biochemical markers. These tests identified discrete genetic factors with a Mendelian



pattern of inheritance which could be used for parentage testing and comparing populations and breeds of horses (Bowling & Ruvinsky, 2000).

A major difficulty when using blood typing methods was the fragility of blood. Tests usually required fresh blood, preferably less than a week old, and long-term storage of samples for use in future test was only possible for serum or plasma. Furthermore, after 60 years genetic variation had been identified for less than 100 loci. With the advent of DNA testing, it became possible to identify dramatically more genetic variation.

#### **1.2.2.2 Genetic Characterization Studies Using Microsatellite Markers**

Perhaps the most common tool used in molecular genetic characterization of equine breeds has been microsatellite markers. Ellegren et al. reported the first use of microsatellite markers in horses (Ellegren, Johansson, Sandberg, & Andersson, 1992). One of the first studies which paved the road to use microsatellite markers in genetic diversity and characterization of breeds, was the publication by Behara et al. (1998). The aim of that study was to evaluate the potency of microsatellite markers in studies of genetic relationships among equine populations. Eleven microsatellite markers in 903 horses from 11 breeds were analyzed to study the genetic relationship among populations as well as to compare heterozygosity between common versus rare/endangered breeds. Initial analysis on the effect of sample size on clustering patterns showed that resolution of clustering in extremely isolated breeds as well as closely-related breeds is less dependent on the distance size and sample quantity; such that in 98% of cases only 30 samples would be sufficient to properly assign horses of isolated breeds into their respective genetic clusters. Therefore, the outcomes of this study advocated usefulness of microsatellite markers for genetic distance and diversity studies (Behara, Colling, Cothran, & Gibson, 1998).

Following that, the potential capability of standard sets of microsatellite markers was confirmed in a study of genetic relationships among riding breeds (Arabian and Hanoverian), primitive breeds (Exmoor and Soraia) and six German draft breeds (Aberle, Hamann, Drögemüller, & Distl, 2004). Likewise, in a study by Bigi et al. (2007), using only 12 microsatellite markers was sufficient to show significant clustering of Thoroughbred and Anglo-Arabian (Thoroughbred-Arabian cross) as opposed to Haflinger, Bodaglino and Italian heavy draft breeds (Bigi, Zambonelli, Perrotta, & Blasi, 2010). Luis et al (2007) combined data from 17 protein markers and 12 microsatellite markers to distinguish 8 breed groups among 33 breeds, where clustering of four groups was well-justified: 1) Andalusian and Lusitano; 2) Friesian and two pony breeds; 3) Morgan, Standardbred, Rocky Mountain and the American Saddlebred; 4) Thoroughbred, Quarter Horse, Hanoverian, Holstein and Irish Draft (Luis, Juras, Oom, & Cothran, 2007).

Microsatellite markers have also been used to investigate the origin of certain breeds. For example, Kakoi et al. (2007) found evidence of origin from Mongolian pony to the Japanese breeds (Kakoi, Tozaki, & Gawahara, 2007). Also, data from 26 microsatellite loci provided weak evidence of an ancestral relationship between Mongolian pony and Norwegian breeds, although making sense with morphological traits of Norwegian breeds (Bjørnstad, Nilsen, & Røed, 2003).

Microsatellite markers have been employed in some studies for breed assignment purposes. In a study by Canon et al (2000), 481 horses randomly selected from Spanish breeds as well as 60 Thoroughbred horses were genotyped for 13 microsatellite loci. According to their results, only 8% of the genetic variation was assigned to the among-population. Nevertheless, the markers used in this study were able to distinguish horses

into defined breed groups. Their study advocated the usefulness of microsatellite markers for breed assignment purposes, although the precision rate in breed assignments varied for different breeds (Canon et al., 2000).

### **1.2.2.3 Population Genetic Studies Using Single Nucleotide Polymorphism Markers**

Single Nucleotide Polymorphism (SNP) markers became the most common genetic polymorphism used in genetic studies of horses after the completion of the whole genome sequence (McCue et al., 2012; Wade et al., 2009). After development of the commercial Equine SNP50 Genotyping Array (Illumina Inc, San Diego, CA) which simultaneously genotypes for 54,602 SNPs in horses, McCue et al performed a series of analyses to evaluate the efficiency of the SNP chip. One of these analyses was the study of inbreeding, genetic distance and relationship among breeds. Their analysis showed that between 43,287 to 52,085 of those markers are capable of demonstrating the individual differences within breeds. Also in this study the relationship and distinction among 14 breeds was evaluated using the SNP chip which showed its high potency in distinguishing and identifying the genetic relationship among horse breeds. According to their results from the Equine SNP50 array, individuals of each breed formed a distinctive group from other breeds (McCue et al., 2012).

Perhaps the most comprehensive population genetics study in horses was conducted by Petersen et al (2013) who also utilized the Equine SNP50 array in their analyses. The study which aimed to investigate genetic diversity within and among horse breeds, included 814 horses from 36 breeds. Genotype data was obtained from the Equine SNP50 array, which after pruning the 54,602 SNPs, a total of 10,536 SNPs remained for analysis, which included  $F_{ST}$  and genetic distance calculations as well as Parsimony Analysis. The

use of these 10,536 markers resulted in clear distinction of horses into defined breed groups and accurate formation of breed relationship diagrams. Out of the 814 horses studied, only 7 horses were mis-allocated to a place other than their actual breed group (Petersen et al., 2013).

#### **1.2.2.4 Population Genetics Studies on Persian Horse Breeds**

There have been only a few scientific studies designed to characterize genetic relationships among Persian horse populations. An example of them is a study on the Caspian Horse by Shahsavarani and Rahimi-Mianji (2012). The study aimed at assess the genetic variation in the population of Caspian horses in Iran as well as looking for evidence of bottleneck events. One hundred horses from 5 locations were sampled and genotyped for 16 Microsatellite loci. Observed heterozygosity was 0.52, which was lower than the expected heterozygosity (0.82). Also, there were multiple cases of significant deviations from the Hardy-Weinburg equilibrium. Wright's Fixation Index, known as  $F_{IS}$  was 0.367 which was indicative of lowered heterozygosity. Ultimately, their results did not find evidence for a bottleneck event in the recent history of Caspian horse population (Shahsavarani & Rahimi-Mianji, 2012).

Other population genetic studies in different countries have been conducted that included Persian samples, although those Persian breeds have not been the focus of their study. As an example, in the study by Petersen et al. (2013) on the genetic diversity of the modern horse populations, samples from Caspian, Arabian and Akhal-Teke (one of the sub-populations of the Turkoman) are seen within the list of 36 breeds studied (Petersen et al., 2013). Another example would be the study by Khanshour et al. (2013) on the various populations of Arabian horse, including 682 horses from 7 different geographical

populations. Out of those seven populations, three belonged to Middle Eastern region (252 Syrian Arabian, 33 Saudi Arabian and 40 Persian Arabian), two from Europe (21 Shagya-Arabian and 36 Polish Arabian) and one population was from the America (155 American Arabian). Genotyping on 15 microsatellite loci was performed to calculate genetic distance, Analysis of Molecular Variance (AMOVA), Factorial Correspondence Analysis. Their overall results showed that genetic diversity in Middle Eastern horses was higher than that of Western populations. Also, the AMOVA showed that eastern Arabian populations are the primary source of variation in the studied populations. Genetic divergence among the Middle Eastern populations was not evident, but clear divergence of American populations from the Middle Eastern was found, and that the American population was genetically more uniform (Khanshour, Conant, Juras, & Cothran, 2013).

Sadeghi et al (2019) conducted a scientific study with a focus on the Persian Arabian horse, to explore genetic diversity, identify signatures of selection, as well as study their relationship with other Persian horse populations. Their study used the Equine SNP array which genotypes for 670,000 markers simultaneously (Axiom Equine Genotyping Array, ThermoFisher Scientific). Observed heterozygosity was 0.43, comparable to the expected 0.45. They observed low measures of average inbreeding, which indicated a high genetic diversity in Persian Arabian horses. Their data suggested that Persian Arabian horses can be divided into 3 groups. Using methods of Tajima's D, H, and H12, they found 15 genomic regions as potentially under selection in Persian Arabian horses. They included SNP genotypes from 30 horses of 4 other Persian breeds, which indicated a distinct population structure among Persian Arabian, and Turkoman and Caspian horse breeds (Sadeghi et al., 2019).

Genetic diversity of Persian Arabian horse has also been studied in a broader research study, which looked at the genetic structure of the Arabian horse populations across the globe, using the newer Equine 670K SNP Array and whole-genome sequence data. Like other studies, this research also confirmed a high level of genetic diversity in the Middle Eastern populations, including horses coming from Iran. Their data overall support the Middle East as the origin of the Arabian horse, without evidence of reduced genetic diversity across global population of the breed (Cosgrove et al., 2020).

## **CHAPTER 2. FINE MAPPING STUDIES OF HORSE CHROMOSOME 20 REGION ASSOCIATED WITH JUVENILE ONSET LORDOSIS IN AMERICAN SADDLEBRED HORSES**

### **2.1 Summary**

Previous studies reported an association of Juvenile Onset Lordosis (JOL), as determined by Measurement of Back Contour (MBC), with a region on horse chromosome 20 (chr20:42,430,946-43,503,716). Their data suggested a single recessive variant on chr20 to be associated with the trait. The present research aimed to find the causal mutation on chr20 using Whole-Genome Sequencing (WGS), testing the hypothesis that a single recessive variant on chr20 causes high-MBC. Eleven Saddlebreds were Whole-Genome Sequenced for two experiments. For one experiment, WGS variants from 3 high-MBC horses and 3 low-MBC horses were compared in the region chr20:41M-44M and filtered for variants characteristic of a recessive mode of inheritance. None were found to support the hypothesis for a recessive mode of inheritance. In another experiment, WGS variants were compared for a full-sib family including sire, dam, and three offspring where two of the offspring had high-MBC values. Again, no variants were found supporting presence of a variant in this region with recessive mode of inheritance for the High-MBC. These results together reject the hypothesis, suggesting more than one major variant to be involved in swayback. To determine whether or not the original association of high-MBC with chromosome 20 was a statistical artifact, additional markers were tested in 109 horses, 34 high-MBC and 75 low-MBC. Re-evaluation of the association on chr20 was performed via genotyping for tag markers on chr20 haplotype, where a chi-square comparison confirmed that the association of chr20 with high-MBC. In addition, a subjective and qualitative evaluation of the back conformation of two of the low-MBC horses led to the conclusion

that having an MBC of less than 7.0 centimeters does not necessarily mean that the horse is unaffected/normal. In other words, while the high-MBC designation is effective in identifying affected horses, low-MBC does not mean they are unaffected.

Keywords: American Saddlebred Horse, Swayback, Lordosis, Whole Genome Sequencing.

## **2.2 Introduction**

Juvenile Onset Lordosis (JOL), also known as swayback, soft-back, low-back, or extreme lordosis, is a conformation defect in which the back curvature is exaggeratively deviated from the normal structure. It is distinguished from geriatric lordosis in that it develops within the first two years of life and is tied to hereditary condition, whereas lordosis that occurs in senior horses is believed to be primarily a consequence of aging. Cook et al. mention an overrepresentation of the phenotype in the American Saddlebred horses (5%) compared to other equine breeds (1%) (D. G. Cook, 2014). Gallagher et al (2003) developed an objective measure of back conformation called, Measurement of Back Contour (MBC) and used it on a population of Saddlebred horses. The population had a bimodal distribution with two peaks at 4cm (indicative of low-MBC non-affected) and 10cm (representative of high-MBC swayback), and a measurement of 7cm or greater was designated high-MBC. Their results also suggested a recessive hereditary factor to be responsible for the trait (Gallagher et al., 2003). Following that study, Cook et al. (2010) used MBC measurements to compare the genome of Saddlebreds with high-MBC (>8.0cm) as case versus low-MBC controls (<5.0cm) in a genome wide association study (GWAS). Their study identified a region on chromosome 20 associated with the trait (D. Cook et al.,



2010). A haplotype defined by four SNPs spanning chr20:42.4-43.5M were found to be most strongly associated with the occurrence of High MBC and referred to as “TGTG”. Their results corroborated a recessive mode of inheritance for the studied trait. However, the association with the TGTG was not complete. While most horses with high-MBC were homozygous for the haplotype, the low MBC horses include homozygotes (TGTG/TGTG) and heterozygotes (TGTG/other). This suggested a recessive mode of inheritance for the trait and that some TGTG haplotypes included a variant responsible for the trait while others did not. To find the variant responsible, it would be necessary to conduct fine mapping of the region using whole genome sequencing (WGS). At the time, costs of WGS were prohibitive for the study. Since then, the cost of WGS has reduced to the point where it can be used as a screening tool.

The present study was aimed to take advantage of this technology in fine-mapping the Chromosome 20 haplotype suggested by Cook et al (2010) and find the causal mutation affecting lordosis in Saddlebred horses. Our study considered the recessive mode of inheritance as the foundation of the research, being developed over the following hypothesis:

*“There is a single recessive variant on chromosome 20 causing high-MBC in the American Saddlebred horses”*

## **2.3 Materials and Methods**

### **2.3.1 Experimental Animals for Whole-Genome Sequencing**

Eleven registered American Saddlebred horses were selected for Whole Genome Sequencing. The horses were privately owned and maintained on farms in Kentucky, Ohio,

Louisiana, and Wisconsin. Six horses were selected based on phenotypic measures of MBC plus genotypes for the 4 SNPs used to define the TGTG haplotype from the study by Cook et al. (D. Cook et al., 2010). Five horses were selected as part of a core family segregating for low MBC. IACUC 2019-3247 was approved in connection with study of these horses.

### 2.3.2 Experimental Animals for Population Genotyping

Archived DNA samples of 109 American Saddlebred horses with MBC measurements and TGTG genotypes were available from the study of Cook et al. as well as newly collected samples. Of these 27 high-MBC and 73 low-MBC belonged to the association study of Cook et al. (2010) and 7 high-MBC and 2 low-MBC were newly collected samples. Table 2.1 identifies the MBC phenotypes as well as the zygosity for TGTG haplotype of the studied horses.

Table 2.1 MBC phenotype and zygosity of the TGTG haplotype in the horses included in the association validation study. “other” represents non-TGTG haplotypes at these sites, e.g., CACT, CATT, CACG, etc.

	TGTG/TGTG	TGTG/other	other/other	Sum
Low-MBC	24	15	36	75
High-MBC	25	8	1	34

### 2.3.3 Phenotyping

Measurement of Back Contour (MBC) was conducted as described previously (Gallagher et al., 2003). MBC is based on the distance between point of the withers and point of the hip. Straight distance between the two points is subtracted from the contour distance and the result will be the MBC. For 9 of the horses the author measured the horses while for one the measurement was made by owners. One horse (number 3603) was not

measured but was visibly extremely affected by JOL. For horses in the archive, their MBC were previously measured by Gallagher et al. (Gallagher et al., 2003)

IACUC 2019-3247 was approved in connection with study of all experimental animals.

#### **2.3.4 DNA Isolation**

Whole blood was drawn from each horse into blood collection vacutainer tubes containing EDTA as an anti-coagulant. Blood samples were refrigerated until DNA extraction. DNA isolation was performed using Genra Puregene Blood Kit by Qiagen Inc. following the manufacturer's protocols. DNA quantity and quality was measured using a NanoDrop 2000 instrument before submitting the samples to for Whole-Genome Sequencing.

#### **2.3.5 Whole Genome Sequencing**

Samples from each horse were submitted to Psomagen Inc (Rockville, MD) to conduct Whole Genome Sequencing. Basically, a minimum of 500 ng of DNA with DNA Integrity Number (DIN) of 7.0 or higher was provided from each horse to pass the DNA quality requirement, which was measured using a Bioanalyzer instrument at the company. Library preparation was performed using the Illumina TruSeq DNA PCR-Free kit. Paired-end short-read sequencing (read size of 150 bp) was executed using an Illumina NovaSeq instrument to generate at least 50 Giga base of sequence data per horse. This was expected to yield an average of 20X of sequencing coverage, given the reference genome size of 2.5 Gb in the horse, based on EquCab3.0 reference assembly. Sequence information was provided as raw FASTQ files.

### **2.3.6 Analysis of Whole Genome Sequence**

Sequencing adapters were trimmed using TrimGalore software (Krueger, 2012). BWA aligner was used to map a total of 4,630,641,790 reads to the equine reference genome assembly (EquCab 3.0) (Kalbfleisch et al., 2018; H. Li & Durbin, 2010; H. J. a. p. a. Li, 2013). Genome Analysis Toolkit (GATK) was used to call the variants and genotype them to generate the final VCF file (McKenna et al., 2010).

### **2.3.7 Genomic Region under Study**

The Haplotype (TGTG) region that Cook et al found to be the most discriminative between the affected vs non-affected cohorts, spans chr20:42.4-43.5M (D. Cook et al., 2010). To ensure that our WGS scans do not miss anything within and upstream/downstream of the implicated region, we widened the analyses to chr20:41.0-44.0M. This region was scrutinized both visually using Integrative Genome Viewer (IGV) Software (Robinson et al., 2011), as well as computer programs customized to detect the variants that met our variant filtering criteria described below.

### **2.3.8 Experimental Design**

Two experiments using short read whole genome sequencing were designed and conducted to test the hypothesis of a recessive mode of inheritance of high-MBC. Experiment 1 was a comparison of 3 high-MBC versus 3 low-MBC horses with various haplotype structures on chromosome 20. Experiment 2 was the study of a core family of 5 horses including normal-backed sire and dam, together with their three full-sib offspring where two of them were high-MBC swayback and the other was a normal-backed low-MBC.

**2.3.8.1 Experiment 1: Whole-genome comparison of  
unrelated horses of low versus high MBC**

**2.3.8.1.1 HORSES IN EXPERIMENT 1**

Six ASB horses were selected based on their MBC phenotypes and TGTG haplotypes for WGS (Table 2.2). One horse (3527) had low MBC and no TGTG; one horse (3519) had high MBC and was a TGTG homozygote; two horses (3517 and 3529) had high MBC and were heterozygous for TGTG and another haplotype (CACT); two horses (3520 and 3542) had low MBC but were homozygotes for TGTG. No high-MBC horses without TGTG were available for sequencing.

Table 2.2 Horses with available whole genome sequence data. The red highlight marks the haplotype with highest frequency in the swayback horses in the study by Cook et al. Red color in the last column marks the alleles that were found associated with high-MBC in Cook et al.

Animal ID	Relationship	Category	MBC (cm)	Haplotypes in TGTG sites
3517	Unrelated	High MBC	7.5 in 2000 9.0 in 2019	CACT/TGTG
3519	Unrelated	High MBC	14.0	TGTG/TGTG
3529	Unrelated	High MBC	8.0	CACT/TGTG
3527	Unrelated	Low MBC	3.5	CACT/CATT
3520	Unrelated	Low MBC	5.5	TGTG/TGTG
3542	Unrelated	Low MBC	4.5	TGTG/TGTG

### 2.3.8.1.2 VARIANT FILTERING CRITERIA FOR EXPERIMENT 1

Following the model which focuses on a recessive mode of inheritance for the chr20 region, we looked for variants that were homozygous for the alternate allele in the high-MBC horses (3517, 3519 and 3529) while being heterozygous or homozygous for the reference allele in low-MBC horses (3527, 3520 and 3542).

*Definition of Reference and Alternate alleles:* reference allele refers to the base that is found in the reference genome (Kalbfleisch et al., 2018). Since the reference is derived from an individual's genome, it is not always the major allele. In contrast, the alternate allele refers to any base, other than the reference, that is found at that locus.

### 2.3.8.2 Experiment 2: Whole-genome study of the core family

#### 2.3.8.2.1 HORSES IN EXPERIMENT 2

A core family of Saddlebred horses including normal-backed sire and dam, together with their three full-sib offspring where two of them were high-MBC swayback. Table 2.3 describes the members of the family.

Table 2.3 Specifications of the members of the core family. Same descriptions in the caption of table 2.2 apply to this table.

ID	Relationship	Category	MBC (cm)	Haplotypes in TGTG sites
3604	Sire	Low MBC	4.0	CACG/TGTG
3535	Dam	Low MBC	5.0	TGTG/TGTG
3601	Full-sib sister	Low MBC	3.5	CACG/TGTG

3602	Full-sib brother	High MBC (Swayback)	9.0	TGTG/TGTG
3603	Full-sib sister	High MBC (Swayback)	Not available (identified visually)	TGTG/TGTG

Figure 2.1 depicts the actual pictures of the studied horses with their back conformation in a diagram of familial relationship.

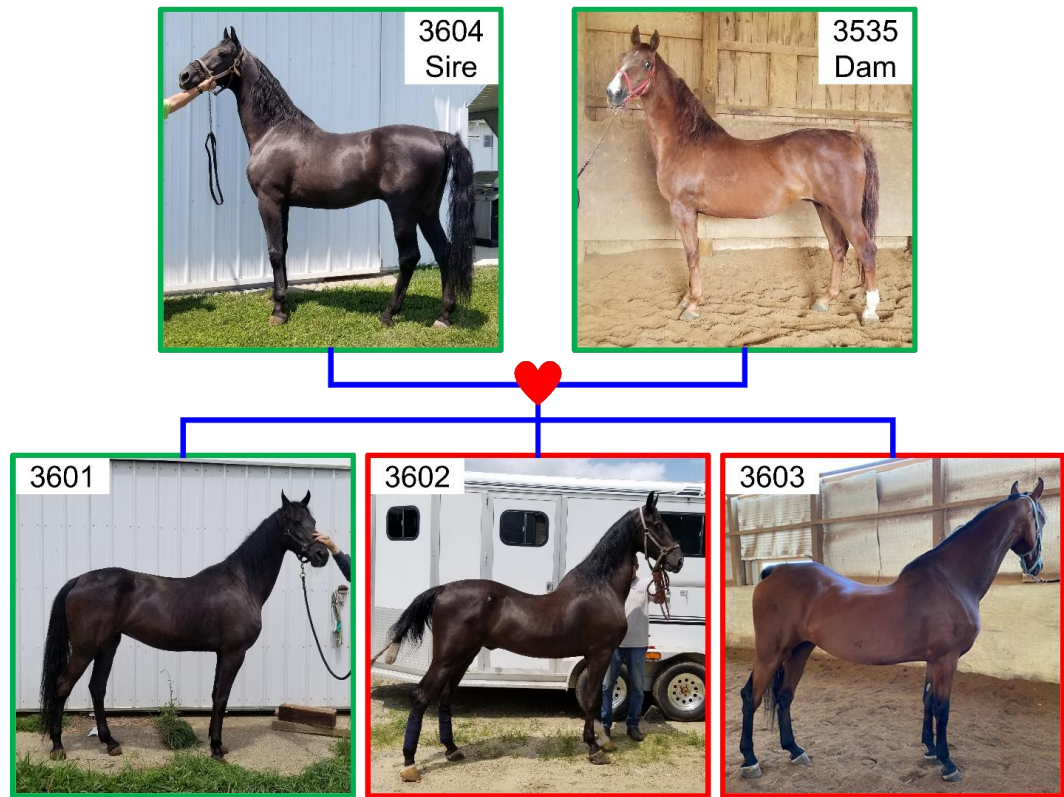


Figure 2.1 Side photos showing back conformation of the horses in the core family study.

### 2.3.8.2.2 VARIANT FILTERING CRITERIA FOR EXPERIMENT 2

For Experiment 2, we looked for variants that were homozygous for the alternate allele in the high-MBC offspring (3602 and 3603), while being heterozygous or

homozygous for the reference allele in the rest of the family members, including sire (3604), dam (3535), and the normal sister (3601).

### **2.3.9 Genetic Marker Genotyping for Re-evaluation of the chr20 Association**

Three candidate variants were identified by WGS analysis and appeared to be common to the haplotype associated with high-MBC. These markers were used for population genotyping to re-examine the association of chr20 with the MBC phenotype.

These three markers include:

- 1- Chr20:42,222,093. Deletion of a 224 bp Equine Repetitive Element 1 (ERE1), also known as SINE or Short Interspersed Nuclear Element located at the 3' Untranslated Region (UTR) of the *TAF8* gene.
- 2- SNP chr20:42,247,262G>A: This is a missense variant in the fourth exon of the gene *C6orf132* which has a calculated SIFT score of zero, interpreted to have a deleterious effect on gene function.
- 3- Chr20:42,399,504. Deletion of a 215 bp segment within the second intron of the *TRERF1* gene. This is the closest gene to the peak of association (chr20:42,430,946) found by Cook et al (D. Cook et al., 2010).

Genotyping method of each of these variants are explained in the following.

#### **2.3.9.1 ERE1 deletion at chr20:42,222,093**

PCR primers were designed for the region flanking the ERE1 element as Forward 5'-TTGATGAGCAGTGCCATGTC-3' and Reverse 5'-CCGTGGCCGAGTGGTAAAGT-3'. The amplicon was investigated by gel electrophoresis to identify the presence or absence of the ERE1.



### **2.3.9.2 SNP chr20:42,247,262G>A**

A custom TaqMan SNP Genotyping Assay by Thermo Fisher Inc. was used as a genotyping tool, with the 3130 Genetic Analyzer (Applied Biosystems Inc.) as the instrument.

### **2.3.9.3 215bp Deletion at chr20:42,399,504**

PCR primers were designed for the region flanking the deletion as Forward 5'-GAAGAAATTACTCAGAGTTTCAGCA-3' and Reverse 5'-AGTGATCTGGCATTCTCTCTG-3'. The amplicon was investigated by gel electrophoresis to identify the presence or absence of the deletion.

### **2.3.10 Analysis of Genotype Distribution and Association Tests**

A chi-square test was conducted comparing genotype distributions of each of the three variants between high versus low MBC Saddlebred horses. An online tool for chi-square calculations was available from the Social Science Statistics website (<https://www.socscistatistics.com/tests/chisquare2/default2.aspx>) which was used for this purpose.

## **2.4 Results**

### **2.4.1 Overall WGS Results**

The coverage of Whole-Genome Sequencing on the 11 horses ranged from 20.8 to 41.2 X with an average of 25.4 X (Table 2.4).

Table 2.4 Summary statistics of the Whole Genome Sequence data generated for each of the horses in the study. Depth of coverage has been calculated by dividing the Total read bases by the total sequence length of the EquCab3.0 reference genome assembly reported by NCBI to be 2,506,966,135 base pairs.

Sample ID	Total Read Bases	Total reads	Depth of Coverage (X)
3517	62,300,730,620	412,587,620	24.9
3519	68,007,072,262	450,377,962	27.1
3529	60,725,922,024	402,158,424	24.2
3527	61,063,568,896	404,394,496	24.4
3520	52,262,754,658	346,110,958	20.8
3542	53,273,128,878	352,802,178	21.3
3604	59,478,823,292	393,899,492	23.7
3535	53,975,832,142	357,455,842	21.5
3601	66,089,044,592	437,675,792	26.4
3602	58,881,960,760	389,946,760	23.5
3603	103,168,072,166	683,232,266	41.2

#### 2.4.2 WGS results for the target region: chr20:41M-44M

The Variant Call Format (VCF) file contained a total of 9,691 variants called and genotyped within the range of chr20:41.0-44.0M in the cohort of all studied animals. Table 2.5 shows the summary data of genotype counts for each horse in the target region chr20:41M-44M.

Table 2.5 Distribution of genotypes for each of the horses in the target region of chr20:41M-44M.

Sample ID	homozygous reference (0_0 in VCF) count	Heterozygous (0_1 in VCF) count	Homozygous alternate (1_1) count
3517	2,797	3,194	3,392
3519	4,452	190	4,781
3529	2,475	4,749	2,172
3527	5,302	3,957	148
3520	4,428	196	4,797
3542	4,402	555	4,479
3535	4,473	142	4,851
3604	3,065	4,031	2,328
3601	3,068	4,051	2,324
3602	4,499	113	4,857
3603	4,512	108	4,866

In both experiments, independently and in aggregate, none of the variants passed the variant filtering criteria governing the hypothesis of single recessive. Table 2.6 shows the breakdown of variants shared as alternate-homozygous among high-MBC horses and compares them to each of the individual control horses in the Experiment 1. Table 2.7 shows the same for horses in Experiment 2.

Table 2.6 Alternate-homozygous variants shared among high-MBC horses and their comparison to individual low-MBC horses in Experiment 1.

Alternate-homozygous variants shared among high-MBC horses	1,617
--	-------

Alternate-homozygous variants shared among high-MBC horses and 3527	95
Alternate-homozygous variants shared among high-MBC horses and 3520	1,609
Alternate-homozygous variants shared among high-MBC horses and 3542	1,553
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in 3527	1,522
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in 3520	8
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in 3542	64
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in low-MBC horses	0

Table 2.7 Alternate-homozygous variants shared among high-MBC horses and their comparison to individual low-MBC horses in Experiment 2.

Alternate-homozygous variants shared among high-MBC horses	4,838
Alternate-homozygous variants shared among high-MBC horses and 3535	4,818
Alternate-homozygous variants shared among high-MBC horses and 3604	2,298
Alternate-homozygous variants shared among high-MBC horses and 3601	2,297
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in 3535	20
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in 3604	2,540
Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in 3601	2,541

Alternate-homozygous variants shared among high-MBC horses, and not alternate-homozygous in low-MBC horses	0
--	---

Figure 2.2 shows the haplotype structure of the target zone chr20:41.0-44.0M on the horses included in both experiments.

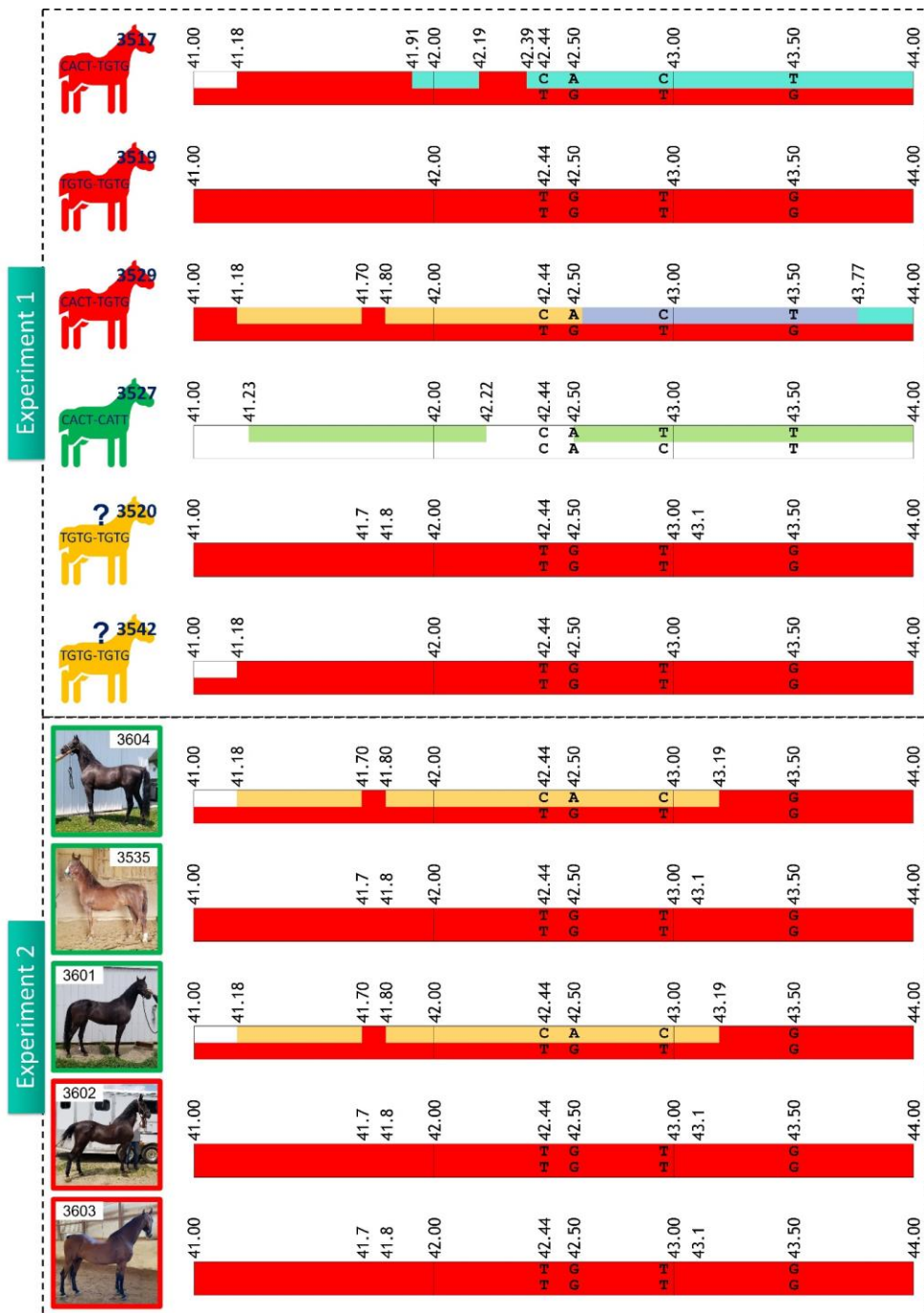


Figure 2.2 Haplotype structure of the target genomic region chr20:41.0-44.0M on the horses studied in both experiments. Red color identifies the haplotype containing the TGTG alleles that are most commonly found in swayback horses. White color marks the reference haplotype found in the reference genome. Other colors identify the haplotypes unique to the horses in our study, which were minor modifications of the reference haplotype. Relative distances are proportional. Body colors or photo outlines of the horses in the diagram of the Experiment 1: Red) high-MBC swayback, Green) low-MBC control, Yellow) low-MBC with abnormal back structure.

As depicted in figure 2.2, the low-MBC dam in the core family (Experiment 2) shared homozygosity of the swayback-associated haplotype with her two affected offspring across the entire target region. As the diagram shows, all the animals in the core family were homozygous for the alternate alleles at chr20:43.19M and retain that configuration until the end of the target region (chr20:44.0M).

### 2.4.3 Observations about the MBC phenotypic measurement

Since the low-MBC horses 3520 and 3542 in Experiment 1 were homozygous for the haplotype associated with High-MBC horses for almost the entire target region, we examined their back curvature more closely. Figure 2.3 shows the conformation of their back. While the MBC measurements for 3520 and 3542 were 4.5 and 5.5, respectively, their conformation might be called lowback, softback or swayback by some horsemen.



Figure 2.3 Back conformation of the horses 3542 (left) with MBC of 4.5 and 3520 (right) with MBC of 5.5.

Treating these horses as affected and comparing 3527 (normal back) to the other 5 horses in Experiment 1 led to identification of 1,455 variants in the target region that match the criteria for a hypothesis of a single recessive variant causing the extreme lordosis. However, this did not change the overall outcome of the analysis aggregating the two

experiments or considering experiment 2 only, meaning that there is no variant passing the single recessive hypothesis. When the family from Experiment 2 was added, considering 3604, 3535 and 3601 as normal backed horses, all 1,455 variants were eliminated.

#### 2.4.4 Population Genotyping of Genetic Markers to Test Significance of Association of chr20 Region

Several variants identified in the previous section were tested in a large population of Saddlebred horses that had been characterized for MBC. These data were used to test whether significance of the association of these variants with the occurrence of high-MBC (Table 2.8).

Table 2.8 Distribution of genotypes for ERE1 deletion at 42,222,093, Missense SNP of *C6orf132* variant at 42,247,262, and 215bp deletion at 42,399,504, and comparison of between high-MBC to low-MBC (Chi-Square test) for each marker. For the ERE1 and 215bp deletion, D corresponds to the deleted (alternate) allele, whereas N signifies the intact (reference) allele. For *C6orf132* SNP, A is the alternate and G is the reference allele. Numbers in parentheses are horses from that group that were not used in the previous GWAS study (Cook et al., 2010), e.g., data is shown that 28 horses possessed the DD genotype and 9 of those were newly sampled for this study.

Variant	Genotype	Number high-MBC	Number low-MBC	P from Chi-Square test
<i>ERE1 Deletion</i> chr20:42,222,093	<i>DD</i>	28 (9)	33	0.000186
	<i>DN</i>	6 (1)	8 (1)	
	<i>NN</i>	0	14 (1)	
SNP chr20:42247262G>A	<i>AA</i>	27 (10)	34	0.000899
	<i>AG</i>	7 (1)	27 (1)	
	<i>GG</i>	0	14 (1)	
<i>215bp Deletion</i>	<i>DD</i>	20	40	



chr20:42,399,504	<i>DN</i>	4	15	0.013
	<i>NN</i>	0	18	

## 2.5 Discussion

### 2.5.1 Results from filtering VCF files using criteria for a recessive mode of inheritance

Horses were selected for sequencing and compared based on their MBC scores. If high-MBC had a recessive mode of inheritance based on a variant in the region chr20:41M-44M, then we would expect to find a variant for which all high-MBC horses were homozygous for the non-reference variant and all the low-MBC horses were either homozygous or heterozygous for the reference allele (never homozygous alternate). No such variants were found. This is evidence against the hypothesis for a single recessive allele in this region causing the high-MBC phenotype.

### 2.5.2 Construction of Haplotypes Across chr20:41M-44M

The VCF files from the 11 horses used for WGS were compared and used to construct haplotypes across the region chr20:41M-44M. Comparison of haplotypes again showed that the genetic influence of this region on high-MBC could not be a simple, Mendelian recessive. Among the 6 horses identified as having high or low MBC, no variants were found in the region chr20:41M-44M that were homozygous for the non-reference haplotype in the high-MBC and either heterozygous or homozygous reference haplotype among the three horses with low-MBC.

Perhaps the most striking evidence for rejection of the hypothesis is the low-MBC dam in Experiment 2 having identical haplotype structure to the swayback offspring which eliminated all possibilities for any variant to meet the criteria. Likewise in Experiment 1, two of the low-MBC horses showed the same haplotype pattern as the dam in Experiment 2, plus mutually exclusive haplotype structure among high-MBC individuals (3517 and 3529) led no common variants to meet the criteria (Figure 2.2).

Together, these observations suggest that, while high-MBC is associated with a genetic factor in this region, the genetic factor does not exhibit a recessive mode of inheritance. Several low-MBC horses were homozygous for a haplotype that appeared commonly among high MBC horses suggesting that other factors contribute to the phenotype. These factors could be epigenetic factors on chromosome 20 or they could involve interaction with genes at other loci (epistatic gene interaction).

### **2.5.3 Population Genotyping to Consider possibility of Statistical Artifact**

Variants were identified in this region and tested on a larger number of Saddlebred horses to determine whether the results from the horses selected for WGS reflected the distribution of variants in the general population. Furthermore, to determine whether the original results were a statistical artifact, the markers were used to compare the association among 109 horses. The results confirmed the association, and the distribution of markers reflected the structure subsequently determined in haplotype analyses.

### **2.5.4 Insights about the Measurement of Back Contour**

Another observation emanating from this study was the potential limits of the MBC measure. The MBC is a quantitative measure ranging from 1.0 to 14.0 centimeters (Gallagher et al., 2003). Based on the population distribution of MBC measurements,

Gallagher et al. (2003) chose an MBC of 7.0 or higher as the definition of affected horses. This threshold for high MBC was a conservative choice in that every horse with such a score would be affected. However, this does not mean that horses with lower MBC (i.e., below 7.0) are not afflicted with JOL. Indeed, when two of the horses identified as low-MBC had genetic profiles shared with the high-MBC horses, their back conformation was re-assessed. Subjectively, their back profiles deviate from normal. Attempts were made to devise new methods to measure the back and capture the unique variation they exhibited. However, none were found to be satisfactory. In any case, the analyses described here are confounded by the uncertainties surrounding the phenotypic characterization. We are unable to reliably identify horses as being free of the trait.

### **2.5.5 Approaches to Identifying Epistatic Genetic Factors**

A hypothesis that follows from these observations is that genetic influences in this region predispose horses to high-MBC, but that genetic effects elsewhere in the genome lead to the extreme lordosis seen as JOL. GWAS could be used to identify other genetic factors. Essentially, horses could be identified as being homozygous for the high-MBC associate haplotype, measured for MBC then compared in a GWAS to see what other chromosome regions might play a role. Since all horses would be homozygous for the region on chr20:41-44, then the only differences would be loci at other locations.

## CHAPTER 3. INVESTIGATION OF THE GENETIC EFFECT OF A CHROMOSOME 20 FACTOR ON DEVELOPMENT OF THE JUVENILE ONSET LORDOSIS IN AMERICAN SADDLEBRED HORSES

### 3.1 Summary

Juvenile Onset Lordosis, or swayback, is a common hereditary conformation defect in Saddlebred horses where the back topline curvature drops within the first two years of life. The phenotype has been quantified using a Measurement of Back Contour (MBC), where horses of  $MBC > 7.0\text{cm}$  are considered high-MBC swayback. Genome-wide association studies identified a recessive haplotype on chromosome 20 to be associated with the high-MBC. Whole-genome sequence was conducted on 11 horses and sequences were compared in the target region for high-MBC and low-MBC horses. No variants were found that fit the criteria for a single, recessive mode of inheritance. This led to a new hypothesis, specifically that high-MBC is caused by multiple genetic factors with a major effect from chr20. The aim of the present study was to evaluate the entire genomic variation in the target region of chr20:41,000,000-44,000,000 to identify variants that might alter gene function, contributing to the high-MBC phenotype. A total of 9,691 variant loci were detected that make 21,463 transcript variations. Of these, 599 made coding sequence variations, including 315 synonymous, 250 missense, 14 frameshift, 9 in-frame deletion, 7 in-frame insertion, and 2 splice-donor and 2 start-loss variants. The distribution among affected and unaffected horses was compared along with computer predictions for impact of the variants on gene function. Based on these evaluations, potentially deleterious variants were found for 4 candidate genes, specifically *MDFI*, *C6orf132*, *PTK7* and *NCR2*. The strongest candidate based on predicted function was a frameshift deletion of 7bp in the exon 1 of the *MDFI* gene, at 20:41873061-41873068. *MDFI*-knockout mice show defects

in the formation of thoracic vertebrae and ribs, which restrains fusion of the spinous processes. This is consistent with necropsy reports on juvenile lordotic horses, where the spinous processes of the vertebral bones are underdeveloped. Further RNA expression studies are suggested to compare the expression of the *MDFI* gene in the affected spinous process of the vertebral bones with healthy bone tissues. Also, an additional Genome-Wide Association Study to control for the chr20 factor between affected and unaffected animals could reveal further loci affecting JOL.

### **3.2 Introduction**

In the previous chapter evidence was presented rejecting the hypothesis that high-MBC in Saddlebred horses, associated with JOL, was caused by a recessive variant within the region CHR20:41M-44M. Specifically, no variant was found in that region that was homozygous for the non-reference variant for all horses with high-MBC but heterozygous or homozygous for the reference variant in horses with low-MBC measurements. At the same time, the distribution of haplotypes among these horses also ruled out a simple dominant effect. When genetic variants in the region were compared among high-MBC and low-MBC horses, the association was reconfirmed. Therefore, genetic variation in this region does have an impact on the occurrence of high-MBC but its nature is complex. Part of the complexity may be involvement of other genes at other loci in an epistatic effect to cause the high-MBC. At the same time, subjective evaluation of horses, as described in the previous chapter, suggested that horses with low MBC may also have the gene. The situation appears to be that a high-MBC definitely identifies affected horses but horses with a low MBC may be predisposed but not affected. Future studies may be appropriate to

identify other loci associated with high-MBC. However, the genome sequence data generated in this study may allow us to identify variants that would be predicted to impact gene function in the region. The approach would be to investigate the genetic variants associated with annotated genes and look for those that may be deleterious to function.

This chapter aims to examine all the variants in the target region of ch20:41M-44M, regardless of their genotype distribution between the high/low MBC groups. In other words, to identify the best candidates from the genomic variation within the target window of ch20:41M-44M as likely effector on lordosis based on either the biological impact or any genes with functions that might be relevant to lordosis (e.g. genes controlling skeletal or connective tissues). Targets include large structural variations (large indels, inversions, translocations, etc.), changes in the exons of protein coding genes (e.g. frameshifts, premature stop codons, missense variants, etc.), changes leading to splice variants in introns, changes that affect regulatory regions of annotated genes or changes that affect poorly annotated regulatory regions or genes.

### **3.3 Materials and Methods**

#### **3.3.1 Phenotype**

The curvature of the back associated with extreme lordosis was measured as described by Gallagher et al (Gallagher et al., 2003). This method is based on two measurements: 1) the straight distance between two reference points on the horse's topline: point of the withers and point of the hip, called the "Straight Back Length". 2) the distance between those same two points along the contour line of the back, called "Contour Back

Length” (Figure 3.1). The Straight Back Length is subtracted from the Contour Back Length, and the result is the Measurement of Back Contour (MBC).

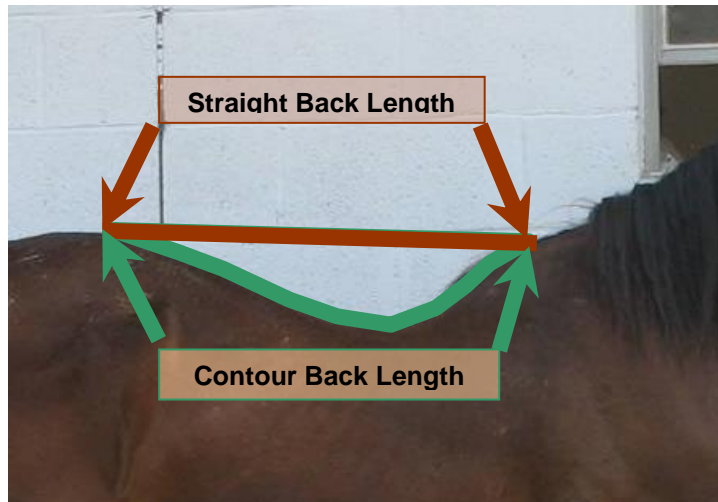


Figure 3.1 Straight Back Length and Contour Back Length in Measurement of Back Contour (= Contour Back Length – Straight Back Length).

Gallagher et al. (2003) defined an MBC of greater than 7.0cm as indicative of affectedness with Swayback and (known as high-MBC), whereas  $MBC < 7.0\text{cm}$  is categorized as low-MBC. This definition was mainly based on the bimodal distribution of the MBC measurements in the population of Saddlebred horses, where the cutoff point of  $MBC = 7.0\text{cm}$  best distinguished the two groups (Gallagher et al., 2003).

#### Experimental Animals

Whole-Genome sequence data for 11 Saddlebred horses was available from the study described in Chapter 2. Of these, 6 horses were selected based on phenotypic measures of MBC (3 high-MBC and 3 low-MBC) plus genotypes for the 4 SNPs used to define the TGTG haplotype from the study by Cook et al. (D. Cook et al., 2010). The remaining five horses were members of a family, which included two low-MBC parents, together with their three full-sib offspring, two of which were high-MBC.

To study the genotype distribution of selected variants in Saddlebred horse population, archived DNA samples were available for 109 American Saddlebred horses with MBC measurements and TGTG genotypes available from the study of Cook et al. (D. Cook et al., 2010) as well as newly collected samples from 7 high-MBC and 1 low-MBC horses. In total, our sample set comprised of 75 low-MBC and 34 high-MBC horses.

IACUC 2019-3247 was approved in connection with study of all experimental animals.

### **3.3.2 Bioinformatic analysis of the variants identified from the Whole Genome Sequence Data**

Variants called and genotyped by Genome Analysis Tool Kit (GATK) (McKenna et al., 2010) from whole genome sequence data of the 11 experimental animals (described in chapter 2) were filtered to include only the variants within the target region of chr20:41,000,000-44,000,000. The resulting VCF file was submitted to the Variant Effect Predictor online tool on the Ensembl website (McLaren et al., 2016) to identify the biological consequence of the genotyped variants.

Integrative Genomics Viewer (IGV) (Robinson et al., 2011) was used for visually browsing of the target region chr20:41,000,000-44,000,000 to identify or verify variants that are difficult to detect by GATK algorithms. Many of these variants are structural variations larger than 150 bp including, but not limited to, deletions, insertions, translocations. It should be acknowledged that not all structural variations were visible by IGV browsing; for example, insertions of larger than 150bp usually are not detectable by visual browsing of reads in IGV.



### 3.4 Results

The target region of chr20:41,000,000-44,000,000 had a total of 9,691 variants called using GATK and genotyped from the Whole Genome Sequence data on the 11 animals involved in the Whole-Genome Sequencing project. Table 3.1 presents a summary statistic of these variants.

Table 3.1 Summary statistics of the variants included in the VCF file at the target region of chr20:41M-44M.

<b>Category</b>	<b>Count</b>
Variants identified	9,691
Novel variants (percentage of total)	2,057 (21.2%)
Existing variants (percentage of total)	7,634 (78.8%)
Overlapped genes	96
Overlapped transcripts	172

Here are the definition of each category in the table: 1) Variants identified: the total number of variants that were successfully called and genotyped by GATK in the cohort of whole-genome sequenced horses; 2) Novel variants: the variants that have not been previously identified/annotated in the variant repository of Ensembl and have been newly identified in our studied horses; 3) Existing variants: the variants already identified/annotated in the variant repository of Ensembl; 4) Overlapped genes: the number of genes that harbored one or more of the variants in the VCF file; 5) Overlapped transcripts: the number of RNA transcript that harbored one or more of the variants in the VCF file.

Figure 3.2 categorizes the 9,691 variants based on their biological consequences.

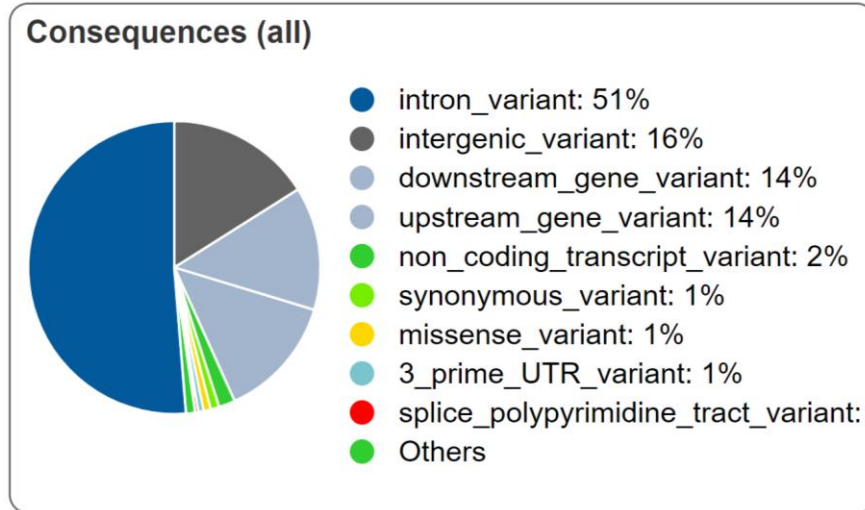


Figure 3.2 Biological consequence of the 9,691 variants identified in the target region of chr20:41M-44M.

Since each genomic variant can influence different transcripts of the same gene in different ways, the 9,691 were inferred by the Variant Effect Predictor to make 21,463 transcript variations. Of these, 599 made coding sequence variations, which include 315 synonymous, 250 missense, 14 frameshift, 9 in-frame deletion, 7 in-frame insertion, and 2 splice-donor and 2 start-loss variants. We prioritized investigating the variants based on the magnitude of biological impact they are known to have.

### 3.4.1 Investigation of Frameshift Variants

Table 3.2 provides a summary description of the frameshift variants identified by the GATK Haplotype Caller Tool.

Table 3.2 Frameshift variants identified within the target region of chr20:41M-44M.

Location	Allele	Gene	Transcript ID	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Variant Accession
20:41873061-41873068	-	<i>MDFI</i>	ENSECAG00000014590	1/5	725-731	725-731	242-244	GGA/X	GGGGGCGCG/GG	rs3435459097
20:41873061-41873068	-	<i>MDFI</i>	ENSECAG00000014590	1/4	725-731	725-731	242-244	GGA/X	GGGGGCGCG/GG	rs3435459097

20:42916977-42916977	G	-	ENSECAG00000010795	4/4	841-842	841-842	281	R/RX	CGG/CGGG	rs3433322688
20:43467645-43467645	G	<i>GTPBP2</i>	ENSECAG00000024984	1/13	56-57	56-57	19	P/PX	CCG/CCCG	-
20:43602365-43602365	GA	<i>VEGFA</i>	ENSECAG00000009402	6/6	1311-1312	1311-1312	437-438	-/X	-/GA	rs3432305126
20:43602365-43602365	GAGAGAG A	<i>VEGFA</i>	ENSECAG00000009402	6/6	1311-1312	1311-1312	437-438	-/ERX	-/GAGAGAGA	rs3432305126
20:43972073-43972074	-	<i>CAPN11</i>	ENSECAG00000000758	20/20	2380	2092	698	T/X	ACC/CC	-
20:43972073-43972074	-	<i>CAPN11</i>	ENSECAG00000000758	21/21	2455	2167	723	T/X	ACC/CC	-

Table 3.3 shows the genotype distribution of the above frameshift variants as called by GATK Haplotype Caller.

Table 3.3 Genotype distributions of the frameshifts in the target region Chr20:41M-44M called by GATK Haplotype Caller in the cohort of 11 horses whole-genome sequenced in this study. In the genotype cells, 0 codes for the reference allele and 1, 2 and 3 code for the alternate non-reference alleles, and dot “.” stands for unknown basecalls. For ease of visualization, homozygous reference genotypes have been colored in light green, heterozygous 0\_1 in light orange, and homozygous for the non-ref in light red. Also at the top row, high-MBC horses have been color shaded as dark red, control low-MBC horses in dark green, and the horses with low-MBC but questionable back structure in dark orange. Positions highlighted in blue identify the variants whose genotype distribution is concordant with haplotype pattern associated with the high-MBC.

Gene	Chr20 Position	Reference allele	Alternate allele	3517	3519	3529	3527	3520	3542	3535	3604	3601	3602	3603
<i>MDFI</i>	41873061	GGGGGCGC	G	1_1	1_1	0_1	0_0	1_1	1_1	1_1	0_1	0_1	1_1	1_1
-	42916977	C	CG	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>GTPBP2</i>	43467645	C	CG	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_1	0_0	0_0
<i>VEGFA</i>	43602365	C	CGAGAGA, CGAGAGAGA, CGA	0_2	0_0	0_3	1_2	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>CAPN11</i>	43972073	CA	C	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_1	0_0	0_1	0_0
<i>CAPN11</i>	43972074	A	C,*	0_0	._	0_1	0_0	._	0_1	0_1	0_2	0_1	0_2	0_0

Looking at the genotype distributions, only the frameshift variant in the *MDFI* frameshift genotype distribution was consistent haplotype pattern associated with high-MBC in the sequenced animals, and the rest of the variants have been called due to being

heterozygous in one or a few more animals, and the remainder being homozygous for the reference allele.

*MDFI* (*MyoD Family Inhibitor*), also known as *I-Mfa*, codes for a transcription factor that functions as a negative regulator of myogenic family proteins (Stelzer et al., 2016). Knockout mice show defects in the formation of thoracic vertebrae and ribs, which restrains fusion of the spinous processes (Kraut, Snider, Chen, Tapscott, & Groudine, 1998). Since the biological function of this gene could be directly related to the development mechanism of swayback in horses, this gene has been selected for discussion as a likely effector candidate for swayback development.

### 3.4.2 Investigation of Missense Variants

Missense Variants happened in 94 unique locations in the target region of chr20:41M-44M. They made a total of 250 transcript variations and pinned 39 annotated genes. 165 of those 250 transcript variations had a calculated SIFT score available, where only 38 of them were had a deleterious effect interpreted, and the remaining 126 had been marked as tolerated. These 38 putatively deleterious transcript variants mark 18 unique genomic locations inside 12 annotated genes, which are listed in the table 3.4.

Table 3.4 Missense variants with a putative deleterious SIFT scores in the target region of chr20:41M-44M.

Chr20 Location	Deleterious Allele	Gene	Gene Identifier	EXON	position cDNA	position CDS	Protein position	Amino acids	Codons	Variant Accession	SIFT
4224492 1	G	<i>C6orf132</i>	ENSECAG0000003493 0	4/6	2825	2825	942	G/A	gGg/gCg	rs343184626 8	Deleterious (0)
4237473 4	G	<i>TRERF1</i>	ENSECAG0000001697 1	3/15	1683	1229	410	K/T	aAg/aCg	-	Deleterious (0)

4305747 4	T	<i>PTK7</i>	ENSECAG0000001110 1	3/20	469	469	157	R/W	Cgg/ Tgg	rs114226442 9	Deleterious (0)
4319302 0	G	<i>SLC22A7</i>	ENSECAG0000000855 1	1/10	750	46	16	F/V	Ttc/ Gtc	-	Deleterious (0)
4330972 0	T	<i>ABCC10</i>	ENSECAG0000000516 9	14/2 1	3206	3206	1069	P/L	cCg/ cTg	rs113600144 1	Deleterious (0.01)
4160545 7	G	<i>NCR2</i>	ENSECAG0000001890 7	3/7	899	370	124	I/V	Atc/ Gtc	rs395823520	Deleterious (0.02)
4329836 9	A	<i>ABCC10</i>	ENSECAG0000000516 9	3/21	1390	1390	464	V/M	Gtg/ Atg	rs113776059 0	Deleterious (0.02)
4331288 7	T	<i>ABCC10</i>	ENSECAG0000000516 9	19/2 1	4174	4174	1392	L/F	Ctc/ Ttc	rs114111144 7	Deleterious (0.02)
4396384 2	G	<i>CAPN11</i>	ENSECAG0000000075 8	4/20	758	470	157	Q/R	cAg/ cGg	rs114068600 9	Deleterious (0.02)
4285638 3	C	<i>BICRAL</i>	ENSECAG0000002050 0	3/10	1000	995	332	V/A	gTv/ gCt	rs395604580	Deleterious (0.04)
4329754 2	G	<i>ABCC10</i>	ENSECAG0000000516 9	3/21	563	563	188	V/G	gTg/ gGg	rs115154623 7	Deleterious (0.04)
4188386 2	G	<i>MDFI</i>	ENSECAG0000001459 0	4/4	1373	1373	458	P/R	cCc/ cGc	rs343157911 5	deleterious low confidence (0)
4224726 2	A	<i>C6orf132</i>	ENSECAG0000003493 0	4/6	484	484	162	P/S	Cca/ Tca	rs114169861 2	deleterious low confidence (0)
4392437 9	G	<i>TMEM63 B</i>	ENSECAG0000001305 5	2/24	581	581	194	L/R	cTa/ cGa	rs114419718 2	deleterious low confidence (0)
4319951 0	G	<i>CRIP3</i>	ENSECAG0000001773 2	6/7	698	698	233	K/T	aAa/ aCa	rs396205752	deleterious low confidence (0.01)
4350944 0	C	<i>MRPS18A</i>	ENSECAG0000000487 0	1/7	26	26	9	T/R	aCg/ aGg	rs114980614 8	deleterious low

													confidence (0.01)
4160500 0	C	<i>NCR2</i>	ENSECAG0000001890 7	2/7	234	68	23	F/S	tTc/ tCc	rs114236231 1			deleterious low confidence (0.03)
4397208 3	C	<i>CAPN11</i>	ENSECAG0000000075 8	20/2 0	2389	2101	701	T/P	Acc/ Ccc	rs342962501 5			deleterious low confidence (0.03)

Table 5 shows the genotype distribution of the missense variants listed above in the table 4 among the 11 whole genome sequenced horses.

Table 3.5 Genotype distributions of the missense variants in the Table 3.4 called by GATK Haplotype Caller in the cohort of 11 horses whole-genome sequenced in this study. In the genotype cells, 0 codes for the reference allele and 1 code for the alternate non-reference alleles, and dot “.” stands for unknown basecalls. For ease of visualization, homozygous reference genotypes have been colored in light green, heterozygous 0\_1 in light orange, and homozygous for the non-ref in light red. Also at the top row, high-MBC horses have been color shaded as dark red, control low-MBC horses in dark green, and the horses with low-MBC but questionable back structure in dark orange. Positions highlighted in blue identify the variants whose genotype distribution is concordant with haplotype pattern associated with the high-MBC.

Gene	Chr20 Pos.	REF	ALT	3517	3519	3529	3527	3520	3542	3535	3604	3601	3602	3603
<i>C6orf132</i>	42244921	C	G	0_0	0_0	0_1	0_1	0_0	0_0	0_0	0_1	0_1	0_0	0_0
<i>TRERF1</i>	42374734	T	G	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>PTK7</i>	43057474	C	T	0_1	1_1	0_1	0_0	1_1	1_1	1_1	0_1	0_1	1_1	1_1
<i>SLC22A7</i>	43193020	T	G	0_0	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>ABCC10</i>	43309720	C	T	0_0	0_0	0_0	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>NCR2</i>	41605457	A	G	0_0	0_0	0_1	0_1	0_0	0_0	0_0	0_1	0_1	0_0	0_0
<i>ABCC10</i>	43298369	G	A	0_1	1_1	0_1	0_0	1_1	1_1	1_1	1_1	1_1	1_1	1_1
<i>ABCC10</i>	43312887	C	T	0_0	0_0	0_0	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>CAPN11</i>	43963842	A	G	0_0	0_0	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>BICRAL</i>	42856383	T	C	0_0	0_0	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>ABCC10</i>	43297542	T	G	0_1	1_1	0_1	0_0	1_1	1_1	1_1	1_1	1_1	1_1	1_1
<i>MDF1</i>	41883862	C	G	.	.	.	.	.	.	0_0	.	0_0	1_1	0_1
<i>C6orf132</i>	42247262	G	A	1_1	1_1	0_1	0_0	1_1	1_1	1_1	0_1	0_1	1_1	1_1

<i>TMEM63B</i>	43924379	T	G	0_1	0_1	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>CRIP3</i>	43199510	T	G	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>MRPS18A</i>	43509440	G	C	0_1	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>NCR2</i>	41605000	T	C	1_1	1_1	0_1	0_0	1_1	1_1	1_1	0_1	0_1	1_1	1_1
<i>CAPN11</i>	43972083	A	C	0_0	0_0	0_1	0_0	0_0	0_1	0_1	0_1	0_1	0_0	0_0

As highlighted with blue in table 5, only three variants with a deleterious SIFT score happened to have their non-ref allele associated with the high-MBC haplotype on chromosome 20. These three variants tag three genes, *PTK7*, *C6orf132* and *NCR2*.

*PTK7* (Protein Tyrosine Kinase 7) encodes a member of the receptor protein tyrosine kinase family of proteins that transduce extracellular signals across the cell membrane. The encoded protein lacks detectable catalytic tyrosine kinase activity, is involved in the Wnt signaling pathway and plays a role in multiple cellular processes including polarity and adhesion. It is involved with mental disorders like panic disorders like Anxiety, Panic, fear of open spaces and phobia of going out (Stelzer et al., 2016).

*C6orf132* (Chromosome 6 open reading frame 132) is an uncharacterized protein coding gene and its function is unknown.

*NCR2* (Natural Cytotoxicity Triggering Receptor 2), is a protein coding gene that is predicted to enable signaling receptor activity, involved in cellular defense response and signal transduction, located in plasma membrane and integral component of plasma membrane, and to be active in cell surface. Gene Ontology (GO) annotations related to this gene include transmembrane signaling receptor activity. Disorders associated with this gene include immune system disease including Newcastle's Disease (Stelzer et al., 2016).

It is noteworthy that a missense variant showed up in the *MDFI* gene but sequencing errors did not allow proper basecalling and genotyping, which prohibits further

investigation for possible effect, unless a specific assay is developed targeting this variant with a qPCR experiment.

### 3.4.3 Investigation of in-frame insertions/deletions

Table 3.6 lists the in-frame indel variants detected by the GATK Haplotype Caller in the target region of chr20:41M-44M.

Table 3.6 In-frame indel variants called by GATK Haplotype Caller within the genomic region of chr20:41M-44M among the cohort of 11 whole-genome sequenced horses.

Chr20 Location	Allele	Gene	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Variant Accession
42249583	CTC	<i>C6orf132</i>	6/11	1557-1558	258-259	86-87	-/E	-/ GAG	rs3091782435
42249583	CTC	<i>C6orf132</i>	6/8	2049-2050	258-259	86-87	-/E	-/ GAG	rs3091782435
42375022	GCT	<i>TRERF1</i>	5/18	1652-1653	940-941	314	L/QL	CTG/ CAGCTG	rs3091782431
42375022	GCT	<i>TRERF1</i>	5/17	1652-1653	940-941	314	L/QL	CTG/ CAGCTG	rs3091782431
42375022	GCT	<i>TRERF1</i>	5/17	1652-1653	940-941	314	L/QL	CTG/ CAGCTG	rs3091782431
43602365	GAGAGA	<i>VEGFA</i>	6/6	612-613	612-613	204-205	-/ER	-/ GAGAGA	rs3432305126
43993214	GGGGGGGG G	<i>SLC29A1</i>	1/14	120-121	2-3	1	M/MGG G	ATG/ ATGGGGGGGG G	-

Table 3.7 identifies the genotypes assigned to the 11 whole-genome sequenced horses for each of the in-frame insertion variants listed in the table 3.6.



Table 3.7 Genotype distributions of the in-frame indel variants in the Table 6 called by GATK Haplotype Caller in the cohort of 11 horses whole-genome sequenced in this study. In the genotype cells, 0 codes for the reference allele and 1, 2 and 3 code for the alternate non-reference alleles. For ease of visualization, homozygous reference genotypes have been colored in light green, heterozygous 0\_1 in light orange, and homozygous for the non-ref in light red. Also at the top row, high-MBC horses have been color shaded as dark red, control low-MBC horses in dark green, and the horses with low-MBC but questionable back structure in dark orange. Positions highlighted in blue identify the variants whose genotype distribution is concordant with haplotype pattern associated with the high-MBC.

Gene	Chr20 Position	Ref	Alt	3517	3519	3529	3527	3520	3542	3535	3604	3601	3602	3603
<i>C6orf132</i>	42249583	T	TCTC	1_1	1_1	1_1	0_1	1_1	1_1	1_1	1_1	1_1	1_1	1_1
<i>TRERF1</i>	42375022	A	AGCT	1_1	1_1	0_1	0_0	1_1	1_1	1_1	0_1	0_1	1_1	1_1
<i>VEGFA</i>	43602365	C	CGAGAGA, CGAGAGAGA, CGA	0_2	0_0	0_3	1_2	0_0	0_0	0_0	0_0	0_0	0_0	0_0
<i>SLC29A1</i>	43993214	T	TGGGGGGGGG, TGGGGGGGGGGGGG	0_0	0_0	0_1	0_0	0_0	0_0	0_0	0_0	0_2	0_0	0_0

As shown in table 3.7, only the genotypes of *TRERF1* insertion matched with the haplotype pattern associated with high MBC. *TRERF1* is a zinc-finger transcriptional regulating protein which interacts with CBP/p300 to regulate the gene *CYP11A1* in humans. Diseases associated with *TRERF1* in humans include Estrogen Resistance and Breast Cancer (Stelzer et al., 2016). This gene has been investigated further in this chapter.

#### 3.4.4 Visual Investigation of the Structural Variation

Visual browsing of the chr20:41M-44M target region using Integrative Genomics Viewer (IGV) resulted in identification of the following variants as candidates for having effect on the lordosis phenotype.

- 1- Chr20:42,222,093. Deletion of a 224 bp Equine Repetitive Element 1 (ERE1), also known as SINE or Short Interspersed Nuclear Element located at the 3' Untranslated Region (UTR) of the *TAF8* gene.

- 2- Chr20:42,399,504. Deletion of a 215 bp segment within the second intron of the *TRERF1* gene. This is the closest gene to the peak of association (chr20:42,430,946) found by Cook et al (D. Cook et al., 2010).

The above variants were the only structural variants with significant size of over 120 bp (larger than sequencing read size) to be inside genes or within regulatory regions of genes.

### **3.5 Discussion**

As described in the Chapter 2, examination of every existing variant within the target region of chr20:41M-44M by whole genome sequencing showed that no variant could entirely distinguish the case group from controls, i.e. high versus low MBC groups. At the same time, association tests with new markers in the target region showed that the effect of chromosome 20 haplotype is not a statistical artifact, therefore the chromosome 20 factor is still involved in the formation of lordosis. With these pieces of evidence, it is inferable that the lordosis is genetically caused by interaction of genetic factors on chromosome 20 with other factors, possibly other loci in the genome which complete the action of chromosome 20 factor to cause high-MBC. Another complicating factor is the phenotype used to characterize the trait, namely high-MBC. All phenotype assessments in the chain of studies performed so far have been founded on a case/control basis, i.e. comparing a group defined as 'high' MBC versus a 'low' MBC group. However, there could be a spectrum of phenotypic variation for back curvature, where many of the horses in the population of Saddlebred horses could not fit within any of those two categories. Horses defined as unaffected ( $MBC < 7.0$ ) may still be affected. This could be the case

with the dam in the Experiment 2 of chapter 2, where she had the same genetic structure as her high-MBC offspring across the entire target window of chr20:41M-44M.

This raises two questions. Firstly, what other factors interact to cause high-MBC? At the end of chapter 2, approaches to this question were suggested but fall outside this study. Secondly, what genetic factors in the region Chr20:41M-44M might contribute to high-MBC phenotypes? This second question is addressed in this chapter. Specifically, genetic variants within this region were investigated, using current gene annotation, to identify variants that might impact gene function. Due to complications in genotype distributions between the high and low MBC groups, causal variant discovery based on association with the phenotype seems impossible. Hence the next alternative approach would be to screen the variants based on their biological impact according to the relevance that they may have to lordosis, or the profoundness of their consequence (for example, prioritizing missense variants over synonymous variants).

Four candidates were chosen for further ontology study based on their distribution among the affected and non-affected horses as well as SIFT score predictions. These include 1) frameshift variant in the *MDFI* gene, 2) ERE1 deletion at the 3'UTR of the *TAF8* gene, 3) missense variant in the *C6orf132* gene and 4) a 215 bp deletion within an intron of the *TRERF1* gene. Each of these variants have been discussed in the following.

### **3.5.1 *MDFI* gene variant**

Perhaps the strongest candidate gene in the entire target region of chr20:41M-44M is the *MDFI* (*MyoD Family Inhibitor*), which was initially spotted based on a putative frameshift variant that was detected in the exon 1 of one of its transcripts. The most interesting features of this gene are the frameshift nature of the variant which could have

profound effects downstream, as well as the biological relevance of the function of the gene, as documented in other species. Kraut et al (1998) studied the function of the *MDFI* gene in mice. They showed that *MDFI*-knockout mice are incurred with defects in formation of thoracic vertebrae and ribs. Based on their report, the underdevelopment in the vertebral bones is localized around the spinous processes, which consequently restrains fusion of the spinal bones (Figure 3.3) (Kraut et al., 1998).

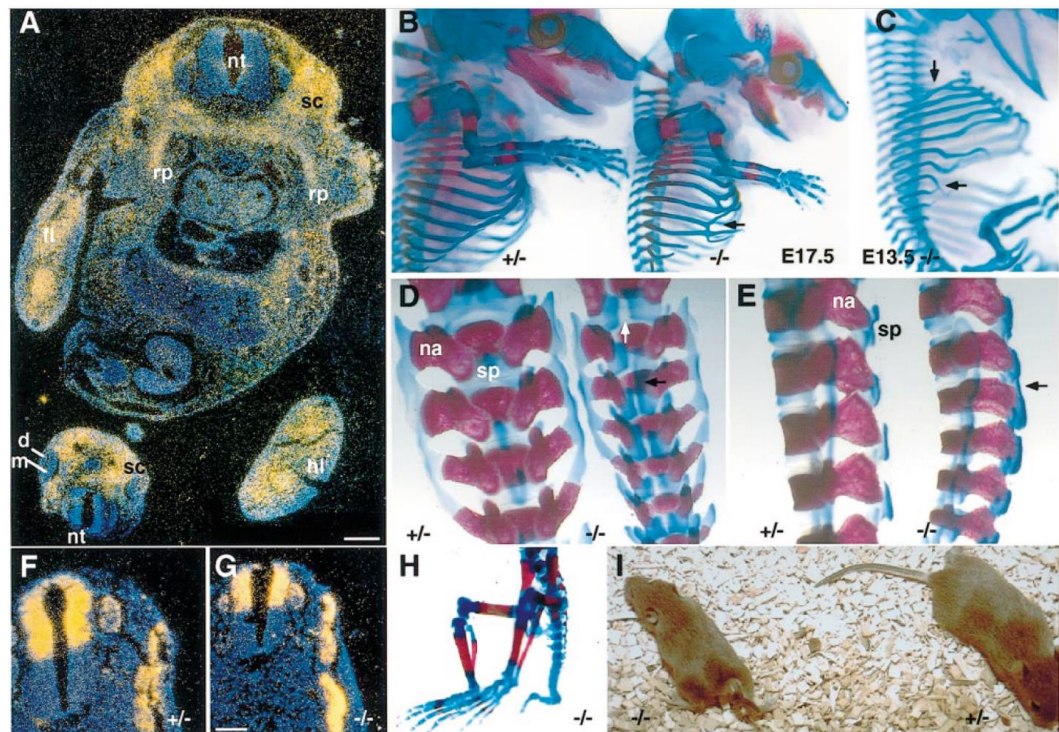


Figure 3.3 Courtesy picture from Kraut et al. (1998), showing defects in formation of ribs and thoracic vertebral bones in mice knocked out for the *MDFI* gene (-/- being homozygous knockout, +/- being heterozygous). Most notable are the sections D and E of the figure, which show that in *MDFI* mutant newborns, there are abnormal fusions of spinous processes. The blue staining of spinous processes in the mutant, which is shown on the right (-/-), do not merge together medially as indicated by the white arrow. However, the spinous processes can still fuse in a cranial-caudal direction, as demonstrated by the black arrow.

Their findings in mice perfectly aligns with what Rooney et al observed in the necropsy report of mix-bred lordotic horses, as they found the spinous processes of the

thoracic vertebral bones to be hypoplastic (underdeveloped) (Rooney & Prickett, 1967). Since the biological function of this gene could be directly related to the development mechanism of swayback in horses, it becomes the best candidate gene as a likely effector in swayback development.

The *MDFI* gene is conserved across various species including chimpanzees, Rhesus monkeys, dogs, cows, mice, rats, chickens, zebrafish, and frogs. In fact, 456 organisms have orthologs with the human *MDFI* gene. Additionally, there is evidence to suggest that *MDFI* is implicated in Spondylocostal Dysostosis 4, an Autosomal Recessive disorder that exhibits phenotypes such as Myelomeningocele, Rib fusion, Vertebral segmentation defect, Hemivertebrae, and Spina bifida occulta (OMIM:613686).

The *MDFI* frameshift variant detected in our sequenced horses results from deletion of 7 base pairs spanning chr20:41,873,062-41,873,068. This happens inside the boundaries of Exon 1 in one of the two transcripts that the *MDFI* gene, known as MDFI-201 transcript (ENSECAT00000063568.2). This exon which is identified as ENSECAE000000306796, is 2,116 bp long and spans chr20:41,871,081-41,873,196. At the first look at the genotypes of the *MDFI* frameshift variant that was assigned by GATK Haplotype Caller to the 11 sequenced animals in the study (Table 3, first row), the frameshift deletion sounds to be perfectly embedded within the chr20 haplotype that was found associated with the high-MBC. However, visual evaluation of the sequences imply that the animals may have a different genotype than that assigned by GATK Haplotype Caller automatically. Figure 3.4 visualizes the arrangement of reads locally realigned by the GATK Haplotype Caller around the 7 bp deletion in the horse 3519, which was initially genotyped as homozygous for the deletion.



Figure 3.4 An IGV screenshot of the local realignment of sequencing reads by GATK Haplotype Caller around the 7bp frameshift deletion of the *MDFI* gene in the horse 3519.

As it appears from the figure 3.4, although the GATK Haplotype Caller genotyped the horse 3519 as homozygous for the 7bp deletion, the visual evaluation of the reads locally realigned in this region suggest that the horse is actually a heterozygous. It becomes ambiguous to visually evaluate the genotype of the rest of the animals in this study. For example, in the horse 3603 who has deepest sequencing coverage around the variant, 24 reads agree on presence of the deletion, while the other 6 reads endorse absence of the deletion. In 6 of those reads, 3 cover the variant location at the middle of the read. In cases like the horse 3603, it is unclear if the horse is a real heterozygous or if the observation is due to an error in alignment of the reads to the reference sequence. Noteworthy, genotyping errors are expectable as the variation site is surrounded by repeats of G( $\times$ N)C sequence, which should bring difficulty at both sequencing steps (due to high GC content) as well as alignment (because of repeats). Further investigation of *MDFI* frameshift deletion is

warranted using custom genotyping assays that specifically target the variant, to reveal the actual genotype of the high and low MBC horses. There has also been a missense variant detected in the *MDFI* gene (chr20:41,883,862), whose genotype calls look ambiguous too, so custom genotyping is required to examine the potency as a likely effector. Also interesting to investigate is for the presence of any genomic imprinting on the gene; this is particularly because the sequence composition around the frameshift variant contains several CpG islands, so it could be subject to methylation. If this is the case, it could resolve the discrepancies in genotype distributions of the variants between the low and high MBC groups, particularly the cases like the dam in the fullsib family, who is homozygous for the entire high-MBC haplotype, identical to her high-MBC offspring. There is evidence for methylation of the *MDFI* gene in human genome, where the methylation was found involved in colorectal cancer (J. Li et al., 2017). Ultimately, an RNA expression study on an affected bone tissue (ideally spinous processes) from a juvenile swayback Saddlebred to be compared to any healthy bone tissue from the same horse as well as a vertebral bone from an unaffected horse could reveal if any aberrant transcript of the *MDFI* gene exists in the affected tissue. Interesting about *MDFI* expression, evaluation of the RNA-Seq data from the FAANG project shows the exon harboring the frameshift variant is not expressed in most tissues, but it is expressed in sesamoid bones (the only bone tissue available from the FAANG data), which sounds promising to be expressed in the vertebral bones (Burns et al., 2018). Figure 3.5 shows RNA-Seq data from sesamoid bone tissues around the frameshift site.

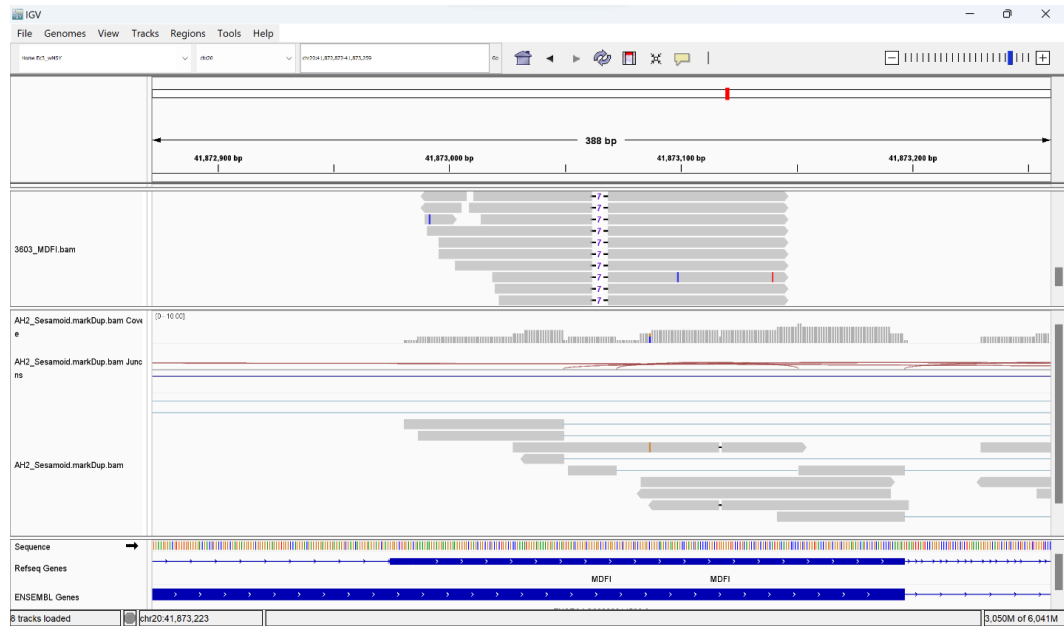


Figure 3.5 IGV screenshot of FAANG RNA-Seq data (bottom panel) and DNA sequence data around the frameshift deletion from the horse 3603 (top panel).

None of the other variants detected in the region had a biologically relevant function to development of lordosis, as *MDFI* did. Aside from the *MDFI* variant though, the three variants selected for population genotyping are picked because of the depth of impact they could have on gene function. Each of these likely impacts is discussed in the following.

### 3.5.2 ERE1 deletion at chr20:42,222,093 around 3'UTR of the *TAF8* gene

Equine Repetitive Element 1 are equine-specific SINE elements (standing for Short Interspersed Nuclear Element) that are spread throughout the genome in 45,713 unique places. Some of them have been identified to have a phenotypic effect by disrupting the functional sequences of coding genes. An explicit example of these is an ERE1 insertion within the promoter of the Myostatin gene which is known to affect optimal racing distance in Thoroughbred horses by altering body muscle mass proportions (Hill et al., 2010; Santagostino et al., 2015). In the case of ERE1 deletion at chr20:42,222,093, it happens



closer to the 3' end of the *TAF8* gene, which at the first look could be a potential disruptor of the 3'UTR of the gene. However, consulting with the expression data available from Functional Annotation of Animal Genome Project, the ERE1 occurs outside the expressed region. Hence, it will make it less likely to alter *TAF8* gene expression by disrupting its 3'UTR.

### **3.5.3 Missense SNP chr20:42,247,262G>A within the *C6orf132* gene**

This is the one of the few missense SNPs in the entire target region to have a calculated SIFT score of zero which is interpreted by the Ensembl to have a deleterious effect on the protein product of the gene. The genome annotation shows this variant to be in the fourth exon of the *C6orf132* gene. However, comparing the genomic annotation of this gene to the actual RNA-Seq data available from the FAANG project, it seems that the location of exons in the annotation are several base pairs away from what the RNA-Seq shows to be the actual location of the exon. This could lead to misinterpretation of the SNP to alter codons deleteriously or even happen within an actual exon.

### **3.5.4 215bp Deletion at chr20:42,399,504-42,399,718 within *TRERF1* gene**

At the first look on this variant, it seems interesting that it is relatively large in size and its sequence does not show up anywhere else in the genome, including ERE elements. It happens within the second intron of the *TRERF1* gene, which is the closest gene to the peak of Association in Cook's Genome Wide Association Study (D. G. Cook, 2014). Although being an intronic variant, if it happened to alter splicing of any of its two neighboring exons, it could be a strong candidate to have a potential effect on gene function and consequently lordosis. However, the deletion is 1,997 bp away from the Exon 2 (chr20:42,401,715-42,401,829) and 23,284 bp away from the Exon 3 (chr20:42,374,520-

42,376,220) of the gene. Since the deletion is far away from both exons, it seems unlikely to have any effect on exon splicing. Also consulting FAANG data from expression of a few tissues (Sesamoid bones and Muscle tissue), no RNA expression is found within or around the location of the deletion (Figure 3.6) (Burns et al., 2018).



Figure 3.6 RNA-Seq on the equine Sesamoid bones and Muscle Tissue in the IGV around the 215 bp deletion of the *TRERF1* gene. The bottom track marks the location of the deletion in the horse 3519 who is homozygous for the deletion.

The above observations lower expectations from the latter three candidate variants to have a functional effect on the development of lordosis. Hence, it looks unlikely for them to have a potential causative variant. The other way of approaching the causal variant would follow a reverse direction, by investigation of gene expression. For example, if any gene in the region happens to have an aberrant transcript in the affected individuals while having a normal transcript in none-affected horses, it could be marked as a potentially causal gene. This not only applies to any aberrant transcripts but also quantification of the transcripts. This means differential expression levels of any gene in the affected individuals as compared to the unaffected horses. The *MDFI* gene sounds to be the strongest candidate,

if looking at a single gene is the only option. A challenge in conducting expression studies is choosing a tissue that would be directly affected by lordosis, plus the most relevant tissues (thoracic vertebral bones) are very difficult and invasive to access. However, if aberrant transcripts are constitutively expressed in all body tissues, any accessible tissue including blood can distinguish the affected individuals from non-affecteds. FAANG data indicates that all three candidate genes (*TAF8*, *C6orf132* and *TRERF1*) are expressed in the equine Peripheral Blood Mononuclear Cells (PBMC) (Burns et al., 2018), which makes it a readily-available tissue for transcript detection. However, this is not the case with the *MDFI* gene, as it is not expressed in PBMC.

### **3.5.5 Future Directions**

The analyses presented above are attempts to identify best candidates on the chromosome 20 as a major factor implicating lordosis. However, they do not test the new hypothesis developed in the introduction of this chapter (high-MBC as a complex multigenic trait with a major effect from chr20) as an alternative to the initial hypothesis governing Chapter 2 (single recessive variant on chr20 causing lordosis with mendelian inheritance). Looking at the overall layout of the results obtained in the whole project, the main complicating factors seem to be: 1) definition of the phenotype based on a categorical basis of high versus low MBC is defective, meaning that it cannot capture all the variation in the back curvature and distinguish the lordotic individuals from non-affected animals; 2) mode of inheritance on chromosome 20 is still unknown, which is likely obscured by the complementary action of other loci elsewhere in the genome, making almost all chr20 genotypes being observed in both low and high MBC groups. The only observable differences in the distribution of genotypes between the two groups are 1) Larger

proportion of homozygous non-ref individuals are in the high-MBC group, 2) larger proportion of heterozygous individuals are in the low-MBC group, 3) no high-MBC individuals were found to be homozygous for the ref allele. This makes it difficult to predict the high/low MBC phenotype of an individual based on its chr20 genotype. For example, if a horse happens to be heterozygous for the chr20 haplotype, it will not be possible to say whether it is going to be high or low MBC. Now the idea is that a complementary genotype on another locus elsewhere in the genome could resolve the ambiguity of genotype distribution on chromosome 20. A hypothetical example is presented here about genotype complementarity between two loci that could explain the distinction between an affected versus non-affected individuals with the same genotype on chromosome 20:

Phenotype	Chr20	Chr [unknown]
Affected with lordosis	AA	BB/Bb
Not affected with lordosis	AA	bb

In this case, although both animals are homozygous for the A allele on chromosome 20, dominant action on an unknown chromosome locus could be determinative of lordosis phenotype.

In this scenario, a specific aim to test the new hypothesis of multigenic nature of the trait would be a Genome Wide Association Study that controls for the effect of chromosome 20 could reveal a secondary peak that is indicative of the second locus on the genome to be involved in formation of lordosis. Luckily the haplotype on chr20 that is associated with the high MBC is also the most prevalent haplotype in the population of Saddlebred horses (D. Cook et al., 2010), with a frequency of 0.43. This means that non-affected Saddlebreds that are homozygous for the entire region of chr20:41M-44M should

be accessible in sufficient numbers to conduct a GWAS. Plus, the chr20:41M-44 is small enough to find horses like 3535 to be homozygous for the chr20 haplotype. A hypothetical study would involve at least 20 swayback and 20 non-swayback horses that are all homozygous across the region chr20:41M-44M. Tag SNPs could be selected to verify the haplotype zygosity of the individuals to be selected for the GWAS. Any of the SNP genotyping arrays (Equine SNP70 by Illumina or Axiom 670K SNP Array by Affymetrix) can be employed to conduct genome-wide SNP genotyping.

Since a strong candidate gene (*MDF1*) has been introduced in this study based on the relevance of its function to the development of lordosis, an RNA expression study that was described earlier in the discussion of the gene, sounds to be promising. Since similar studies in mice as well as necropsy studies on swayback horses indicate the spinous process of the thoracic vertebral bones to be affected, this tissue is proposed to be ideal for comparison between a juvenile lordotic and an unaffected Saddlebred, or the articular process of thoracic vertebral bone of a lordotic individual to another unaffected bone of the same individual. This will account for genetic differences that arise from individual to individual.

## CHAPTER 4. GENOMIC COMPARISONS OF PERSIAN KURDISH, PERSIAN ARABIAN AND AMERICAN THOROUGHBRED HORSE POPULATIONS

### 4.1 Summary

The present research aimed to characterize the Persian Kurdish horse population relative to the Persian Arabian and American Thoroughbred populations using genome-wide SNP data. Fifty-eight Kurdish, 38 Persian Arabian and 83 Thoroughbred horses were genotyped across 670,796 markers. After quality control and pruning to eliminate linkage disequilibrium between loci which resulted in 13,554 SNPs in 52 Kurdish, 24 Persian Arabian and 58 Thoroughbred horses, the Kurdish horses were generally distinguished from the Persian Arabian samples by Principal Component Analyses, cluster analyses and calculation of pairwise  $F_{ST}$ . Both Persian breeds were discriminated from the Thoroughbred. Pairwise  $F_{ST}$  between the two Persian samples (0.013) was significantly greater than zero and several fold less than those found between the Thoroughbred and Kurdish (0.052) or Thoroughbred and Persian Arabian (0.057). Cluster analysis assuming three genetic clusters assigned the Kurdish horse and Thoroughbred to distinct clusters (0.942 in cluster 2 and 0.953 in cluster 3 respectively); the Persian Arabian was not in a distinct cluster (0.519 in cluster 1), demonstrating shared ancestry or recent admixture with the Kurdish breed. Diversity as quantified by expected heterozygosity was the highest in the Kurdish horse (0.342), followed by the Persian Arabian (0.328) and the Thoroughbred (0.326). Analysis of Molecular Variance showed that 4.47% of the genetic variation was present among populations ( $P < 0.001$ ). Population-specific inbreeding indices ( $F_{IS}$ ) were not significantly different from zero in any of the populations. Analysis of individual inbreeding based on runs of homozygosity using a larger SNP set suggested greater

diversity in both the Kurdish and Persian Arabian than in the Thoroughbred. These results have implications for developing conservation strategies to achieve sound breeding goals while maintaining genetic diversity.

This work was published: Yousefi-Mashouf N, Mehrabani-Yeganeh H, Nejati-Javaremi A, Bailey E, Petersen JL. Genomic comparisons of Persian Kurdish, Persian Arabian and American Thoroughbred horse populations. *PLoS One*. 2021 Feb 16;16(2):e0247123. doi: 10.1371/journal.pone.0247123. PMID: 33592064; PMCID: PMC7886144.

## **4.2 Introduction**

The Kurdish horse of Iran is one of the four major Iranian horse breeds. The other Iranian horse breeds include Caspian, Turkoman, Persian Arabian (also known as Assil). Historical literature chronicles a developmental history of more than 2500 years for Kurdish horses, relating them to an ancient, now-extinct population of horses called “Nesayee.” The Nesayee horses have been documented to have served as transportation for the army of Medes tribe, whose realm was congruent with today’s homeland of Kurdish horses (west of Iran) (Diakonoff, 1956).

No formal registry exists for the Kurdish horse, however, the breed is a well-known landrace, selected for agility, dressage gaits, mountain riding and resistance to harsh environmental conditions. We previously characterized the Kurdish horse from a phenotypic perspective by establishing the breed standards, which describe the ideal characteristics used as selection criteria by breeders (Yousefi-Mashouf, Mehrabani-Yeganeh, Nejati-Javaremi, & Maloufi, 2020).

The Persian Arabian horse originated in the southwestern part of Iran but currently is a geographically neighboring population to the Kurdish horse and one might suspect some admixture or ancestral relationship between these two populations, as also endorsed by the available historical information. Preliminary genetic studies suggested an immediate common ancestor between Kurdish and Persian Arabian populations (Ovchinnikov et al., 2018). In addition, as these breeds occupy overlapping geographical areas, the question has arisen as to whether they are distinct. Hence, the present research aimed to characterize the diversity of and relationships between these Persian breeds using genome-wide single nucleotide polymorphism (SNP) data from Kurdish horses and Persian Arabians as well as data from the more distantly related American Thoroughbred. We hypothesized that the current population of Kurdish horses can be considered as a unique and homogeneous population distinct from Persian Arabian horses from the genomic standpoint.

### **4.3 Materials and Methods**

#### **4.3.1 Sampled Individuals**

We sampled 58 Kurdish horses (43 males and 15 females) distributed over a wide geographic range to ensure the highest level of diversity, including five provinces of Kermanshah, Kurdistan, Western Azerbaijan, Isfahan and Kerman. The 38 Persian Arabian samples (11 males and 27 females), all registered in the Persian Arabian studbook, were obtained in the provinces of Khouzestan, Yazd and Kerman (Figure 4.1). The Kurdish horses sampled from the central locations (Isfahan and Kerman) did not originate there, rather they (or their ancestors) were imported from the three western provinces (Kermanshah, Kurdistan and Western Azerbaijan). Similarly, the Persian Arabian horses



sampled from the central locations (Yazd and Kerman), have their origin from the southwest (Khouzestan). DNA samples were provided from the archive at the University of Kentucky for 83 American Thoroughbred horses (44 males and 39 females). The Thoroughbred horses were randomly sampled from 8 farms in central Kentucky.

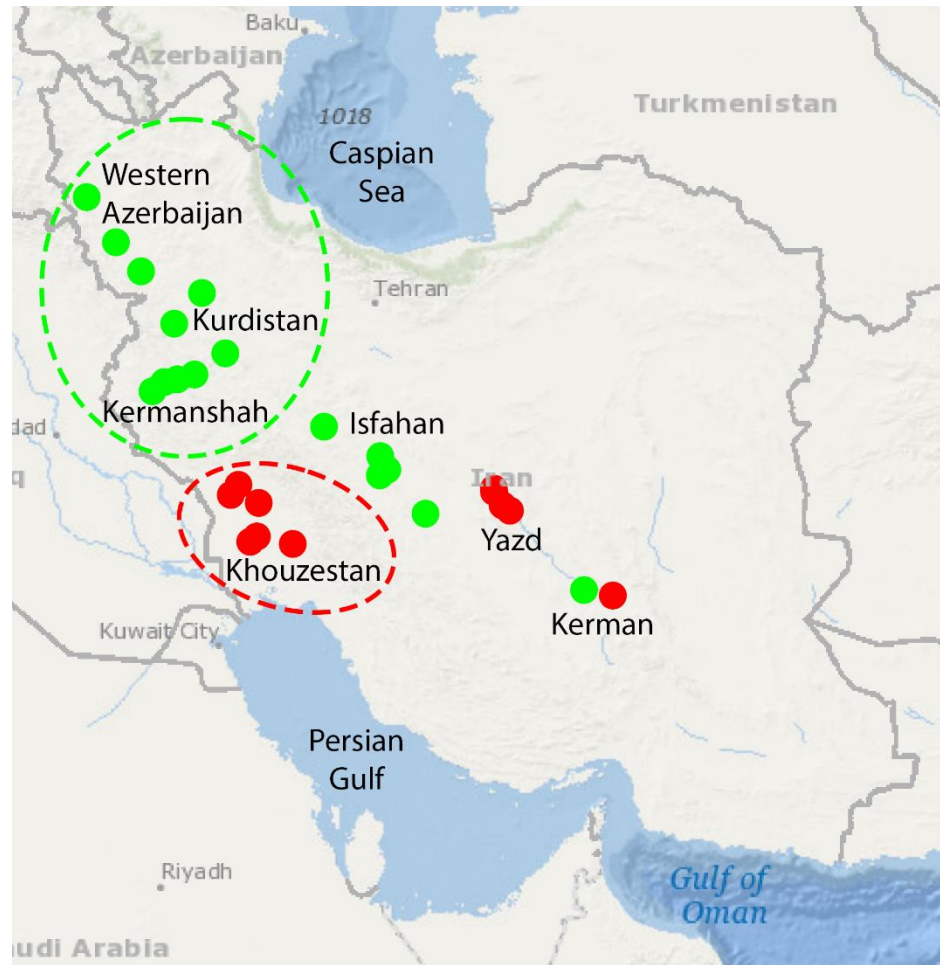


Figure 4.1 Sampling locations for Iranian populations. Green and red points signify sampling locations for Kurdish and Persian Arabian horses, respectively. The regions circled in green and red identify the original homeland of Kurdish and Persian Arabian horses, respectively. All the point locations outside the ovals represent horses that were descendants of, or were themselves imported from the original homelands. Map imported from the USGS National Map open resources.

### **4.3.2 Blood collection and DNA extraction**

Blood samples drawn from the jugular vein were collected in 6 ml EDTA vacuum tubes, transported to the lab cooled, and were kept frozen until DNA isolation. DNA extraction was carried out using Phenol-Chloroform protocol (Sambrook & Russell, 2006).

### **4.3.3 Ethical Statement**

The IACUC committee at the University of Kentucky waived review by the ethics committee in 2010 since the Thoroughbred horse samples were obtained from privately owned horses and provided by the owners. Likewise, the samples from Persian horses came from horses that were privately owned and managed, samples were provided by owners so a formal review of the study protocol for ethical treatment of horses was deemed unnecessary. The Department of Animal Science at the University of Tehran is responsible for evaluations of ethical use of animals.

### **4.3.4 Genotyping**

DNA samples were submitted to Neogen GeneSeek (Lincoln, Nebraska) for genotyping with the Axiom Equine Genotyping Array (Affymetrix Inc.) which harbors 670,796 SNP markers (Schaefer et al., 2017).

### **4.3.5 Data Analysis**

SNP & Variation Suite version 8 (Golden Helix, Inc., Bozeman, MT, [www.goldenhelix.com](http://www.goldenhelix.com)) software (Golden Helix) was used for basic quality control of the genotype data, in which the markers were disqualified if the Minor Allele Frequency (MAF) was  $< 0.05$  and per-SNP Call Rate  $< 0.95$ . All the markers on the X chromosome or with unknown genomic location (categorized as CHR\_UN) were removed from the

dataset. To avoid inclusion of closely-related individuals, the dataset was pruned for Identity by Descent (IBD), such that between each pair of individuals with  $IBD > 0.2$ , the one that had more genetic relationship with the other individuals in the sample set was removed. Lastly, SNPs that were in LD across samples were also removed, pruning for an LD threshold of  $r^2=0.25$ , considering 100 SNP windows and moving 25 SNPs per set (LD computation method: CHM). As the unequal number of samples among populations could bias LD pruning towards the populations with higher number of samples, we randomly selected 24 individuals from each population and performed LD pruning on the 72-individual subset ( $24 \times 3$ ), then applied the selected markers to the whole population. To analyze runs of homozygosity, however, LD pruning was not applied although markers with a genotyping rate  $< 0.95$  and the loci on the X chromosome or contigs unassigned chromosomes were removed.

Principal Component Analysis was carried out in the SNP & Variation Suite v8 (SVS). Pairwise  $F_{ST}$  values between breeds were calculated using SVS and Arlequin version 3.5 (Excoffier & Lischer, 2010). Arlequin was also used to obtain expected heterozygosity ( $H_E$ ) and population-specific inbreeding ( $F_{IS}$ ) values as well as to perform an analysis of molecular variance (AMOVA). Average inbreeding coefficients for each population were calculated using SVS v8 and PLINK 1.07 (Purcell et al., 2007). To calculate runs of homozygosity (ROH) and obtain ROH-based inbreeding coefficients for each individual, the R package detectRuns was used (Biscarini, Cozzi, Gaspa, & Marras, 2018) with parameters: windowSize=15, threshold=0.1, minSNP=15, maxOppWindow=1, maxMissWindow=1, maxGap=1000000, minLengthBps=250000, minDensity=1/10000. To quantify ROH of various lengths (0 to  $\geq 48$ Mb), 20 individuals from each sample were

randomly selected in three iterations with the mean and standard deviation in counts/class calculated for each sample. Clustering of breeds into genetic groups was examined using the STRUCTURE program version 2.3.4 assuming K values of 1 to 5, replicating the analysis of each K value five times. The STRUCTURE algorithm assumed the admixture model and correlated allele frequencies. Burn-in iterations of 10,000, 25,000, 100,000, 150,000 and 300,000 reps were tested along with different MCMC repetitions of 100,000, 200,000, 250,000 and 600,000 to confirm convergence. Also, to see if the Thoroughbred samples biased the clustering patterns of Kurdish and Persian Arabian groups, we ran STRUCTURE without the Thoroughbred samples at K=2. To determine the optimal value of K using the Evanno method (Evanno, Regnaut, & Goudet, 2005), the online program STRUCTURE Harvester (Earl, 2012) was employed.

## **4.4 Results**

### **4.4.1 Data Pruning**

After data pruning, the final dataset included 13,554 SNPs for a total of 134 individuals out of the 180 horses genotyped. The IBD filtering left 50 out of 58 Kurdish, 24 out of 38 Persian Arabian and 58 out of 83 Thoroughbred horses. For ROH, after removing SNPs with genotyping rate < 95%, 262,390 autosomal SNPs were included in the analysis.

### **4.4.2 Principal Component Analysis (PCA)**

The first Principal Component explained 6.45% of the variance, which discriminated the Thoroughbred from both Iranian breeds. The second PC, capturing

2.05% of the variance, showed divergence between the Kurdish and Persian Arabian samples (Figure 4.2).

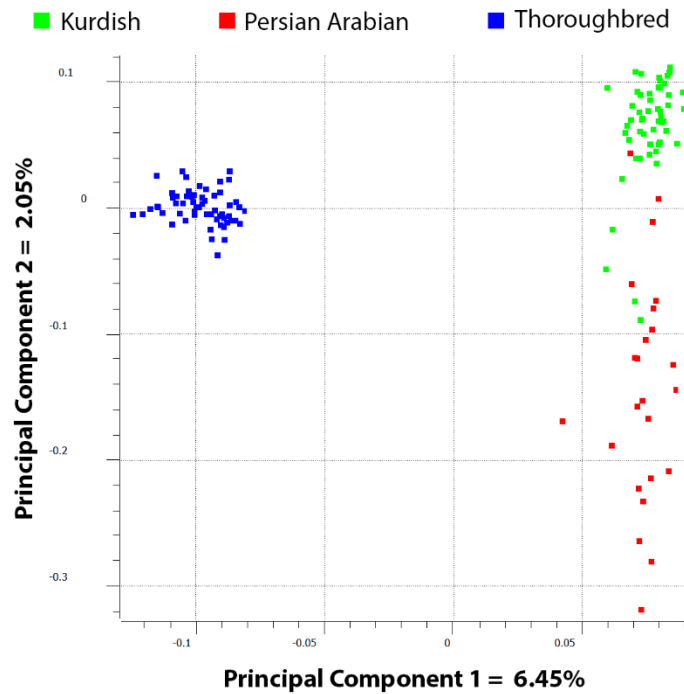


Figure 4.2 Plot of principal components 1 versus 2 for the 134 horse representing 3 breeds.

The plot also showed that Thoroughbred horses formed a tight cluster. The Kurdish horses clustered in a manner similar to the Thoroughbred, while the Persian Arabian horses had a wider distribution across PC2, where the Kurdish and Persian Arabian clusters were adjacent to one another, with several individuals lying in overlapping regions.

#### 4.4.3 $F_{ST}$

Pairwise  $F_{ST}$  values were all significantly greater than zero with the least divergence observed between the Kurdish horse and Persian Arabian (Table 4.1).

Table 4.1 Pairwise  $F_{ST}$  values between breed groups. All of the P values were significant ( $P < 0.05$ ).

Population Pair	$F_{ST}$	P Value
-----------------	----------	---------

Kurdish – Persian Arabian	0.013	<0.0001
Kurdish – Thoroughbred	0.052	<0.0001
Thoroughbred – Persian Arabian	0.057	<0.0001

#### 4.4.4 Population Specific Inbreeding

Although greatest in the Kurdish horses, no  $F_{IS}$  value for any sample was significantly greater than zero (Table 4.2).

Table 4.2 Population-specific  $F_{IS}$  values.

Population	$F_{IS}$	P (Rand $F_{IS} \geq$ Obs $F_{IS}$ )
Kurdish	0.005	0.386
Persian Arabian	-0.018	0.620
Thoroughbred	-0.008	0.636

#### 4.4.5 Analysis of Molecular Variance (AMOVA)

AMOVA identified 4.47% of the variation present among populations ( $P < 0.001$ ), -0.41% among individuals within populations ( $P = 0.585$ ), and 95.94% within individuals ( $P = 0.065$ ). By eliminating the Thoroughbred from the analysis, the among-population variation remained significant ( $P < 0.001$ ), explaining 1.29% of the variation between the Kurdish and Persian Arabian samples.

#### 4.4.6 Runs Of Homozygosity Analysis

The Thoroughbred samples had a greater number of longer runs of homozygosity than the Persian and Kurdish horses (table 4.3). The Kurdish horse had fewer ROH  $\geq$  6Mb than either the Persian Arabian or Thoroughbred, with no ROH longer than 24Mb.

Table 4.3 Summary of the total number of runs of homozygosity by each size class. Due to unequal sample size, three iterations, each of 20 randomly selected individuals was evaluated. Given is the mean and standard deviation (in parenthesis) of the replicates.

ROH size (Mb)	Kurdish	Persian Arabian	Thoroughbred
0-6	23321.0 (138.7)	22207.3 (140.7)	15594.7 (129.9)
6-12	26.0 (2.6)	88.0 (3.5)	304.7 (8.0)
12-24	6.7 (0.6)	41.3 (7.6)	100.3 (11.9)
24-48	0.0 (0)	11.0 (4.4)	18.0 (1.0)
>48	0.0 (0)	1.3 (1.2)	0.3 (0.6)

The mean individual inbreeding coefficient for the Kurdish horse was less than that of the Persian Arabian (Figure 4.3), both of which were less than that of the Thoroughbred. Variation among samples was greatest in the Iranian samples, and notably in the Persian Arabian.

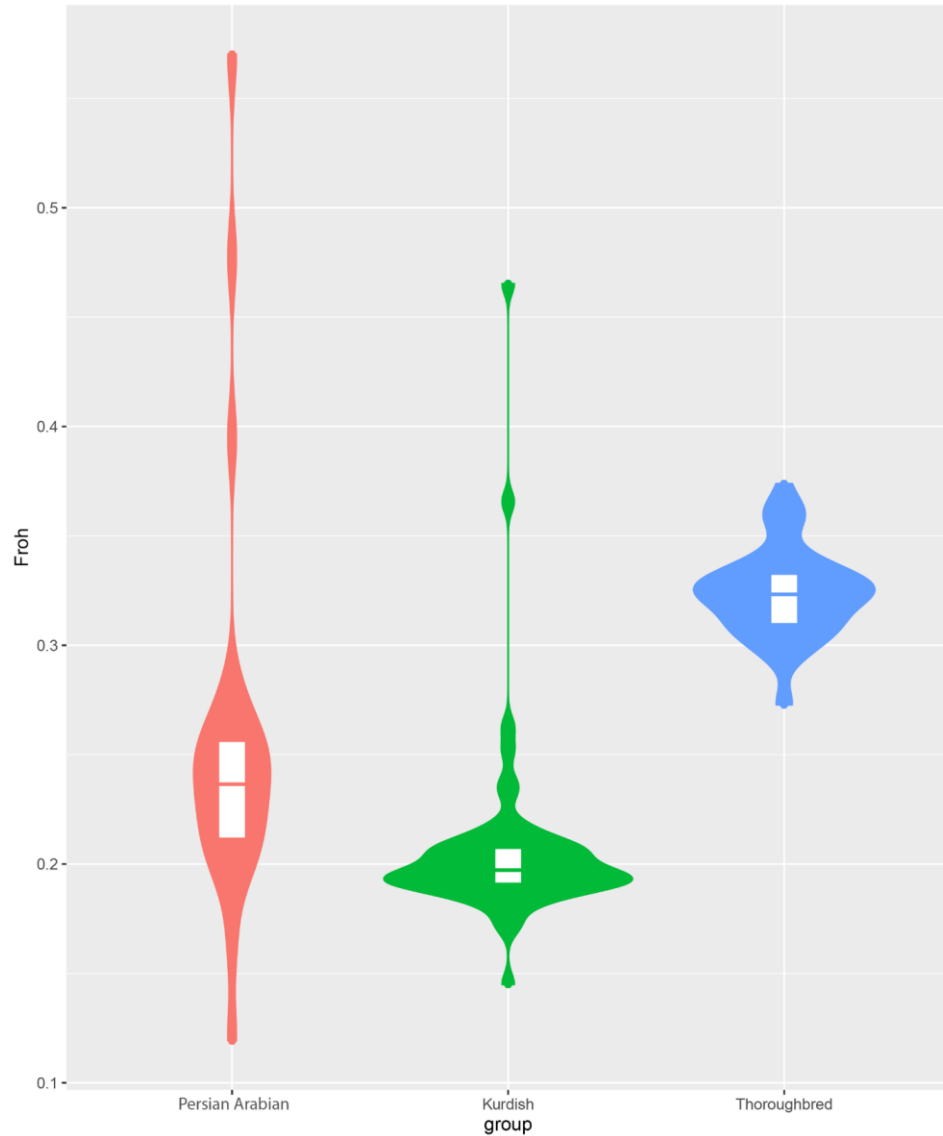


Figure 4.3 Violin plot of mean, quartiles and the frequency (the width of the plot) of the ROH-based inbreeding coefficient ( $F_{ROH}$ ) for each breed group.

The violin plot identifies three outliers in the Persian Arabian group as well as two in the Kurdish population. The same individuals were identified as outliers by PLINK calculations of individual inbreeding. The pedigree of the three outlier Persian Arabian horses available from the studbook confirmed presence of common ancestors in both their paternal and maternal lines and conformed with their higher inbreeding values. For the two



outlier Kurdish horses, however, limited pedigree information was available and the extent of shared ancestry could not be determined. Lastly, the correlation between the  $F_{ROH}$  and the inbreeding values ( $F$ ) calculated in PLINK was 0.765.

#### 4.4.7 Expected Heterozygosity ( $H_E$ )

Expected heterozygosity was greatest in the Kurdish, followed by the Persian Arabian and Thoroughbred (Table 4.4).

Table 4.4 Average Expected Heterozygosity values for each breed group.

Population	Expected Heterozygosity
Kurdish	0.342
Persian Arabian	0.328
Thoroughbred	0.326
Total	0.341

#### 4.4.8 Cluster Analysis

For cluster analyses, burn-in and MCMC iterations of 150,000 and 250,000 (respectively) produced consistent results and converged at the highest value of  $K$  examined ( $K=5$ ). Of the five  $K$  values (i.e. the number of subpopulations hypothesized to exist within the entire sample set) tested in this analysis,  $K=2$  and  $K=3$  were most informative and were further scrutinized in more detail. Little change was observed when 4 or 5 clusters were considered. The Evanno method identified  $K=2$  as the best fit for these data (Table 4.5).

Table 4.5 Results on the comparisons of K=1 to 5 tested by STRUCTURE Harvester. The highlighted row belongs to the K value (=2) that maximizes Delta K per the Evanno method of determining the best fit for the data.

K	Reps	Mean LnP(K)	Stdev LnP(K)	Ln'(K)	Ln''(K)	Delta K
1	2	-1881113.550000	2.899138	—	—	—
2	2	-1843270.200000	9.192388	37843.350000	33267.516667	3619.028712
3	3	-1838694.366667	1369.204007	4575.833333	3094.616667	2.260157
4	4	-1837213.150000	332.264478	1481.216667	1915.700000	5.765588
5	3	-1837647.633333	1541.276556	-434.483333	—	—

Assuming two clusters, the Thoroughbreds were assigned to a single cluster and the Persian horses (Kurdish + Persian Arabian) to the other (Table 4.6, Figure 4.4).

Table 4.6 Average proportion of membership of each pre-defined population in each of the 2 clusters at K=2.

Population	Cluster 1	Cluster 2
Kurdish	0.971	0.029
Persian Arabian	0.971	0.029
Thoroughbred	0.038	0.962

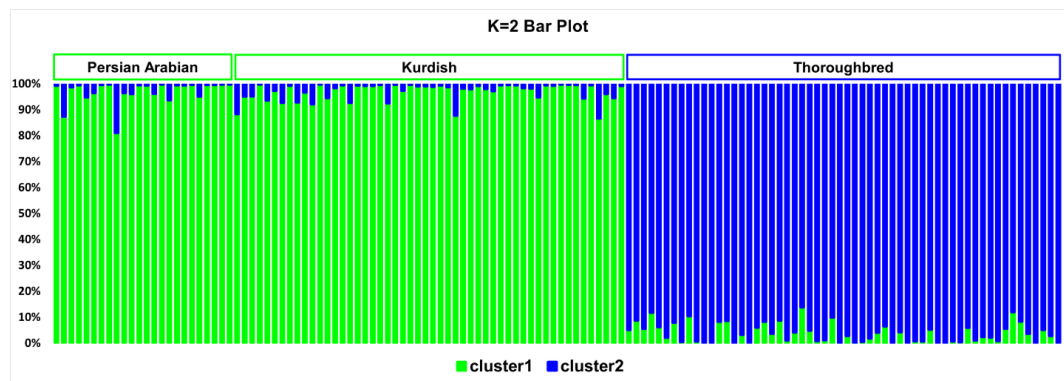


Figure 4.4 Bar plot of the K=2 results. The green color designates cluster 1 (which mostly harbored Kurdish and Persian Arabian horses) and the blue color signifies cluster 2 (which mostly contained Thoroughbred horses). Each individual is represented by a single vertical line broken into K colored segments, with lengths proportional to each of the K inferred clusters.

Assuming three genetic clusters (K=3), the Kurdish and Thoroughbred horses were each assigned to distinct clusters (Figure 4.5). The third genetic cluster was found primarily in Persian Arabians, which still showed shared ancestry with the Kurdish horses (Table 4.7).

Table 4.7 Average proportion of membership of each pre-defined population in each of the 3 clusters at K=3.

Population	Cluster 1	Cluster 2	Cluster 3
Kurdish	0.042	0.942	0.016
Persian Arabian	0.519	0.463	0.017
Thoroughbred	0.023	0.023	0.953

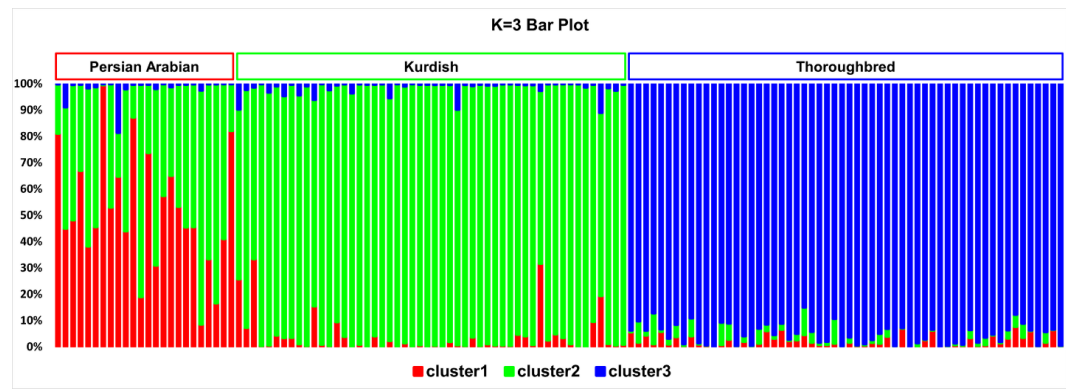


Figure 4.5 Bar plot of the K=3 results. The blue color designates cluster 3 (which mostly harbored Thoroughbred horses), the green color signifies cluster 2 (which mostly contained Kurdish horses and covered a part of Persian Arabian’s genome), and the red color represents cluster 1, which is attributable to Persian Arabian. Each individual is represented by a single vertical line broken into K colored segments, with lengths proportional to each of the K inferred clusters.

Lastly, running the STRUCTURE program excluding Thoroughbred samples from the dataset (at K=2) returned similar clustering values for Kurdish and Persian Arabian groups to the original K=3 results (Table 4.8).

Table 4.8 Average proportion of membership of each pre-defined population (excluding the Thoroughbred samples from the dataset) in each of the 2 clusters at K=2.

Population	Cluster 1	Cluster 2
Kurdish	0.969	0.031
Persian Arabian	0.457	0.543

For a comparison of outliers/overlapping individuals between Kurdish and Persian Arabian samples, we first identified the individuals in the PC plot that were positioned between the two main breed clusters and ranked them in the order of being closest to the opposing population (based on their PC2 values). Then we examined the individual Structure output for each of the individuals that were intermediate between the two breeds and ranked them in the order of having the highest membership in the opposing cluster. The number and ranking of the outlier/overlapping individuals were identical between the two analyses (6 Persian Arabian and 4 Kurdish individuals). All horses that appear intermediate between the two sets of Persian horse samples were assigned more strongly to the cluster representing the other breed.

#### **4.5 Discussion**

Several genetic studies have incorporated horses from Iranian populations (Amirinia, Seyedabadi, Banabazi, & Kamali, 2007; Evrigh, Omri, Boustan, Seyedsharifi, & Vahedi, 2018; Khanshour et al., 2013; Petersen et al., 2013; Rafeie, Amirinia, Javaremi, Mirhoseini, & Amirmozafari, 2011; Rahimi-Mianji, Nejati-Javaremi, & Farhadi, 2015; Sadeghi et al., 2019; Shahsavarani & Rahimi-Mianji, 2010), this is the first, however, to have focused specifically on the Persian Kurdish horse. With respect to our hypothesis, we

found evidence that the Kurdish horse population is distinct from but shares ancestry with the Persian Arabians. Additionally, although some individual horses have genomic evidence of inbreeding, overall both Persian breeds appear to be more diverse than the sample of American Thoroughbreds to which they were compared. As our Thoroughbred sample set was limited to a small number of farms in Kentucky, the Thoroughbred analyses in this report may not be generalizable to the whole population of Thoroughbred, but were intended to serve as a comparison to the two Iranian populations. Overall, the results of the present study serve as baseline data characterizing the diversity and genetic make-up of these breeds to allow the development of conservation strategies to achieve sound breeding goals while maintaining genetic diversity of these breeds.

#### **4.5.1 Analyses of Population Structure**

Our hypothesis that the two Persian breeds are distinct was supported by significant pairwise  $F_{ST}$  values and AMOVA analyses support the distinction of the Kurdish and Persian Arabian samples. As visualized by the PC plot and cluster analyses, however, there is evidence of gene flow between the two breeds. There are two explanations for this observation. The first is that Persian Arabians might have originated from the Kurdish horse and thus Arabians maintain genomic ties to the Kurdish horses resulting from that founding event, while the Kurdish population has independently evolved. The alternative explanation, considering the Kurdish population may have originated from the Persian Arabian, is based upon the Persian Arabian being a diverse and large breed, relative to the Kurdish horse. A founder effect could have led to the isolation of a portion of this diversity into a new genetic group, that led to, or integrated with Kurdish horses. Gene flow from the Persian Arabian to the Kurdish horse may have resulted from recent population

admixture. At the time when the first studbooks for Persian Arabian horses were established around 1976, there was no official breed registry for Kurdish horses; it may be that horse owners with a Kurdish-Persian Arab crossbred or even a Kurdish horse, preferred to have their horses registered with a breed registry, rather than leaving it officially unknown. This would lead to the presence of Kurdish characteristics among Persian Arabian horse populations. Nevertheless, a caveat of cluster analyses and measures of relationships is that directionality of gene flow is not defined; regardless, these data support gene flow between the two Persian populations.

Cluster and PC analyses both suggested that the Kurdish horses are genetically more homogeneous than the Persian Arabians. The observation of higher diversity in the Arabian population as compared to the Thoroughbred is in agreement with reports by Khanshour *et al.* (2013) (Khanshour et al., 2013), Sadeghi *et al.* (2018) (Sadeghi et al., 2018) and is concordant with the latest population study on Arabian horses, which identified a high degree of genetic variation and complex ancestry of those horses from the Middle East, including the Persian group (Cosgrove et al., 2020). Diversity of the Persian Arabians in this study is supported by the evidence of gene flow with the Kurdish horse in cluster analysis. Given the observation that the inbreeding at the population level is not significantly different from zero, it may be interpretable that adverse effects of inbreeding depression is less of a concern in the studied populations and there is still a good wealth of genetic diversity which can be used towards breeding if proper management is applied. This is further supported by the analysis of runs of homozygosity. Despite their relatively small population size and likely ascertainment bias of SNP loci favoring detection of rare variants in the Thoroughbred, the Thoroughbred showed a greater proportion of long (>6

Mb) runs of homozygosity, which were less numerous in the Persian Arabian and absent in the Kurdish horse.

#### **4.5.2 Analyses of Individual Diversity**

With a small population size and lack of formal record keeping, it was of interest to calculate individual estimates of inbreeding based upon both heterozygosity (PLINK) and runs of homozygosity (detectRuns). Although gross calculations of average individual inbreeding present Persian Arabians to be more inbred than Kurdish and Thoroughbred horses, the overall inbreeding level of the Persian Arabian population appears to be driven by the presence of three outliers. Averaging the inbreeding coefficients of the Persian Arabian samples without those outliers, brings the mean inbreeding of the Persian Arabian population lower than that of the Thoroughbred and Kurdish populations. Nevertheless, inbreeding has always been an apparent issue among Persian Arabian horse breeders, as the majority of Persian Arabian lines are descendants of two famous stallions, namely Haddad and Samarghand. Despite considerable inbreeding at the individual level, however, the population has still maintained a good diversity, as implied from their heterozygosity values. Although, their high genetic diversity ensures a healthy population from the genetic standpoint, it is advisable to the breeders to relieve individual inbreeding in their breeding practices, to avoid negative consequences of inbreeding depression. The results also provide another piece of evidence that the Kurdish horse is relatively diverse. This might be due to the fact that the population of Kurdish horses has been mostly shaped by natural selection over a long time rather than human artificial selection.

## 4.6 Conclusions

- There is evidence that the Kurdish horse population forms a distinct genetic cluster with some individuals showing mixed ancestry.
- There is evidence of gene flow between the Persian Arabian and Kurdish horses.
- The overall diversity parameters in the Kurdish horses resembled that of Thoroughbreds. Kurdish horses are a smaller population and it may be appropriate to develop conservation strategies and sound breeding goals to maintain their genetic diversity.



## REFERENCES

- Aberle, K. S., Hamann, H., Drögemüller, C., & Distl, O. (2004). Genetic diversity in German draught horse breeds compared with a group of primitive, riding and wild horses by means of microsatellite DNA markers. *Animal genetics*, 35(4), 270-277.
- Amirinia, C., Seyedabadi, H., Banabazi, M. H., & Kamali, M. A. J. P. J. o. B. S. (2007). Bottleneck Study and Genetic Structure of Iranian Caspian Horse. *10*(9), 1540-1543.
- Behara, A., Colling, D., Cothran, E., & Gibson, J. (1998). *Genetic relationships between horse breeds based on microsatellite data: applications for livestock conservation*. Paper presented at the Proceedings of the 6 th world congress on genetics applied to livestock production, Armidale, Australia.
- Bigi, D., Zambonelli, P., Perrotta, G., & Blasi, M. (2010). The Ventasso Horse: genetic characterization by microsatellites markers. *Italian Journal of Animal Science*, 6(1s), 50-52.
- Biscarini, F., Cozzi, P., Gaspa, G., & Marras, G. (2018). detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes. In: CRAN (The Comprehensive R Archive Network).
- Bjørnstad, G., Nilsen, N., & Røed, K. (2003). Genetic relationship between Mongolian and Norwegian horses? *Animal genetics*, 34(1), 55-58.
- Bowling, A. T., & Ruvinsky, A. (2000). *The genetics of the horse*: CAB International.
- Boyko, A. R., Brooks, S. A., Behan-Braman, A., Castelhano, M., Corey, E., Oliveira, K. C., . . . Ainsworth, D. M. J. B. g. (2014). Genomic analysis establishes correlation between growth and laryngeal neuropathy in Thoroughbreds. *15*(1), 1-9.
- Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller, K. E., . . . Vonholdt, B. M. J. P. b. (2010). A simple genetic architecture underlies morphological variation in dogs. *8*(8), e1000451.
- Brooks, S., Makvandi-Nejad, S., Chu, E., Allen, J., Streeter, C., Gu, E., . . . Sutter, N. J. A. G. (2010). Morphological variation in the horse: defining complex traits of body size and shape. *41*, 159-165.
- Burns, E. N., Bordbari, M. H., Mienaltowski, M. J., Affolter, V. K., Barro, M. V., Gianino, F., . . . Katzman, S. A. J. A. g. (2018). Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *49*(6), 564-570.
- Canon, J., Checa, M., Carleos, C., Vega-Pla, J., Vallejo, M., & Dunner, S. (2000). The genetic structure of Spanish Celtic horse breeds inferred from microsatellite data. *Animal genetics*, 31(1), 39-48.
- Coates, J. W., & McFee, R. C. J. T. C. V. J. (1993). Congenital lordosis in three Haflinger foals. *34*(8), 496.
- Cook, D., Gallagher, P., & Bailey, E. (2010). Genetics of swayback in American Saddlebred horses. *Animal Genetics*, 41, 64-71.
- Cook, D. G. (2014). *Use of genomic tools to discover the cause of champagne dilution coat color in horses and to map the genetic cause of extreme lordosis in American Saddlebred horses*: University of Kentucky.
- Cosgrove, E. J., Sadeghi, R., Schlamp, F., Holl, H. M., Moradi-Shahrbabak, M., Miraei-Ashtiani, S. R., . . . Stefaniuk-Szmukier, M. J. S. r. (2020). Genome diversity and the origin of the Arabian horse. *10*(1), 1-13.

- Diakonoff, I. M. (1956). *The History of Media: Moscow-Leningrad*.
- Earl, D. A. J. C. g. r. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *4*(2), 359-361.
- Eberlein, A., Takasuga, A., Setoguchi, K., Pfuhl, R., Flisikowski, K., Fries, R., . . . Kühn, C. J. G. (2009). Dissection of genetic factors modulating fetal growth in cattle indicates a substantial role of the non-SMC condensin I complex, subunit G (NCAPG) gene. *183*(3), 951-964.
- Ellegren, H., Johansson, M., Sandberg, K., & Andersson, L. (1992). Cloning of highly polymorphic microsatellites in the horse. *Animal genetics*, *23*(2), 133-142.
- Evanno, G., Regnaut, S., & Goudet, J. J. M. e. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *14*(8), 2611-2620.
- Evrigh, N. H., Omri, M., Boustan, A., Seyedsharifi, R., & Vahedi, V. J. J. o. E. V. S. (2018). Genetic Diversity and Structure of Iranian Horses' Population Based on Mitochondrial Markers. *64*, 107-111.
- Excoffier, L., & Lischer, H. E. J. M. e. r. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *10*(3), 564-567.
- Fages, A., Hanghøj, K., Khan, N., Gaunitz, C., Seguin-Orlando, A., Leonardi, M., . . . Albizuri, S. J. C. (2019). Tracking five millennia of horse management with extensive ancient genome time series. *177*(6), 1419-1435. e1431.
- Gallagher, P., Morrison, S., Bernoco, D., & Bailey, E. (2003). Measurement of back curvature in American Saddlebred horses: Structural and genetic basis for early-onset lordosis. *Journal of Equine Veterinary Science*, *2*(23), 71-76.
- Golden Helix, I. SNP & Variation Suite™. Version 8. Retrieved from <http://www.goldenelix.com>
- Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., . . . Steinberg, S. J. N. g. (2008). Many sequence variants affecting diversity of adult human height. *40*(5), 609-615.
- He, S., Zhang, L., Li, W., & Liu, M. J. A. b. (2015). BIEC2-808543 SNP in the LCORL gene is associated with body conformation in the Yili horse. *26*(4), 289-291.
- Hill, E. W., Gu, J., Eivers, S. S., Fonseca, R. G., McGivney, B. A., Govindarajan, P., . . . MacHugh, D. J. P. o. (2010). A sequence polymorphism in MSTN predicts sprinting ability and racing stamina in thoroughbred horses. *5*(1), e8645.
- Jones, P., Chase, K., Martin, A., Davern, P., Ostrander, E. A., & Lark, K. G. J. G. (2008). Single-nucleotide-polymorphism-based association mapping of dog stereotypes. *179*(2), 1033-1044.
- Kakoi, H., Tozaki, T., & Gawahara, H. (2007). Molecular analysis using mitochondrial DNA and microsatellites to infer the formation process of Japanese native horse populations. *Biochemical genetics*, *45*(3-4), 375-395.
- Kalbfleisch, T. S., Rice, E. S., DePriest Jr, M. S., Walenz, B. P., Hestand, M. S., Vermeesch, J. R., . . . Saremi, N. F. J. C. b. (2018). Improved reference genome for the domestic horse increases assembly contiguity and composition. *1*(1), 197.

- Khanshour, A., Conant, E., Juras, R., & Cothran, E. G. J. J. o. H. (2013). Microsatellite analysis of genetic diversity and population structure of Arabian horse populations. *104*(3), 386-398.
- Kim, J.-J., Lee, H.-I., Park, T., Kim, K., Lee, J.-E., Cho, N. H., . . . Han, B.-G. J. J. o. h. g. (2010). Identification of 15 loci influencing height in a Korean population. *55*(1), 27-31.
- Kraut, N., Snider, L., Chen, C.-M. A., Tapscott, S. J., & Groudine, M. J. T. E. J. (1998). Requirement of the mouse *I-mfa* gene for placental development and skeletal patterning. *17*(21), 6276-6288.
- Krueger, F. J. U. h. w. b. b. a. u. p. t. g. (2012). Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisulfite-Seq) libraries.
- Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., . . . Raychaudhuri, S. J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *467*(7317), 832-838.
- Li, H., & Durbin, R. J. B. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *26*(5), 589-595.
- Li, H. J. a. p. a. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
- Li, J., Chen, C., Bi, X., Zhou, C., Huang, T., Ni, C., . . . Duan, S. J. G. (2017). DNA methylation of *CMTM3*, *SSTR2*, and *MDFI* genes in colorectal cancer. *630*, 1-7.
- Luis, C., Juras, R., Oom, M., & Cothran, E. (2007). Genetic diversity and relationships of Portuguese and other horse breeds based on protein and microsatellite loci variation. *Animal genetics*, *38*(1), 20-27.
- Makvandi-Nejad, S., Hoffman, G. E., Allen, J. J., Chu, E., Gu, E., Chandler, A. M., . . . Brooks, S. A. J. P. O. (2012). Four loci explain 83% of size variation in the horse. *7*(7), e39929.
- McCue, M. E., Bannasch, D. L., Petersen, J. L., Gurr, J., Bailey, E., Binns, M. M., . . . Hill, E. W. (2012). A high density SNP array for the domestic horse and extant *Perissodactyla*: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet*, *8*(1), e1002451.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., . . . Daly, M. J. G. r. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *20*(9), 1297-1303.
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. J. G. b. (2016). The ensembl variant effect predictor. *17*(1), 1-14.
- McMaster, M. J., & Singh, H. J. J. (1999). Natural history of congenital kyphosis and kyphoscoliosis. A study of one hundred and twelve patients. *81*(10), 1367-1383.
- Metzger, J., Schrimpf, R., Philipp, U., & Distl, O. J. P. o. (2013). Expression levels of *LCORL* are associated with body size in horses. *8*(2), e56497.
- Mostafavi, A., Fozzi, M. A., Koshkooieh, A. E., Mohammadabadi, M., Babenko, O. I., & Klopenko, N. I. J. A. S. A. S. (2019). Effect of *LCORL* gene polymorphism on body size traits in horse populations. *42*.

- Ovchinnikov, I. V., Dahms, T., Herauf, B., McCann, B., Juras, R., Castaneda, C., & Cothran, E. G. J. P. o. (2018). Genetic diversity and origin of the feral horses in Theodore Roosevelt National Park. *13*(8), e0200795.
- Petersen, J. L., Mickelson, J. R., Cothran, E. G., Andersson, L. S., Axelsson, J., Bailey, E., . . . Brama, P. J. P. o. (2013). Genetic diversity in the modern horse illustrated from genome-wide SNP data. *8*(1), e54997.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. J. T. A. j. o. h. g. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *81*(3), 559-575.
- Rafeie, F., Amirinia, C., Javaremi, A. N., Mirhoseini, S. Z., & Amirmozafari, N. J. A. J. o. B. (2011). A study of patrilineal genetic diversity in Iranian indigenous horse breeds. *10*(75), 17347-17352.
- Rahimi-Mianji, G., Nejati-Javaremi, A., & Farhadi, A. J. R. j. o. g. (2015). Genetic diversity, parentage verification, and genetic bottlenecks evaluation in Iranian turkmen horse. *51*(9), 916-924.
- Raudsepp, T., Finno, C. J., Bellone, R. R., & Petersen, J. L. J. A. g. (2019). Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post-genome era. *50*(6), 569-597.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. J. N. b. (2011). Integrative genomics viewer. *29*(1), 24-26.
- Rooney, J., & Prickett, M. (1967). Congenital lordosis of the horse. *The Cornell Veterinarian*, *57*(3), 417-428.
- Sadeghi, R., Moradi-Shahrbabak, M., Miraei Ashtiani, S. R., Schlamp, F., Cosgrove, E. J., & Antczak, D. F. J. J. o. H. (2018). Genetic Diversity of Persian Arabian Horses and Their Relationship to Other Native Iranian Horse Breeds. *110*(2), 173-182.
- Sadeghi, R., Moradi-Shahrbabak, M., Miraei Ashtiani, S. R., Schlamp, F., Cosgrove, E. J., & Antczak, D. F. J. J. o. H. (2019). Genetic diversity of Persian Arabian horses and their relationship to other native Iranian horse breeds. *110*(2), 173-182.
- Sambrook, J., & Russell, D. W. J. C. S. H. P. (2006). Purification of nucleic acids by extraction with phenol: chloroform. *2006*(1), pdb. prot4455.
- Santagostino, M., Khoriauli, L., Gamba, R., Bonuglia, M., Klipstein, O., Piras, F. M., . . . Mazzagatti, A. J. B. g. (2015). Genome-wide evolutionary and functional analysis of the Equine Repetitive Element 1: an insertion in the myostatin promoter affects gene expression. *16*(1), 1-16.
- Schaefer, R. J., Schubert, M., Bailey, E., Bannasch, D. L., Barrey, E., Bar-Gal, G. K., . . . Fries, R. J. B. g. (2017). Developing a 670k genotyping array to tag~ 2M SNPs across 24 horse breeds. *18*(1), 565.
- Shahcheraghi, G. H., & Hobbi, M. J. J. o. P. O. (1999). Patterns and progression in congenital scoliosis. *19*(6), 766.
- Shahsavarani, H., & Rahimi-Mianji, G. (2012). Analysis of genetic diversity and estimation of inbreeding coefficient within Caspian horse population using microsatellite markers. *African Journal of Biotechnology*, *9*(3).
- Shahsavarani, H., & Rahimi-Mianji, G. J. A. J. o. B. (2010). Analysis of genetic diversity and estimation of inbreeding coefficient within Caspian horse population using microsatellite markers. *9*(3).

- Signer-Hasler, H., Flury, C., Haase, B., Burger, D., Simianer, H., Leeb, T., & Rieder, S. J. P. o. (2012). A genome-wide association study reveals loci influencing height and other conformation traits in horses. *7*(5), e37282.
- Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., . . . Mazon, Y. J. C. p. i. b. (2016). The GeneCards suite: from gene data mining to disease genome sequence analyses. *54*(1), 1.30. 31-31.30. 33.
- Tetens, J., Widmann, P., Kühn, C., & Thaller, G. J. A. g. (2013). A genome-wide association study indicates LCORL/NCAPG as a candidate locus for withers height in German Warmblood horses. *44*(4), 467-471.
- Tozaki, T., Sato, F., Ishimaru, M., Kikuchi, M., Kakoi, H., Hirota, K.-I., & Nagata, S.-I. J. o. e. s. (2016). Sequence variants of BIEC2-808543 near LCORL are associated with body composition in Thoroughbreds under training. *27*(3), 107-114.
- Wade, C., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Inslan, F., . . . Bellone, R. J. S. (2009). Genome sequence, comparative analysis, and population genetics of the domestic horse. *326*(5954), 865-867.
- Wallner, B., Palmieri, N., Vogl, C., Rigler, D., Bozlak, E., Druml, T., . . . Tetens, J. J. C. B. (2017). Y chromosome uncovers the recent oriental origin of modern stallions. *27*(13), 2029-2035. e2025.
- Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., . . . Hall, A. S. J. N. g. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *40*(5), 575-583.
- Williams, F., McCall, I. W., O'Brien, J. P., & Park, W. M. J. C. r. (1982). Severe kyphosis due to congenital dorsal hemivertebra. *33*(4), 445-452.
- Yousefi-Mashouf, N., Mehrabani-Yeganeh, H., Nejati-Javaremi, A., & Maloufi, F. (2020). A Novel Approach to Establish Breed Type and Standards for an Equine Breed: Persian Kurdish Horse. *Journal of Agricultural Science and Technology*, *22*(5).
- Yousefi Mashouf, N., Mehrabani Yeganeh, H., Nejati Javaremi, A., Maloufi, F. J. J. o. A. S., & Technology. (2020). A novel approach to establish breed type and standards for an equine breed: Persian Kurdish horse. *22*(5), 1219-1233.

## VITA

### Education

- Sep 2013 - Sep 2016 **University of Tehran**  
Master's degree  
Animal Science – Animal Breeding
- Sep 2009- Jul 2013 **University of Tehran**  
Bachelor's degree  
Animal Science  
Distinguished graduate (Second-rank among 23 graduates)  
Focused on equine sciences

### Publications

#### Book:

- 2016 Navid YousefiMashouf, Mohammad FarzanehFar (2016). *Breed Type and Standards of the Persian Kurdish Horse*. Sarmadi Publications (in Persian language), Tehran, Iran. 55 pages.

#### Scholarly Publications:

- Jul 2023 Navid YousefiMashouf, Theodore Kalbfleisch, Kathryn Graves and Ernest Bailey (2023, July). Investigating the effect of chromosome 20 on lordosis in Saddlebred horses. In *39<sup>th</sup> Conference of the International Society of Animal Genetics (July 2-7, 2023)*. Cape Town, South Africa.
- Jan 2023 Navid YousefiMashouf, Ariana Spina, John Eberth, Rebecca Bellone, Kathryn Graves and Ernest Bailey (2023, January). Investigation of Myostatin gene variants in Thoroughbred and related horse breeds. In *Plant and Animal Genome 30 Conference (January 13-18, 2023)*. PAG.
- Jul 2022 Navid YousefiMashouf, Theodore Kalbfleisch, Kathryn Graves and Ernest Bailey (2022) Genetic Basis for Juvenile Onset Lordosis in Saddlebred Horses. In *13<sup>th</sup> Havemeyer International Horse Genome Workshop*, Cornell University, Ithaca, NY.
- Jul 2022 Navid YousefiMashouf, Ariana Spina, John Eberth, Rebecca Bellone, Kathryn Graves and Ernest Bailey (2022) Linkage disequilibrium for *MSTN* variants in several horse breeds. In *13<sup>th</sup> Havemeyer International Horse Genome Workshop*, Cornell University, Ithaca, NY.
- Feb 2021 Yousefi-Mashouf, N., Mehrabani-Yeganeh, H., Nejati-Javaremi, A., Bailey, E., & Petersen, J. L. (2021). Genomic comparisons of Persian Kurdish, Persian Arabian and American Thoroughbred horse populations. *PloS one*, 16(2), e0247123.
- Sep 2020 Yousefi Mashouf, N., Mehrabani Yeganeh, H., Nejati Javaremi, A., & Maloufi, F. (2020). A Novel Approach to Establish Breed Type and Standards for an Equine Breed: Persian Kurdish Horse. *Journal of Agricultural Science & Technology*, 22(5).

- Feb 2020 Dunuwille, W. M., YousefiMashouf, N., Balasuriya, U. B., Pusterla, N., & Bailey, E. (2020). Genome-wide association study for host genetic factors associated with equine herpesvirus type-1 induced myeloencephalopathy. *Equine Veterinary Journal*.
- Jan 2020 YousefiMashouf, N., Kalbfleisch, T., Graves, K., Bailey, E. (2020, January). Comparison of Horses with Juvenile Onset Lordosis to Normal Horses Using Whole Genome Sequence. In *Plant and Animal Genome XXVIII Conference (January 11-15, 2020)*. PAG.
- Jul 2019 N. YousefiMashouf, J. L. Petersen, H. Mehrabani Yeganeh, A. Nejati Javaremi, T. S. Kalbfleisch, M. Bagher Zandi, and E. Bailey (2019, July). Population structure analysis of the Persian horse breeds and their comparison to worldwide populations using genome-wide SNP genotypes. *Proceedings of the 37th International Conference on Animal Genetics, Lleida, Spain (2019)*.
- Jan 2019 YousefiMashouf, N., Bailey, E., Petersen, J. L., Yeganeh, H. M., & Javaremi, A. N. (2019, January). Genomic Comparisons of the Persian Kurdish Horse to Persian Arabian and American Thoroughbred Populations. In *Plant and Animal Genome XXVII Conference (January 12-16, 2019)*. PAG.
- 2017 Maghsoodi, S. M., Mehrabani, Y. H., Nejati, J. A., & Yousefi Mashouf, N. (2017). Investigating population structure and identifying signatures of selection in Iranian Kurdish and Arabian horses. *Iranian Journal of Animal Science*. 48(3): 429-438.
- Sep 2016 Yousefi Mashouf, N. (2016). Master's Dissertation (in Persian language): Phenotypic and Genetic Characterization of the Iranian Kurdish Horse. *Department of Animal Sciences, University of Tehran*.
- 2014 Promerová, M., Andersson, L.S., Juras, R., Penedo, M.C.T., Reissmann, M., Tozaki, T., Bellone, R., Dunner, S., Hořín, P., Imsland, F., Imsland, P., Mikko, S., Modrý, D., Roed, K.H., Schwochow, D., Vega-Pla, J.L., Mehrabani-Yeganeh, H., Yousefi-Mashouf, N., Cothran, E.G., Lindgren, G., & Andersson, L. (2014). World-wide frequency distribution of the 'Gait keeper' mutation in the *DMRT3* gene. *Animal Genetics*. 45(2): 274–282.
- 2011 Yousefi Mashouf, N., Mehrabani Yeganeh, H., & Nejati Javaremi, A. (2011). Introduction of Partizan Software and the Nationwide Network of Iranian Horsemen (in Persian language). *Proceeding of the First National Congress of Equine Industry*. Golestan: University of Gonbad-Kavous.

### **Awards and Distinctions**

- 2009 - 2012 Received separate official appreciations for managing the Equine Division of the Animal Science Department in 3 consecutive Open Days.
- 2010 - 2014 Received Separate official appreciations for speech and active contribution in 3 national festivals of Caspian Horse.

Nov 2011	Received official appreciation for accepted paper and its oral presentation at the First National Congress of Equine Industry which was held at the University of Gonbad-Kavous, Golestan, Iran.
Sep 2012	Horse Identification Document was elected as premier project in the Third National Festival of “From Science to Practice”.
2013	Distinguished graduate in the Bachelor’s program (ranked second among 23 classmates)
Jul 2013	As a distinguished graduate, awarded the Waiver of the National Graduate Admission Exam, thereby directly enrolled to the Master’s program in Animal Breeding.
Sep 2013 - Sep 2016	First-rank student in the Master’s program (among nine classmates).
2016	Received official appreciation for active contribution to the National Breed Show of Kurdish Horse, Kermanshah, Iran.
2016	Received official appreciation for speech at and management of the First National Conference of Kurdish Horse, Bijar, Kordestan, Iran.
March 2017	Full scholarship for graduate studies at PhD level as Graduate Research Assistant, Department of Veterinary Science, University of Kentucky.
Jul 2022	Travel Bursary to attend the 13 <sup>th</sup> Havemeyer International Horse Genome Workshop, Cornell University, Ithaca, NY.
Jul 2022	1 <sup>st</sup> place Award of Poster Presentation at the 13 <sup>th</sup> Havemeyer International Horse Genome Workshop, Cornell University, Ithaca, NY.
Jul 2022	Approved eligibility for Permanent Residency of the United States based on Academic Achievements through National Interest Waiver program (NIW).
Jul 2023	Travel Bursary to attend the 39 <sup>th</sup> Conference of the International Society of Animal Genetics; Cape Town, South Africa, 2-7 July 2023.