Genetic Factors Underlying Disease and Performance Traits in Standardbreds

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Annette Marie McCoy

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Molly E. McCue, Adviser and Troy N. Trumble, Co-Adviser

May 2014

Acknowledgements

Thank you to everyone who has supported me throughout this process. This is not a journey that I could have taken alone – nor would I have wanted to. My life is richer for having known and worked with you all.

My advisors and committee members: Thank you all for your unflagging support and encouragement throughout my time here. I have learned more from you than I could have imagined. In particular:

Dr. Molly McCue: Thank you for giving me the freedom to grow and the support to be successful. I am unquestionably a better scientist, writer, and teacher because of your guidance, but knowing that you care about me as a person has been your greatest gift.

Dr. Troy Trumble: You help me to remember that I can be both a surgeon and a scientist. Thank you so much for supporting me in my goal to do just that.

Dr. James Mickelson: A willing ear and a critical eye – thank you for sharing both with me without reservation whenever I needed them.

Dr. Cathy Carlson: Thank you for making this all possible – without your belief in me as a T32 candidate, it would have been much harder to even get in the door.

Dr. Denis Clohisy: Thank you for pinch hitting in the ninth inning. I so appreciate your generosity in being a part of all this despite your already packed schedule.

My lab mates and fellow graduate students: Thank you for letting me bounce ideas off of you, commiserate with you when things got frustrating, celebrate successes with you, and learn from you every day.

Our fabulous technicians: Thank you for showing me the ropes and being patient when I asked the same questions over and over. A special thank you to Shea Anderson as well as our undergraduates/lab assistants Leah Freilich and Nicole Tate for providing technical support for the projects reported in this thesis.

To my family: You may not have entirely understood why I wanted to get a PhD when I was "already a doctor," but you never stopped believing in me or encouraging me. Thank you for a lifetime of support and love. In particular, thank you to my wonderful husband Jeffrey Fox – you have been my rock through the highs and lows of the past few years, and without your commitment to me and my goals, I never would have made it this far. Thank you for telling me that I could do it, even when I wasn't sure I could. Finally, to my son Luke, who arrived halfway through this adventure: you are the light of my life, and your smile reminds me that no matter what happens at work, my most important job is always waiting for me at home.

Dedication

Willard George Sampson (November 1927 – December 2013) and
Elizabeth Ann Kimmel Sampson

Dearest grandparents, you helped to teach me how to read a book, sew a seam, solve algebraic equations. More than that, you engendered a love of learning new things and a belief that I could accomplish any goal that I set for myself. For as long as I can remember, you have been interested in what I was doing and how it fit into the world, asking insightful questions and listening with respect to my responses. I could never ask for better role models. So, even though Grandpa is not here to read it, I dedicate this culmination of “what I’ve been doing” to you both with love.

Abstract **Annette M. McCoy** **342 words**

Many diseases and performance characteristics of the horse are considered to be "complex" traits because they are influenced by both genetic and environmental factors. Furthermore, many are polygenic in nature, reflecting the combined effects of multiple genes. Traditional methodological approaches, such as family linkage analysis and candidate gene sequencing are not ideal for identifying the multiple interacting alleles underlying complex/polygenic traits. An alternative investigational approach is needed that can account for environmental risk factors, issues related to population structure in large study cohorts, and epistatic interactions. In the work presented here, whole-genome approaches, including genome-wide association (GWA) analysis, whole-genome sequencing (WGS), and high-throughput genotyping, were used to investigate the genetic factors underlying three complex traits in Standardbred horses, a breed primarily used for harness racing. These were 1) osteochondrosis (OC; a disease of young horses in which the cartilage at the end of long bones does not form normally); 2) pacing (an alternative pattern of locomotion); and 3) performance (using speed as the phenotype).

GWA analysis identified chromosomal regions of association for all three traits of interest, although the significance of the findings for speed was marginal, reflecting the challenge of appropriately phenotyping a complex trait such as performance. WGS performed in eighteen horses identified thousands of variants within chromosomal regions of association identified for OC and pacing, of which a small fraction were predicted to have functional effect. These variants were prioritized and a subset was selected for high-throughput genotyping in the study cohorts (180 horse phenotyped for OC, 500 phenotyped for gait). A few of the markers selected for OC were moderately associated with disease status, while the majority of the markers selected for gait were highly associated with this trait. A crucial next step for interpreting these data will be trying to understand the potential interactions between markers, using a combination of pathway analysis and random forest analysis. Knowledge of gene variants that affect complex traits in the horse – and how they interact with each other – may help reduce the incidence of disease and assist selection for desirable characteristics.

Table of Contents

List of Tables

Chapter 8

Appendix 1

List of Figures

Chapter 1

Introduction and Literature Review

History and Use of the Standardbred Horse

The Standardbred breed traces its ancestry back to a Thoroughbred stallion, Messenger, who was imported from England to the United States in 1788.[1;2] When bred to trotting mares, Messenger produced fast trotter offspring, and the line became popular throughout North America. Messenger's great-grandson Hambletonian (or Hambletonian 10), out of a Norfolk Trotter mare, was the foundation stallion for Standardbreds, but other breeds that contributed to the Standardbred during its early history were the Narrangansett Pacer (now extinct), the Canadian Pacer (also extinct), the Hackney, and the Morgan.[3] This blend resulted in a horse that was of similar build to a Thoroughbred, but slightly shorter, longer, and more muscled, with a large head and docile and willing personality.[1] The breed was widely used for light work in harness, including pulling wagons and sleighs, and informal community competitions eventually developed into organized harness races.[3]

The performance standards that led to the name "Standardbred" were established by the National Association of Trotting Horse Breeders in 1879[2], although harness racing had been popular since the early- to mid-1800s.[1;3] The original requirements for registration as a Standardbred included trotting a mile in 2 minutes 30 seconds for stallions – slightly slower if pulling a wagon[2] – but the two minute mile soon became the mark of an elite trotter.[4] Modern Standardbreds routinely eclipse this "miracle mile" time[4], and today pedigree, rather than speed, is a prerequisite for breed registration[3], but a "Standard Record" of 2:15 or faster (2:20 for 2-year-olds) is still dictated for actively racing individuals.[5]

While the origins of the Standardbred are in trotting horses, the breed is somewhat unique in that some individuals race using an alternative gait, the pace. The ability to pace is traditionally thought to be the legacy of Spanish blood, via the Narrangansett and Canadian Pacers (both of which were "ambling" breeds developed in the 1700s), although this link has never been conclusively proven.[6;7] Trotting and pacing are both 2-beat, symmetrical gaits, but in the trot the diagonal legs move simultaneously while in the pace the ipsilateral limbs move together. Not surprisingly, the biomechanical differences in these gaits result in measurable differences in performance, most notably in that pacers are faster than trotters.[8;9] On the track, most pacers wear hobbles attached to their harnesses to encourage proper limb movement at race speeds, but the gait is natural to these individuals and is demonstrated in young pacing-bred foals before any training occurs.[1] Breeding practices over the past 100 years have distinctly separated pacing lines from trotting, and the genetic differences between the two are now similar to that between other separate breeds.[10] Notably, despite this, approximately 20% of the offspring of trotter stallions subsequently race as pacers.[10] It is not known whether this is due to natural inclination, training, or a combination of both. It is uncommon for these so-called "double-gaited" horses to perform equally well at both gaits, although a handful of individuals have excelled as both pacers and trotters over the past twenty-five years.[4] A recently described mutation in the gene *DMRT3* (an isoform of the doublesex and mab-3 related transcription factor) on equine chromosome (ECA) 23 appears to be permissive for "gaitedness" across breeds, but this mutation is nearly fixed in Standardbred horses despite the fact that not all Standardbreds naturally pace.[11] This suggests that while this mutation is necessary, it is not sufficient for pacing ability, and thus that other modifying

genetic factors that have yet to be described exist in a subset of the population. Pacers, who tend to race more often starting at a young age[8] and thus may be perceived to have greater early earnings potential, make up approximately eighty percent of the racing Standardbred population in North America[3] and are also popular in Great Britain, Australia, and New Zealand. In contrast, only trotters, with their tendency for longer careers[8], compete in Scandinavia and the rest of continental Europe.

Factors Affecting Performance in the Standardbred Horse

Performance in the Standardbred horse is most commonly measured in terms of race outcomes – for example, fastest speed over a mile; number of wins or order of placing in races; or earnings, either cumulative or per start. With nearly $600 million available in purses in the United States and Canada alone in 2012[4], breeders and owners are always looking for the perfect combination of genetic potential and training that will result in the next big winner. Gait, age, and gender have all been established as fixed factors influencing performance[8;9], but other underlying genetic influences have yet to be defined. In an attempt to improve selection of breeding stock, breeding values based on progeny performance (i.e. number of races started, earnings, best racing time) have been introduced by some national breed associations.[12] However, the reported heritability of performance traits varies widely, which calls into question the utility of these breeding values. Heritability of speed, for example, was reported by Thuneberg-Solonen et al. (1999) to be 0.28 in Finnish Standardbred trotters[13] and by Tolley et al. (1983) to be 0.29 in North American Standardbreds (pacers and trotters pooled)[14], but it ranged from only 0.01-0.18 in a population of German trotters described by Bugislaus et al. (2006) when

age was taken into account.[15] Heritability of placement within a race and earnings are reportedly even lower (0.12 and 0.09, respectively).[13] Investigation into the genes and pathways that underlie performance traits in the Standardbred horse is warranted, however, while this new knowledge can only improve genetic selection schemes, it must be acknowledged that even on an optimal genetic background, external influences will always play a major role in the success of any particular individual. These external factors include not only track and race variables such as track surface, the quality of other individuals in the field, and driver experience[8;13;14], but training- and race-related injuries as well.

When considering the performance potential of a yearling, future problems related to the musculoskeletal system are of significant concern to prospective buyers. A system of repository radiographs is not the standard at North American Standardbred yearling sales as it is for Thoroughbreds, but in Europe, where young stock are frequently required to meet certain criteria to be considered for competition and/or breeding, radiographs of yearlings are routinely obtained. Unfortunately, the clinical significance of radiographic findings can be difficult to determine. In large surveys of Thoroughbred yearlings, the majority of radiographic abnormalities have not been found to correlate with early career race performance.[16-20] However, because of the differences between Thoroughbred and Standardbred racing – i.e. under saddle vs. in harness – these findings cannot be directly extrapolated between breeds. There are few comparably large surveys broadly examining radiographic abnormalities and race performance in Standardbreds. Robert et al. (2006) found that in a population of French Standardbred trotters, radiographic abnormalities directly resulted in 31% of individuals failing to qualify for competition.[21] In contrast,

Courouce-Malblanc et al. (2006) reported that neither qualification nor maximal "index of trot" (calculated as the natural logarithm of earnings per start) of French Standardbred trotters were affected by radiographic abnormalities, but that career longevity was reduced.[22] In both of these surveys, two commonly identified lesions were osteochondrosis (OC) of the tarsus (17% and 10.8% of individuals, respectively) and palmar/plantar osteochondral fragments (POF) of the fetlock (22% and 18.3%, respectively).[21;22] These highly prevalent lesions are the focus of the majority of studies examining radiographic abnormalities and performance in Standardbred horses. POF have been found to adversely affect racing performance when examined independently from other fetlock lesions[21;22], although this can be largely mitigated with surgical removal unless secondary osteoarthritic changes have developed.[23] There are conflicting reports regarding the effects of tarsal OC, with some demonstrating impaired performance in OC-affected horses[24-26] and others finding no significant differences in racing success between affected and unaffected individuals.[24;27-29] Differences in cohort selection and definition of OC between these reports make it difficult to directly compare the results of these studies, but it seems clear that tarsal OC has the potential to negatively impact performance, especially in severe and/or untreated cases.

The Importance of Osteochondrosis in the Standardbred Horse

OC and POF are both classified under the recently proposed umbrella category of "juvenile osteochondral conditions" (JOCC) although their pathophysiology is quite distinct.[30] POF occur at ligamentous attachment sites and are traumatic in origin. In contrast, OC is a focal failure of endochondral ossification, the process by which a

cartilage template becomes bone in the limbs of a growing animal, and is thus developmental in origin. While OC is reported across horse breeds (**Table 1**), Standardbreds are considered predisposed to the condition, especially in the tarsus, with an average reported prevalence of 14.8% in this location.[31-38] By comparison, the average reported prevalence of this particular lesion in Thoroughbreds is 5.3%.[19;35;39-41]

Although, as mentioned above, the significance of tarsal OC lesions in terms of race performance is debated in the literature, OC is a concern to breeders due to the heritable component of the disease. Heritability estimates for tarsal OC in the Standardbred range from 0.19[42] to 0.52[32], suggesting that between 20% and 50% of the risk for developing disease can be attributed to genetic factors. Within individual progeny groups, up to 70% of foals have reportedly been affected by OC.[32] Philipsson, et al. (1993) reported significantly higher incidence in tarsal OC in progeny of Standardbred sires known to be affected themselves.[37] Variation in heritability estimates between populations is to be expected for any trait due to differences in population history, gene frequency, and environmental exposures.[43] This is particularly true for OC since it is a complex disease, with known environmental interactions, and likely has multiple genetic alleles conferring susceptibility.

Despite strong evidence demonstrating the heritable nature of OC, the specific genes and alleles underlying OC risk in the horses are, to date, completely unknown. Identification of these genetic risk factors, in addition to environmental manipulation, will be crucial in efforts to reduce disease prevalence. Although affectation with OC is not yet a disqualifying factor for breeding in the Standardbred as it is in the Dutch Warmblood, another highly predisposed breed, a discussion has begun in race circles

about the ethics of perpetuating this condition and the need for more informed breeding decisions.[44]

Challenges to Mapping Complex Traits in the Horse

Susceptibility to diseases, such as OC, as well as economically important performance traits in the horse, such as gait and speed, are complex, or quantitative, traits. In contrast to simple traits that are governed by Mendelian inheritance of a single gene, complex traits involve contributions from multiple alleles and may have important environmental interactions. This represents a significant challenge when trying to identify specific genes and variants that are involved in the expression of such traits. There are two main theories about the genetic architecture of these traits. The "common trait, common variant" theory suggests that many (hundreds to thousands) of common alleles each contribute a very small effect to the phenotypic expression of a trait. In contrast, the "common trait, rare variant" theory posits that a few rare variants of moderate to large effect determine trait expression.[45;46] Genome-wide association (GWA) studies of many complex diseases/traits in humans and animals have thus far only explained a small proportion of heritability, and few specific genes of importance have been identified.[47;48] Proponents of the "common variant" theory could argue that these studies have been underpowered, and that data from tens or hundreds of thousands of individuals would be needed to fully explain heritability of complex traits. Conversely, since GWA by its nature only uses common variants, those who ascribe to the "rare variant" theory could argue that it completely misses causative variants and that an alternative approach will be needed to find the "missing" heritability. Other factors that have been suggested to play a

role in the expression of complex traits include copy number polymorphisms, gene by environment interactions, parent of origin effects, genetic variation in non-coding RNAs, transgenerational genetic effects, and phenotypic robustness.[49;50] Most of these factors have yet to be explored in depth, in part due to limitations of current technology. Evidence from both empirical and simulated data suggests that it is likely that different genetic architectures (or a combination of them) underlie different complex diseases and traits.[45;46] The difficulty lies in determining which architecture best explains the genetic variance of a particular trait of interest, and therefore selection of the investigational approach(es) that might give the best results.

Genome-Wide Association (GWA): GWA studies report the statistical association between a trait of interest and the genotype of an individual at known polymorphic sites throughout the genome. This approach may be equally applied to continuous and dichotomous or "threshold" traits and has been widely applied to complex diseases and traits in both humans and animals.[48;51-53] The statistical power of a GWA relies heavily on marker coverage, the number of samples, appropriate correction for population structure, and the quality of phenotyping.[54;55]

The density of markers needed for adequate coverage of the genome depends on linkage disequilibrium (LD) within a species and is based on the concept of "tagging" markers.[54;56] LD is the nonrandom association of alleles at different genetic loci.[56-58] Alleles that are tightly linked tend to be inherited together, while alleles that are not linked will segregate independently from one another. An unknown disease-causing allele with a low frequency in the population can be identified by its linkage to a more common allele included in a genotyping array. This genotyped allele is known as a

"tagging" marker.[54;56;58] As recombination occurs over generations, the length of LD blocks tends to decrease, and markers must be in closer physical proximity to each other to be reliably inherited together; thus the density of markers in a genotyping array must be increased.[47;54] When compared to humans, LD in domestic animal species is quite long as a result of the relatively recent foundation of separate breeds, genetic isolation and inbreeding, and selection for specific traits.[48;59-61] Thus, successful association mapping can be carried out with tens of thousands of markers, rather than the hundreds of thousands or millions of markers used in human studies. In the horse, average LD, as measured by the correlation coefficient between markers (r^2), remains high ($r^2 > 0.2$) for 100-150kb within most breeds. In the Thoroughbred, r^2 is greater than 0.2 for a distance of 400kb. Long-range LD (greater than 1,200kb) is greatest in Standardbreds and French Trotters, and a conserved haplotype block of up to 4.2Mb within individual breeds is reported at the *MC1R* (chestnut coat color) locus on ECA3.[59] By comparison, most reported haplotype blocks in LD in humans range in size from 5 to 20kb, and the longest reported block is 804kb.[58] Two commercially available genotyping platforms have been developed that capitalize on the extensive LD in the horse. The first generation "SNP chip" (Illumina Equine SNP50) had ~54,000 informative markers (single nucleotide polymorphisms, or SNPs) with average spacing of 43kb between markers.[59] A slightly denser chip (Illumina Equine SNP70) was recently developed to fill in some of the larger gaps in the first chip and contains ~65,000 markers. Although these genotyping platforms have been used successfully to identify loci of interest in a variety of association mapping studies in multiple breeds[59;62-66], they have limitations related to sparse marker density and gaps in coverage for certain regions of the genome, as well as potential ascertainment

bias because SNP discovery was carried out in only a small number of individuals (n = 8).[59] These limitations are of special concern in breeds with greater admixture (e.g. Quarter Horses) or lower LD (e.g. Mongolian Horse).[59]

When it comes to number of samples in a GWA study, the convention is that bigger is better.[55] Indeed, for common human diseases, thousands of unrelated individuals are generally enrolled in a GWA study.[67] However, this is neither a practical nor an economically feasible approach when studying traits in horses. Instead, a more family-based approach can be leveraged in this species, taking advantage of the fact that the creation of breeds from a limited number of founder individuals tends to enhance the number of rare alleles within a population.[56;68;69] This can improve the power of a GWA study to detect association of such rare alleles with disease[68;69], although accounting for population structure, for example, through incorporation of a kinship matrix based on known pedigree or marker identity by state (IBS) or through the use of one or more principal components, is essential to avoid inflation of type I error.[48;55;70;71]

Determining phenotype for certain quantitative production traits in domestic animals (i.e. milk yield in cattle, litter size in pigs) is generally straightforward, but the same cannot be said for many complex traits and diseases in the horse. Individuals may be classified as affected with a disease ("cases") based on a specific constellation of clinical signs, but in fact have conditions with diverse etiologies. Alternatively, diseases with a latent phase, or those that are difficult to diagnose using routine methods, can result in individuals being incorrectly classified as unaffected ("controls"). Accurate assignment of individuals as cases or controls is essential to avoid introducing misclassification bias into a study.[55] Although this bias can be offset by an increase in

study population size[55], this is not always possible. Utilizing a more stringent case definition, such as inclusion of only extreme examples of a phenotype or cases from a single family, may help to improve the power of a case-control GWA (assuming population structure is taken into account).[55] Similarly, use of so-called "hypernormal" controls has been suggested as a way to avoid misclassification bias, although one must be careful not to introduce selection bias from having controls from a different population than cases.[55] This is especially crucial for diseases in which environment is known to play a role. For example, a group of Thoroughbred racehorses from Kentucky that have undergone standing magnetic resonance imaging (MRI) of the lower limb with no significant findings may be convincingly classified as unaffected with fetlock OC, but they would be completely inappropriate as a control group for a cohort of OC-affected Standardbred horses from Michigan. Environmental effects can be difficult to quantify and account for in statistical models, and this is frequently cited as a major reason for failure of GWAS findings to be replicated in independent populations.[47;55] Laboratory conditions are of course ideal for environmental control, but this approach is impractical in most equine research. Selection of individuals from a single farm and/or collection of detailed information regarding environment (including diet and exercise regimens) are potential approaches to overcome this challenge.

Next-Generation Sequencing (NGS): The availability of a high-quality draft reference genome for the horse[72], combined with the ever-decreasing cost of next-generation sequencing technologies has made whole-genome and whole-transcriptome sequencing approaches feasible for the study of complex equine traits. Financially, there is a tradeoff between number of samples and depth of coverage, and the general

consensus seems to be that while low-coverage sequencing of a large number of individuals is adequate for investigation of population-level parameters, deeper sequencing of individuals is preferable for variant discovery.[73;74] A combination of approaches – i.e. sequencing a few individuals with higher coverage and a larger number of individuals with lower coverage – may be ideal for equine studies because it allows a larger population to be sequenced without sacrificing the ability to detect variants of interest. Variants discovered in a small number of individuals can subsequently be genotyped in a larger population to confirm association with the trait/disease of interest.[73]

Although the publically available equine genome, EquCab2, is an invaluable resource, challenges remain related to incomplete annotation of genes and differing gene models. Due to the limited amount of expression data available in the horse, the majority of protein-coding genes are annotated based on *in silico* prediction extrapolated from the gene structure of other species.[75] Different predictor algorithms have resulted in a difference of more than 5,000 called genes between the Ensembl (20,449; http://useast.ensembl.org/Equus_caballus/Info/Index) and NCBI (25,565; http://www.ncbi.nlm.nih.gov/genome/145) databases. Furthermore, evidence from transcriptome data suggests that there are genes that are completely missing from current annotation models, and that many predicted gene models are missing one or more exons (especially exon 1), as well as 5' and 3' untranslated regions.[75] It is not unlikely that variants related to traits and diseases of interest lie within these unannotated regions.

Hypotheses and Specific Aims

Osteochondrosis (OC) in the Standardbred Horse

Objective 1: Identify specific genes and alleles underlying OC susceptibility in the horse. OC is most simply defined as a failure of endochondral ossification, the process by which a cartilage template becomes bone in the limbs of a growing animal. It is characterized by the presence of abnormal cartilage within a joint that may be thickened, soft or collapsed, or separated entirely from the underlying bone.[76] OC is widely recognized in young horses across breeds and is of particular interest because of its potential to cause joint effusion and/or lameness in yearling horses preparing for sales and entering training. Young horses affected with OC may improve with conservative therapy alone, but in many cases surgical intervention is required. Further, severe manifestations of this disease, or inadequate treatment of mild to moderate forms, can lead to long-term debilitative consequences. In these cases, OC can be career- or even life-threatening. Reduction in incidence and, ultimately, prevention of OC is an as-yet unattained goal of the equine industry.

The presence of OC across domestic horse populations, including a feral horse population[77], as well as shared major predilection sites and lesion morphology suggest a unified underlying pathophysiology and shared genetic risk across breeds. The central hypothesis of the first part of this thesis is that one or more genes of major to moderate effect underlie OC susceptibility in horses, and further, that these risk loci are shared across breeds. Standardbred horses were selected as a model population to capitalize on the high prevalence and heritability of hock OC in this breed.

In this objective, chromosomal regions associated with hock OC will be identified using a GWAS in a group of Standardbred yearlings with a shared early environment. The single nucleotide polymorphisms utilized for the GWAS are not expected to directly underlie risk of disease. Thus, whole-genome sequencing in a subset of the population will be performed for the purpose of variant discovery within the regions of interest. Variants will be prioritized by segregation with disease status and predicted functional effect. Putative functional variants will then be genotyped in a larger population to confirm association with disease. This objective represents a crucial step in development of a genetic risk model for OC susceptibility, allowing for genetic testing and quantification of risk in individual horses. Improved risk assessment will facilitate management changes and early intervention in high-risk horses and allow for informed breeding decisions in high-risk breeds/pedigrees.

Gait and Performance in the Standardbred Horse

Objective 2: Identify genetic determinants of pacing in the Standardbred horse. Gaits are specific coordinated patterns of locomotion that can be classified in a variety of ways, including cadence, sequence of foot-fall, and symmetry. Natural gaits in equids include the walk (4-beat, symmetrical), trot (2-beat, diagonal, symmetrical), canter (3-beat, asymmetrical), and gallop (4-beat, asymmetrical). However, among equids, the domestic horse is somewhat unique in that certain breeds have the natural ability to perform additional gaits and have been selected for this ability. These additional gaits have been well-characterized phenotypically, but the developmental physiology and underlying genetic determinants responsible for their expression in specific breeds are

largely unknown. “Gaited” breeds present a diagnostic challenge to practitioners evaluating lameness and performance issues, so an improved understanding of the genetic and physiologic factors playing a role in unique gaits is of significant clinical importance.

The recently described *DMRT3* mutation that appears to be permissive for “gaitedness” across breeds[11] supports the central hypothesis that functional mutations leading to altered neural connections in the spinal cord are responsible for alternative locomotion patterns in the domestic horse, including the pace. However, as this mutation is nearly fixed in Standardbred horses[11], we hypothesize that modifying genetic loci exist that interact with and furthermore that these loci are shared across breeds that have been selected for similar gaits.

In this objective, chromosomal regions associated with pacing will be identified by performing a GWAS in a large population of Standardbred pacers and trotters. Subsequently, variant discovery will be performed by whole-genome sequencing a subset of horses from the GWAS population. Variants will be prioritized by segregation with gait and by predicted functional effect. Putatively functional variants will then be genotyped in the larger population to confirm association with gait.

Objective 3: Identify genetic determinants of performance in the Standardbred horse. Although “performance” can be defined in a variety of ways, including number of starts in a season or over a career, amount of earnings, or some combination of factors[9;78], in this instance, it is defined as the fastest recorded speed over a mile. This measure is considered particularly appropriate in the Standardbred because speed was the primary selection criterion used during creation of the breed.[14] A similar methodological approach to the one described above will be used to address this objective. A GWAS in the same

population of Standardbred pacers and trotters used for objective 2 will be performed to identify chromosomal regions associated with fastest speed over a mile as the outcome of interest. Analysis of whole-genome sequencing data and/or selected sequencing of candidate genes within regions of interest will be performed in the future to identify putative functional variants underlying speed in this breed. Similar work carried out in the Thoroughbred leading to the identification of the so-called "speed gene"[79] has garnered significant attention in the racing community, and it is likely that the results of objective 3 would be of similar interest among Standardbred trainers and breeders. Beyond this, objectives 2 and 3 collectively may provide valuable insight into developmental biology and exercise physiology in the horse.

Table 1: Weighted average prevalence of OC at known predilection sites as reported by breed.

Breed	Lesion Prevalence (%)♦ [range (%)]			Reference
	Fetlock* (MCP/MTP)	**Hock§**	**Stifle‡**	
Quarter Horse	15	5.7	2.7	[80]
Standardbred†	3.3 [2.5-4.8]	14.7 [10.1-26.2]	6.3 [6.2-6.3]	[32-38;42]
Thoroughbred×	12.9 [6.0-20.0]	5.3 [3.0-8.6]	4.7 [2.7-12.0]	[19;35;39-41;81]
Warmblood°	22.3 [18.3-23.5]	11.5 [9.2-22.5]	7.0 [4.4-16.6]	[38;82-87]
South German Coldblood	53.9	40.7	NR	[88]
Draft Breeds∞	0.3	3.0	1.7	[89]
Maremmano	2.8	9.2	5.1	[90]
Feral	3.75	2.5	NR	[77]

MCP = **metacarpophalangeal joint**; **MTP** = **metatarsophalangeal joint**; **NR** = **not reported**
♦ Prevalence presented as a weighted average of the reported prevalence in the associated reference(s)
† Includes populations from the United States, Canada, Denmark, France, and Sweden
× Includes populations from the United States, France, and New Zealand
° Includes Dutch Warmblood, French Warmblood, Swedish Warmblood, Hanoverian, and Holsteiner breeds. These represent populations from the Netherlands, France, Sweden, and Germany.
∞ Includes Clydesdale, Percheron, Belgian, and Shire. These represent populations from the United States and Canada. The horses in this study (n=51) were all affected; prevalence was determined by dividing by the reported total draft horse population (n=1135) seen in the reporting hospitals during the study period.
* This excludes **palmar/plantar osteochondral fragments** (**POF**), which are generally considered to be of traumatic, rather than developmental origin.[30]
§ Includes all lesions within the tarsocrural joint: **distal intermediate ridge of the tibia** (**DIRT**), **lateral and medial trochlear ridges of the talus** (**LTR** and **MTR**, respectively), **medial malleolus** (**MM**).
‡ Includes lesions of the femoropatellar joint, but excludes subchondral bone cysts.

PART I:

Genetic Risk Factors Underlying Osteochondrosis in the Standardbred

Chapter 2

Articular Osteochondrosis: A Comparison of Naturally-Occurring Human and Animal Disease

Articular Osteochondrosis: A Comparison of Naturally-Occurring Human and Animal Disease

Annette M McCoy[1], Ferenc Toth[1], Nils I Dolvik[2], Stina Ekman[3], Jutta Ellermann[4], Kristin Olstad[2], Bjornar Ytrehus[5], and Cathy S Carlson[1]

From the [1]Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA; [2]Department of Companion Animal Clinical Sciences, Equine Section, Norwegian School of Veterinary Science, Oslo, Norway; [3]Department of Biomedicine and Veterinary Public Health, Division of Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden; [4]Department of Radiology, The Center for Magnetic Resonance Imaging Research, University of Minnesota, Minneapolis, MN, USA; [5]Agder District Office, Norwegian Food and Animal Health Authority, Arendal/Kristiansand, Norway

Acknowledgements: This research received no specific funding for scientific objectives. McCoy and Toth were supported by a National Institutes of Health institutional training grant (T32OD10993); Carlson was supported by NIH K18OD010468.

Published as:

McCoy AM, Toth F, Dolvik NI, Ekman S, Ellermann J, Olstad K, Ytrehus B, Carlson CS. Articular osteochondrosis: a comparison of naturally-occurring human and animal disease. (2013) *Osteoarthritis & Cartilage* 21:1638-1647.

Abstract

Background: Osteochondrosis (OC) is a common developmental orthopedic disease affecting both humans and animals. Despite increasing recognition of this disease among children and adolescents, its pathogenesis is incompletely understood because clinical signs are often not apparent until lesions have progressed to end-stage, and examination of cadaveric early lesions is not feasible. In contrast, both naturally-occurring and surgically-induced animal models of disease have been extensively studied, most notably in horses and swine, species in which OC is recognized to have profound health and economic implications. The potential for a translational model of human OC has not been recognized in the existing human literature.

Objective: The purpose of this review is to highlight the similarities in signalment, predilection sites and clinical presentation of naturally-occurring OC in humans and animals and to propose a common pathogenesis for this condition across species.

Study Design: Review

Methods: The published human and veterinary literature for the various manifestations of OC was reviewed. Peer-reviewed original scientific articles and species-specific review articles accessible in PubMed (US National Library of Medicine) were eligible for inclusion.

Results: A broad range of similarities exists between OC affecting humans and animals, including predilection sites, clinical presentation, radiographic/MRI changes, and histological appearance of the end stage lesion, suggesting a shared pathogenesis across species.

Conclusion: This proposed shared pathogenesis for OC between species implies that naturally-occurring and surgically-induced models of OC in animals may be useful in determining risk factors and for testing new diagnostic and therapeutic interventions that can be used in humans.

Introduction

Osteochondrosis (OC) is a developmental orthopedic disease characterized by clinical signs of joint pain, effusion, and dysfunction caused by the formation of clefts extending through the articular cartilage into the subchondral bone. Extensive studies evaluating the clinical aspects of this condition are available in both human and veterinary medicine; however, there is limited information available regarding the similarities and differences between OC in humans and animals.

The majority of studies aimed at describing the etiologic factors and pathogenesis of OC in humans focus on osteochondral fragments removed surgically from adolescents or adults presenting with clinical symptoms of OC.[91] By this time, the fragments have been present for months to years. Understandably, it is nearly impossible to determine the pathogenesis of the disease from examination of these end-stage tissues. Obtaining osteochondral samples from juvenile human cadavers is difficult, and currently there is no established method for screening asymptomatic children or adolescents for OC. Both of these factors have hampered the understanding of the pathogenesis of naturally-occurring human disease. In contrast, in the veterinary literature, OC is defined as a focal disturbance of endochondral ossification[92], the process by which a cartilage template ossifies in the appendicular skeleton of a growing individual. Extensive studies performed in young growing animals of several species have demonstrated early, developing lesions at predilection sites well before the age at which clinical disease manifests.[93-95] We believe that naturally-occurring and surgically-induced OC in animals may provide valuable translational models to help understand the etiology and pathogenesis of human disease. Our review, therefore, aims to highlight the similarities in

signalment, predilection sites and clinical presentation of naturally-occurring OC in humans and animals, and by doing so, propose a common pathogenesis for this condition across species.

Disease Terminology

Evaluation of the literature pertaining to OC is complicated by the variety of terminologies used. In 1887, König proposed the term "osteochondritis dissecans" for an underlying lesion in the joint cartilage facilitating formation of loose bodies in the absence of significant trauma.[96] Subsequent histological studies have not supported a primary inflammatory etiology for the condition, making "osteochondrosis" the more accurate term, as suggested by Howa1d in 1942.[97;98] However, the original phrase has persisted, and in fact, "osteochondrosis" and "osteochondritis" are often used interchangeably. In the clinical literature, when a fissure or fracture in the overlying articular cartilage is present, the condition is nearly universally referred to as osteochondritis dissecans (OCD), although osteochondr*osis* dissecans would be more appropriate. In the veterinary medical field, focal abnormalities of endochondral ossification involving the articular-epiphyseal cartilage complex (AECC) are referred to as osteochondrosis (or osteochondrosis dissecans, as appropriate) regardless of anatomical location. Conversely, in the human literature, manifestations of OC at various anatomical sites are given different names (**Table 1**). Additionally, the phrase "the osteochondroses" includes conditions affecting the AECC, the physis, and various apophyseal locations. This general phrase has also been used to describe diseases of

primary osteonecrotic etiology, such as Legg-Calvé-Perthes disease.[99] The present article will specifically focus on articular manifestations of OC.

Clinical Aspects of Osteochondrosis in Humans and Animals

Human OC is typically not recognized in children or adolescents until the onset of clinical symptoms, at which point the disease is advanced.[100] In many cases, a lag time of months to years may exist between the onset of symptoms and diagnosis of the disease.[101] OC diagnosed prior to the age at which physeal closure occurs is known as juvenile OC; however, lesions diagnosed in adulthood also most likely developed prior to physeal closure.[102] Common presenting clinical complaints include joint pain, especially with extreme flexion or extension, swelling, and catching or locking of the joint. These symptoms may be intermittent, especially early in the course of disease, and may be associated with athletic activity. Continuous or more severe symptoms may be indicative of a loose osteochondral fragment within the joint.[102;103] Bilateral disease is not uncommon, although clinical symptoms are typically worse in one joint than the other.[104] Diagnosis is typically made by radiologic and/or magnetic resonance imaging (MRI) examination of the affected joint. MRI more closely aligns with arthroscopic findings[105] and is also more sensitive for identification of subtle cartilage abnormalities (i.e. prior to formation of overt osteochondral fragments), suggesting that this may be the better imaging modality for OC, especially for early lesions.[102] The preferred initial treatment for OC when the articular surface is intact is non-surgical management, including a combination of non-steroidal anti-inflammatory drugs, physical therapy, and modification of activity, typically with some form of joint immobilization. If conservative therapy

fails, or if partially or completely detached osteochondral fragments are present at the time of diagnosis, then surgical intervention via arthroscopy is pursued.[102;104;106] Although removal of the fragment/flap followed by debridement is most common, reattachment of large osteochondral flaps using internal fixation has also been described.[107] Lesions that are not treated adequately may lead to development of degenerative joint disease with long-term debilitative consequences for the individual; thus, early intervention is recommended.[100;107]

In horses, asymptomatic OC is usually identified at an early age due to extensive radiographic screening aimed at facilitating sale of racehorses as yearlings (before two years of age). In more slowly-maturing breeds that usually do not undergo early radiographic screening, OC is most often identified after 3 years of age as clinical signs, including subtle lameness and joint effusion, develop after the commencement of regular training. This latter presentation is strikingly similar to that noted in cases of juvenile OC in humans, which most frequently affects young athletes and usually presents with poorly localized pain that is exacerbated with exercise.[99;104;108] In horses, OC lesions are most often treated with arthroscopic removal of loose fragments followed by debridement of the fragment bed with or without microfracture.[23] Although many horses go on to perform in their intended capacity after treatment, the prognosis for future athletic career following surgical debridement of OC lesions diminishes as the size of the lesion increases.[109] Novel treatment modalities attempting to salvage and reattach large osteochondral flaps have recently been introduced to address this concern.[23]

In commercially bred pigs, OC is considered to be an important cause of lameness with profound economic implications.[95;110] Clinical signs consistent with OC have been

associated with decreased longevity of young female swine intended for breeding.[111;112] Although histological changes characteristic of early articular OC have been described in the femoral condyles of pigs as young as 6-8 weeks of age[113], clinical signs usually do not become apparent before adolescence.[114] Treatment of OC in pigs is usually not economically feasible and severely affected animals are generally sent to slaughter. As a result, greater emphasis is placed on prevention rather than treatment of disease, although arthroscopic removal of an OC lesion affecting the talus has been reported in this species.[115]

It is worth noting that skeletal maturity is reached much more rapidly in the animal species discussed above than in humans. Ossification ("closure") of the physeal (metaphyseal growth plate) and epiphyseal (AECC) cartilage is the hallmark of skeletal maturity in all species. In young humans, this process occurs between 14 and 25 years of age, depending on anatomical location.[116] In contrast, growth cartilage closure in horses begins around three months of age and is considered complete before 3 years of age.[117] Thus, a yearling horse would be at the approximate maturity of an adolescent human, with the onset of clinically-apparent OC in both species occurring around the time of cessation of growth. The association of athletic activity and onset of clinical signs is also reflected in both humans and horses, although asymptomatic lesions are undoubtedly present earlier.

Prevalence of Osteochondrosis

Global estimates of the prevalence of articular OC are not reported in the human literature, likely due to the tendency to regard manifestations of OC at different

anatomical locations as separate diseases. Prevalence of elbow OC was reported as 4.1% in one radiographic survey of 1,000 Danish men over the age of 15[118], while incidence of knee OC was calculated to be between 15 and 30 per 100,000 in women and men (respectively) between the ages of 10 and 20 in a single Swedish city.[119] In general, OC of the knee is considered to be most common, representing approximately 75% of all cases[102], with manifestations in the elbow (second most commonly affected location), ankle, and hip occurring relatively uncommonly.[101] However, it is likely that any estimate of OC prevalence in humans is an underestimate, given that diagnosis is generally delayed until the onset of clinical signs[120]; many individuals may be asymptomatically affected and never diagnosed. In contrast, in horses, where survey radiographs are routinely taken in many breeds before the onset of clinical signs, global prevalence estimates range from 20% to 80%, although prevalence varies by joint and breed.[39;121] Similarly, in pigs, where prevalence is determined based on post-mortem surveys, up to 70% of animals are reportedly affected.[95;122] Most of these lesions in horses and pigs are subclinical/asymptomatic at the time of diagnosis.

Proposed Pathogenesis and Risk Factors

The underlying etiology and pathogenesis of OC have long been the subject of controversy and speculation, and a variety of environmental and genetic risk factors have been proposed. Historically, the major schools of thought have been divided into those who propose trauma as the primary cause for OC and those who suggest alternative underlying processes, including inflammation, osteonecrosis, vascular abnormalities, and cartilage extracellular matrix abnormalities.[96;97] König himself fell into the latter

category, describing osteochondritis dissecans as occurring in the absence of any significant trauma.[123] However, the inflammatory etiology suggested by the term "osteochondritis" has not been corroborated by subsequent histological studies.[91;124] Histological results from osteochondral fragments removed during surgery also fail to support osteonecrosis of subchondral bone as the primary lesion of OC.[91;124] Instead, necrosis of the subchondral bone is thought to most commonly occur secondary to detachment of the osteochondral fragment, rather than being an inciting cause.[125] The production and accumulation of abnormal extracellular matrix molecules in the endoplasmic reticulum (ER) has been suggested as the underlying cause of abnormal mineralization (failed endochondral ossification) and subsequent OC lesions, based on electron microscopy of surgically removed OC lesions and adjacent "normal" cartilage biopsies from four human patients.[126] The authors hypothesized that an underlying inherited ER storage disorder was responsible for abnormal protein production and accumulation, although no candidate mutation was identified.[126] However, other conditions, including local ischemia, can lead to accumulation of unfolded, but otherwise normal, proteins in the ER.[126] Additionally, the tissues examined were end-stage, making it difficult to determine if the ultrastructural changes were a cause or a consequence of disease. Abnormalities in cartilage extracellular matrix maturation have also been proposed to play a central role in the development of OC. Decreased collagen content and alterations in collagen cross-linking were reported in cartilage samples obtained from foals with OC compared to healthy foals, and this "immature" cartilage was considered potentially more susceptible to external trauma.[127] Similarly, it has been suggested that the marked changes in collagen fibril orientation and density across the epiphyseal

cartilage, especially near nutritive cartilage canals in the ossification front may create focal areas of biomechanical weakness.[128] However, there is little evidence that these matrix changes are primary. Indeed, many of the matrix alterations reported by Lecocq et al (2008) were located in or near focal regions of chondronecrosis.[128] Studies performed in animals demonstrate that these focal regions of chondronecrosis form due to interruption of the blood supply to the nutritive cartilage canals within the epiphyseal cartilage of the AECC during endochondral ossification[129], and this pathogenesis can be reproduced experimentally.[113;130;131] Areas of chondronecrosis resist normal ossification and degenerate, resulting in tissue that is prone to clefting or collapse under the influence of external forces.

Proponents of an etiology for OC involving major trauma have suggested that the preponderance of disease in young boys compared to girls is related to greater athletic activity and tendency towards overuse injuries and trauma in males.[99;120] In most cases of human OC there is no history of a single traumatic event; however, repetitive stress could be important in the development of lesions.[102] This latter hypothesis is supported circumstantially by the fact that most patients affected with juvenile OC have a lengthy history of participation in specific exercise regimens or sports.[125] It is also possible, however, that more active individuals are more likely to become symptomatic and are therefore more likely to have OC diagnosed. In naturally-occurring disease in animals, the role of athletic activity is not clear-cut and is thought to be a secondary factor. For example, in one large study in horses, controlled exercise affected the distribution of OC lesions within joints, but not the total number of lesions.[132] Another study found that regular, limited exercise appeared to reduce the risk of OC development in young

foals.[133] Osteochondral fragments can be elicited in experimental animal models using either repetitive impacts or acute compression and rotation[134], but these models cannot replicate the more commonly recognized early OC lesions with intact overlying articular cartilage.[135] The idea of major trauma as the sole etiologic agent is also brought into question by the occurrence of OC at anatomical locations not exposed to increased stress during physical activity, and because it cannot explain early histologic changes seen at OC predilection sites in young animals.[93;113;136] Thus, while trauma undoubtedly is a key precipitating factor in the onset of clinical signs of OC (i.e. by resulting in disruption of the articular surface and separation of an osteochondral fragment), it is less likely to be the initiating factor in disease development.

A variety of additional environmental factors have been proposed to play a role in the risk for development of OC in veterinary species, including nutrition, exercise, conformation and other biomechanical factors, stress response, *in utero* environment, and hormonal interactions.[137;138] Of these, nutrition has been the subject of the most study. In animal models of disease, dietary factors that have been implicated in OC risk include copper deficiency[139;140], excess phosphorus[141], and excess dietary energy.[142] However, the relationship between disease and nutrition is far from one of straightforward cause-and-effect. For example, in pigs, dietary supplementation of specific amino acids and microminerals reduced the severity, but not the incidence, of OC lesions when compared with a control diet.[143] Similarly, reducing digestible energy and increasing micromineral concentrations reduced the incidence of OC in foals in a prospective study of 17 breeding farms, but did not eliminate the condition.[144] It is likely that "windows of susceptibility" exist during which dietary factors may play key roles in OC manifestation[145], but these

may vary between species and anatomic location, and have yet to be clearly defined. To our knowledge, there are no reports examining a potential relationship between OC development and nutrition in humans.

Genetic risk factors are also thought to play an important role in the development of OC in both humans and animals. Since the initial description of human OC, there have been many reports of families with apparently increased incidence of disease. Many of these initial reports were small, involving a few siblings or a parent and his/her children.[146;147] However, extended families with high incidence of OC over multiple generations have also been reported in the literature[148;149], and these suggested an autosomal dominant pattern of inheritance with varied penetrance. Patients in these families were often affected in multiple joints, and an association with short stature was noted.[148;149] This condition was named Dominant Familial Osteochondritis Dissecans (OMIM 165800), and is caused by a missense mutation in the aggrecan (*ACAN*) gene that results in an aggrecan protein with a reduced ability to interact with other proteins found in the cartilage extracellular matrix.[150] However, this familial form of disease is rare, and the *ACAN* mutation is unlikely to underlie other manifestations of OC. The more common, sporadic, form of OC is likely polygenic in nature; that is, the combined effects of multiple gene variants determine the underlying genetic risk of disease. To our knowledge, there are no reports in the human literature attempting to identify genes contributing to sporadic OC. However, several case reports describing identical lesions, including nearly simultaneous timing of clinical presentation, in monozygotic twins have been reported in the literature.[151-153]

Heritability estimates for OC in horses and pigs range from 0.14 to 0.52, depending on location and disease definition.[32;84;154] Thus, between 14% and 52% of disease risk may be attributed to genetic factors in these species. Up to 70% of foals from a single sire have been reportedly affected with OC[32], and offspring of affected sires were more than twice as likely to develop OC than offspring of non-affected sires.[37] Similarly, boars with affected half-siblings were highly likely to have OC-affected offspring.[95] Two approaches have been taken to try to identify genetic risk factors in these species. The first is a candidate gene approach, where genes known to play a role in skeletogenesis and related processes are identified and subsequently sequenced to try to identify putative risk alleles/causative mutations.[155;156] The second approach is genome-wide association (GWA) analysis, which evaluates statistical association between allele frequency at tens of thousands of known sites of variation throughout the genome and disease status.[53;65;157] GWA studies have been performed examining both the overall occurrence of OC in commercial pigs, and the manifestation of OC in specific locations, including the metacarpophalangeal joint and tibiotarsal joint, in several breeds of horses including Standardbreds, Warmbloods, Thoroughbreds, and French Trotters.[53;63;65;66] Both candidate gene and GWA approaches have limitations and, to date, although several promising candidate chromosomal regions have been identified, specific genes and alleles underlying risk have not been definitively identified. Work in this field is ongoing, however, and improvements in next-generation sequencing technology as well as the formation of cross-institutional consortia will aid efforts. Genes and pathways identified in veterinary species will not only provide insight into OC pathophysiology, but will also become compelling biological candidates for further study in humans.

Evidence for Disturbances of Endochondral Ossification Leading to Osteochondrosis

Evidence from animal models most strongly supports the theory that the primary pathophysiological process underlying the development of osteochondrosis is a disturbance in endochondral ossification, the process by which the epiphyseal growth cartilage of the AECC is gradually replaced by bone. It is likely that several of the etiologic factors described above contribute to this disturbance.

In both humans and animals, nutrients are normally supplied to the epiphyseal cartilage of the AECC and the physis via vessels running in channels called cartilage canals (**Figure 1**).[158-161] The majority of these vessels arise from the perichondrium and run parallel to the articular surface. Physiologically, as endochondral ossification progresses and the ossification front advances in the direction of the articular surface, the blood supply to most vessels in the cartilage canals must change from vessels originating from the perichondrium/periosteum to vessels originating from the medullary cavity of the secondary center of ossification. This transfer enables a consistent supply of nutrients to the ever-thinning AECC and involves the formation of anastomoses between vessels in the epiphyseal cartilage and vessels in the advancing ossification front. Studies performed in animals, however, have demonstrated that this blood supply is prone to failure.[129] This failure may be due to physical instability of the newly formed anastomoses, local effects on the vasculature by growth factors at the ossification front, or inadequate mechanical support for the developing vasculature from the surrounding tissue.[129] The impact of mechanical forces may also be especially high at the transition between two tissue types with very different mechanical characteristics.[129] The lack of anastomoses among vessels

contained within the cartilage canals[162] means that failure of this transition in blood supply often results in avascular necrosis of a well-defined area of epiphyseal growth cartilage. Failure of vascularization and mineralization of the necrotic cartilage causes focal arrest of endochondral ossification, the hallmark of osteochondrosis.[136;163] Indeed, lesions resembling OC have been successfully reproduced in pigs and horses by surgical transection of vessels contained within cartilage canals.[113;130;131] The retained area of necrotic epiphyseal cartilage is inferior to viable epiphyseal cartilage or subchondral bone in providing support to the overlying articular cartilage, predisposing the site to collapse and/or cleft formation that often results in clinical disease.

The focal area of cartilage necrosis in the epiphyseal cartilage is the first histologically apparent lesion during the pathogenesis of OC and is termed *osteochondrosis latens* (**Figure 2**) in the veterinary literature.[159] This lesion has not been described in humans, most likely due to limited access to appropriate tissues for evaluation and because it is not radiographically evident. As the ossification front reaches the necrotic epiphyseal cartilage, the necrotic cartilage resists ossification and becomes radiographically apparent as a radiolucent defect in the subchondral bone, at which point it is designated as *osteochondrosis manifesta* (**Figure 3**).[95] This lesion is observed in human medicine but is not currently recognized as a preclinical form of OC.[164;165] There are two potential fates for this area of necrotic cartilage. In some cases (likely depending, in part, on the size and location of the lesion), it will eventually undergo ossification and resolve with minimal to no visible remnants.[95] Alternatively, if the area of necrotic epiphyseal cartilage is very large or if the overlying articular cartilage sustains excessive trauma, as may occur during athletic activities, a fissure extending from articular cartilage

through the underlying necrotic epiphyseal cartilage may develop, at which point the lesion is described as *osteochondrosis dissecans* (**Figures 4 and 5**).[95] This lesion is most frequently termed *osteochondritis dissecans* in the human literature. In both humans and animals, this stage of the disease results in clinical signs of joint pain/dysfunction and lameness.

As noted earlier, the majority of histological studies evaluating OC/OCD in humans have focused on osteochondral fragments removed during surgery, which are easily accessible but represent the end stage of the disease and rarely include evidence of the early changes affecting the endochondral ossification process. Indeed, human studies regard fibrous/fibro-cartilaginous tissue present at areas of separation of osteochondral fragments from the parent bone as areas of delayed or nonunion.[91] However, the complete absence of a calcified cartilage layer and subchondral bone plate in osteochondral fragments removed from adolescents affected by juvenile OCD[91] indicates that juvenile OCD develops while the endochondral ossification is still ongoing, thus it is unlikely to be a primary disease of the subchondral bone.[166;167] Instead, we believe that this fibrous/fibro-cartilaginous tissue is the remnant of necrotic epiphyseal cartilage and accompanying reactive fibrous tissue which has been present as *osteochondrosis manifesta* well before the development of clinical signs (**Figure 4**). This theory is supported by histological studies in horses and pigs using specimens from young animals undergoing active endochondral ossification and in osteochondral samples obtained from adult animals affected by clinically apparent OCD. Histological studies in foals with ongoing endochondral ossification demonstrated lesions consistent with subclinical OC (*osteochondrosis latens*) at predilection sites of clinically relevant OC.[93;163] Conversely,

studies examining the osteochondral fragments removed from adult horses affected with OCD revealed histological changes consistent with those noted in human studies, including fibrous tissue at the separation border.[168] The similar appearance of end-stage lesions across species suggests that the continuum of disease demonstrated in animals is also likely present in humans.

Shared Aspects of Human and Animal Osteochondrosis

Several factors are suggestive of a shared pathogenesis of OC between humans and veterinary species. In addition to histologically identical end-stage disease as described previously, humans and animals share common predilection sites for development of disease (**Table 1**). In humans, OC is diagnosed in the knee, elbow, and ankle joints with decreasing frequency.[169] Within the knee, the medial femoral condyle is the most commonly affected area, whereas in the ankle, OC affects the talus[170] more commonly than the tibial plafond.[171] OC of the elbow joint usually involves the humeral capitellum.[166] Similarly, in swine, the disease is most commonly seen in the medial condyle of the femur and the medial aspect of the sagittal ridge of the humeral condyle.[92] Lesions are also found, at a lower frequency, in the shoulder, hip, and tibiotarsal (ankle) joints of swine. The most commonly affected sites in horses are the tibiotarsal, stifle (knee), and metacarpo/metatarsophalangeal joints.[39;172] Similar to humans, the most commonly affected sites in the tibiotarsal joint in the horse are the distal intermediate ridge of the tibia and the lateral trochlear ridge of the talus (**Figure 6**).[173] In the stifle, the lateral trochlear ridge is the most commonly affected structure in the horse[39;172], but involvement of the medial femoral condyle has been described as well (**Figure 7**).[174] OC

also infrequently affects the shoulder, carpal, and hip joints of both humans and horses.[172;175]

Clinically apparent bilateral involvement in both humans and animals is widely reported, although the frequency of occurrence varies between joints. Bilateral juvenile OC of the knee is reported in 13 – 30% of human patients[108], and 10% of human subjects are affected bilaterally with osteochondral lesions of the talus.[176] Similarly, in Thoroughbred racehorses, bilateral involvement of the stifle (knee) occurs in 17.5 – 20.5% of affected animals, while incidence of bilateral disease in the tibiotarsal joint is 6 – 10%.[39;40] Many more human patients may have clinically silent lesions in the contralateral joint visible on MRI, although the importance of these lesions has been recently been called into question.[164] However, based on studies in animals, it is highly likely that the majority of the "ossification variants" identified in the MRI studies in patients under the age of eight[165] were, in fact, actually *osteochondrosis manifesta* lesions. The lack of progression of these "ossification variants" into clinically apparent OC, and their occurrence in young patients, corresponds with observations made in animals, where the high ratio of subclinical (*osteochondrosis manifesta*) to clinical (OCD) lesions suggest that the majority of lesions undergo complete healing (**Figure 5**).[95;122] Indeed, healing of radiographically apparent juvenile OCD is reported to occur in approximately 50% of human patients.[125]

The apparent potential for healing of subclinical OC lesions has been demonstrated in swine and horses as well. In young swine, subclinical *osteochondrosis manifesta* lesions affecting the trochlea of the humerus and/or the distal femur were found in up to 70% of animals, whereas clinically apparent OCD lesions were noted in

only 7%.[95] Similarly, in horses, resolution of subclinical but radiographically apparent changes consistent with *osteochondrosis manifesta* of the distal intermediate ridge of the tibia, the lateral trochlear ridge of the talus, and the lateral trochlear ridge of the distal femur may occur before 5, 5, and 8 months of age, respectively. Lesions which remain present beyond these ages, however, become permanent.[172] Experimental studies, in which chondral fractures were created in the cartilage of the femoral condyles in skeletally immature rabbits demonstrated that cartilage flaps are capable of healing if they are stable and have a wide pedicle containing abundant cartilage canals. However, unstable fragments, attached only by a narrow isthmus devoid of cartilage canals, are unlikely to heal completely and result in a fragment that closely resembles OCD.[177]

The gender distribution of OC is similar between humans and swine. In humans, females are affected less frequently than males, accounting for approximately 20-40% of all cases of OC.[108;164] Similarly, in pigs the incidences of *osteochondrosis manifesta* and *osteochondrosis dissecans* are significantly lower in females.[95] Conversely, the incidence of OC in horses does not appear to differ between females and males.[37;133;178] The reason for this difference in gender predilection in the horse compared to other species is unknown.

Conclusion

A broad range of similarities exists between OC affecting humans and animals, including predilection sites, clinical presentation, radiographic/MRI changes, and histological appearance of end-stage (OCD) lesions, suggesting a shared pathogenesis among the various species. Histological findings from examination of sequential early

naturally-occurring and experimentally-induced OC lesion in young animals strongly supports that this pathogenesis is characterized by localized avascular necrosis of the epiphyseal cartilage of the AECC leading to focal retardation and/or failure of endochondral ossification. Further investigation of early, subclinical OC in human subjects using *in vivo* imaging and post mortem histological evaluation of predilection sites should confirm or refute the role of vascular compromise and necrosis of epiphyseal cartilage of the AECC in the development of OC affecting humans. If indeed, this is the case, then naturally-occurring or surgically-induced cases of OC in animals will provide an excellent opportunity to develop and test diagnostic and treatment modalities for this increasingly recognized condition.

Table 1: Disease names for manifestations of osteochondrosis at specific anatomical locations as reported in human literature. For comparison, location of predilection sites in pigs and horses is also presented.

	Disease Name (Human)	Location	Pig	Horse
Articular				
	Theimann's Disease	proximal and distal interphalangeal joints (fingers and toes)		
	Panner's Disease	elbow (humeral capitellum)	X	
	osteochondritis dissecans	elbow (humeral capitellum), knee (medial or lateral femoral condyle), ankle (medial talus)	X	X
	Freiberg's Disease	metatarsophalangeal joint (head of 2nd metatarsal)		X
Non-articular/apophyseal				
	Sinding-Larsen-Johansson Disease	knee (inferior pole of patella)		X
	Köhler's Disease	ankle (tarsal navicular bone)		
	Iselin Disease	ankle (base of 5th metatarsal)		
	medial epicondyle apophysitis	elbow (medial epicondyle)		
Physeal				
	Blount Disease (tibia vara)	proximal tibial physis		
	Scheuermann's Disease	vertebrae	X	X

Human: [99;102;106;179]
Pig: [110]
Horse: [172;180]

All figures reprinted with permission

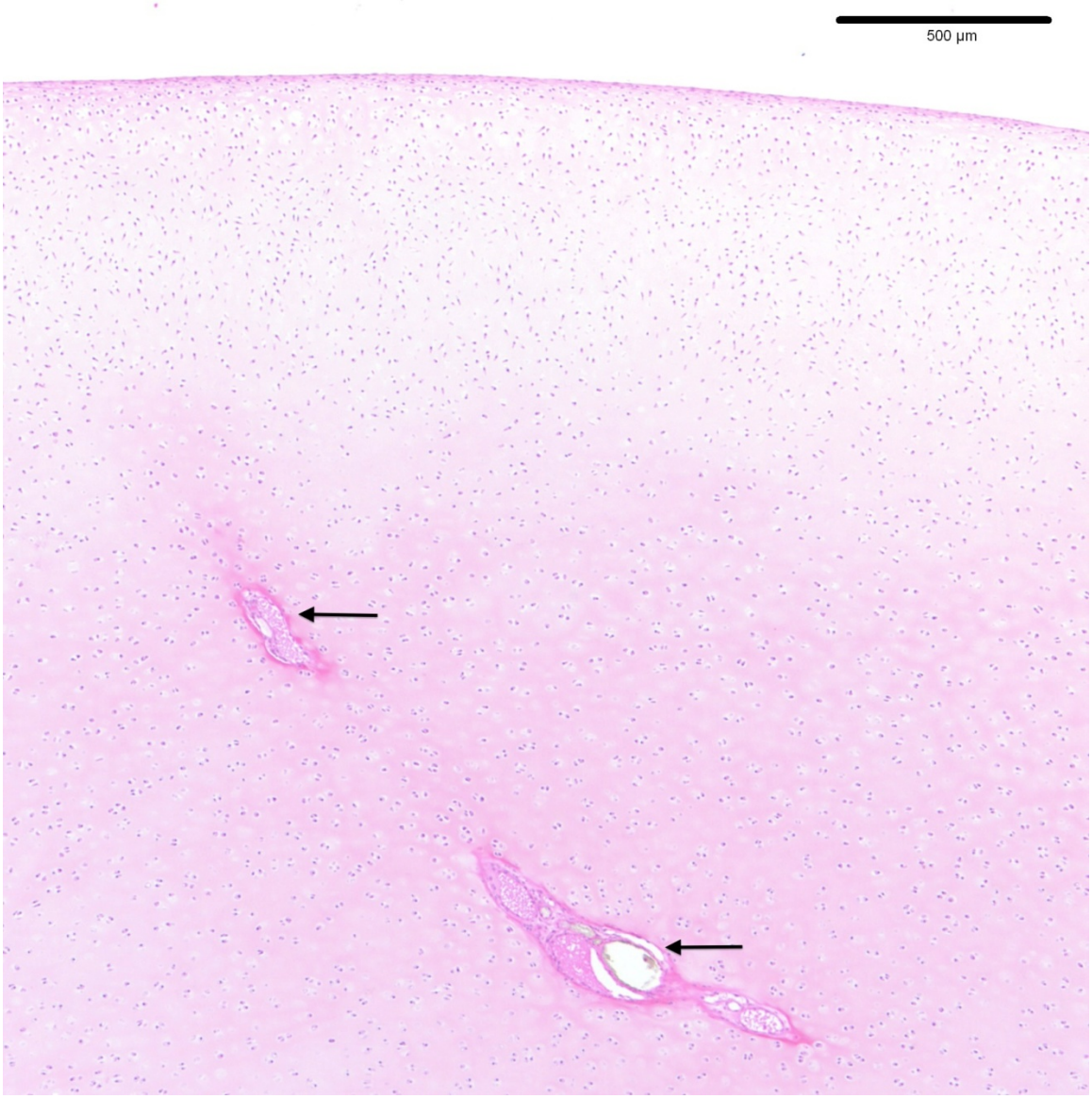


Figure 1: Photomicrograph depicting normal cartilage canals (arrows) containing blood vessels in the articular-epiphyseal cartilage complex in the medial femoral condyle of a 14-week-old pig (hematoxylin and eosin stain).

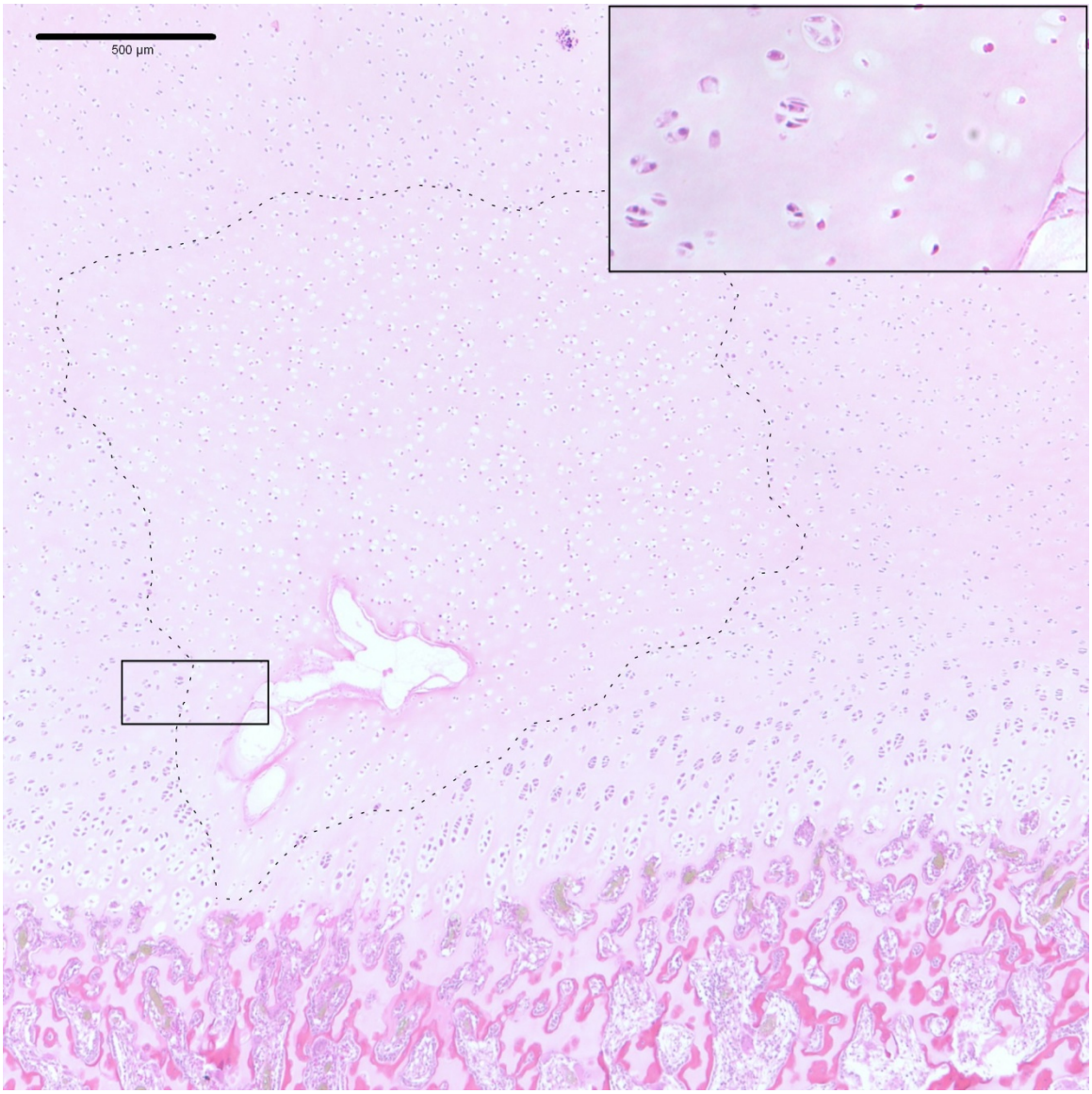


Figure 2: Photomicrograph of an *osteochondrosis latens* lesion (dashed line), including necrotic blood vessels and surrounding necrotic cartilage, involving the medial femoral condyle of a 14-week-old pig. Inset: viable chondrocytes are present in the left half of the image whereas in the right half, chondrocytes are eosinophilic and contain no obvious nucleus, consistent with chondrocyte necrosis (hematoxylin and eosin stain).

Figure 3: Photomicrograph depicting an *osteochondrosis manifesta* lesion in the medial femoral condyle of a 14-week-old pig (same animal as in Figure 2). A large area of necrotic epiphyseal cartilage is present (dashed line) and has resulted in a focal failure of endochondral ossification (arrowheads) (hematoxylin and eosin stain).

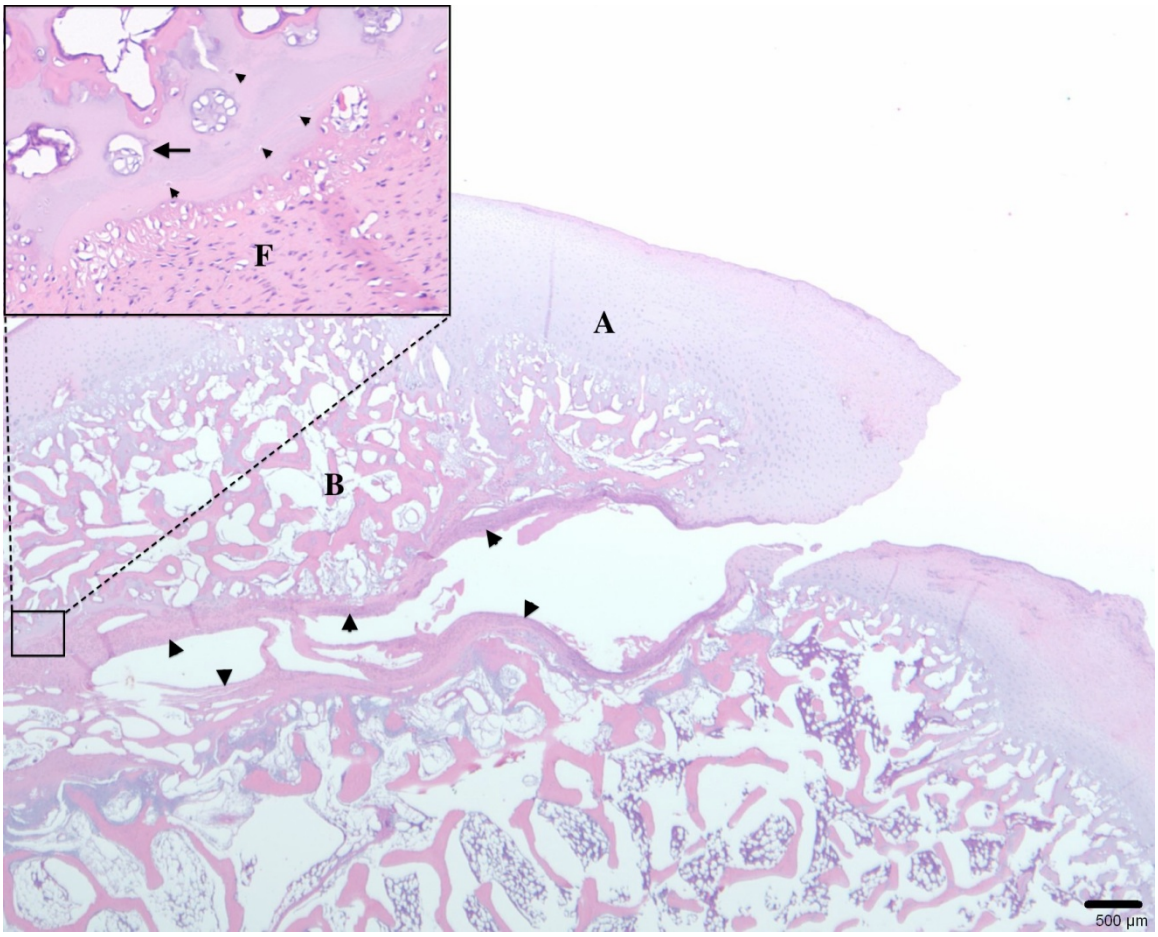


Figure 4: Photomicrograph showing an *osteochondrosis dissecans* lesion involving the medial femoral condyle of a 6-month-old pig. A fissure that is partially lined by fibrous connective tissue extends through the articular cartilage to the subchondral bone, resulting in the formation of an osteochondral cleft. A: articular-epiphyseal cartilage complex; B: subchondral bone; arrowheads: fibrous tissue. Inset: Remnants of necrotic cartilage are present adjacent to the cleft and are accompanied by chondrocyte clones. Arrowheads: necrotic chondrocytes; arrow: chondrocyte clone; F: fibrous tissue (hematoxylin and eosin stain).

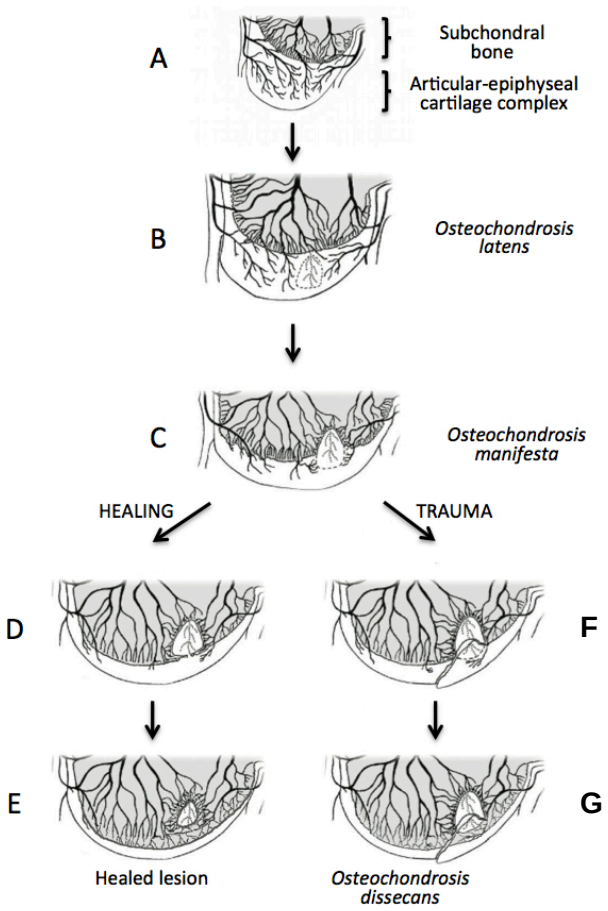


Figure 5: Diagram demonstrating the pathogenesis of *osteochondrosis dissecans* (modified with permission from Figure 7[92]). Panel A: normal endochondral ossification. Panel B: Development of *osteochondrosis latens* lesion due to failure of cartilage canal blood supply causing necrosis of the epiphyseal cartilage (circled area). Panel C: *Osteochondrosis manifesta* lesion appears as a delay in the progression of the ossification front. Panels D and E: healing of *osteochondrosis manifesta* lesion by incorporation into the subchondral bone. Panels F and G: Development of *osteochondrosis dissecans* lesion due to trauma causing collapse of the articular cartilage overlying areas of necrotic epiphyseal cartilage.

Figure 6: *Osteochondrosis dissecans* lesion involving the ankle (tibiotarsal joint). Panel A: posterio-anterior radiographic image of an *osteochondrosis dissecans lesion* (black arrow) of the talus in a juvenile human subject. Panel B: dorsomedial-plantarolateral oblique radiographic image of an *osteochondrosis dissecans* lesion (white arrow) involving the lateral trochlear ridge in a horse. Panel C: lateromedial radiographic image of an *Ooteochondrosis dissecans* lesion (black arrow) involving the distal intermediate ridge of the tibia in a horse.

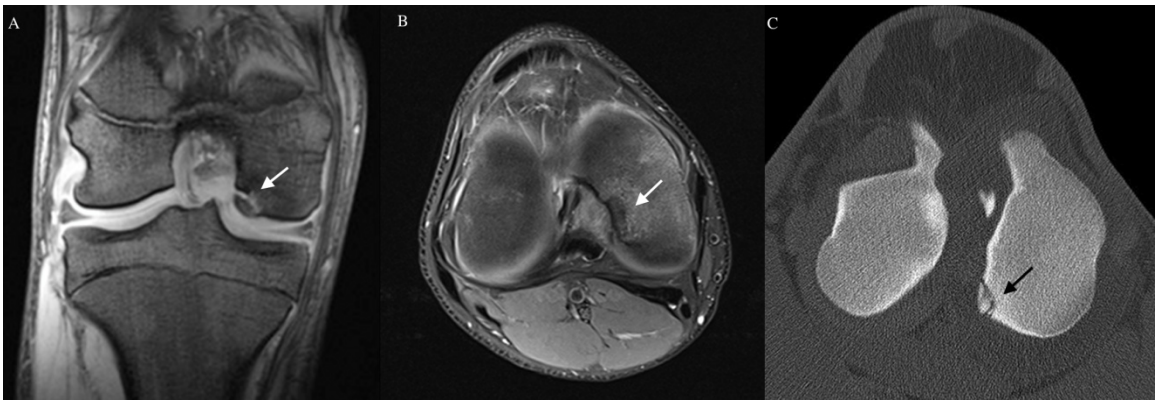


Figure 7: *Osteochondrosis dissecans* lesion of the medial femoral condyle. Panels A (coronal plane) and B (transverse plane) depict MRI findings from a human subject with *osteochondrosis dissecans* of the medial femoral condyle (white arrows). Panel C: CT image of an *osteochondrosis dissecans* lesion (black arrow) of the medial femoral condyle of a horse obtained in the transverse plain. (Image is courtesy of Dr. Bergman, VetCT-Lingehoeve Diergeneeskunde, Netherlands)

Chapter 3

Short- and Long-Term Racing Performance of Standardbred Pacers and Trotters After Early Surgical Intervention for Tarsal Osteochondrosis

Short- and Long-Term Racing Performance of Standardbred Pacers and Trotters After Early Surgical Intervention for Tarsal Osteochondrosis

Annette M. McCoy[1], Sarah L. Ralston[2], Molly E. McCue[1]

From the [1]Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA; [2]Department of Animal Science, Rutgers, The State University of New Jersey, New Brunswick, NJ

Sources of Funding: Dr. McCoy was funded by an institutional NIH T32 Comparative Medicine and Pathology Training Grant (University of Minnesota) and a Doctoral Dissertation Fellowship (University of Minnesota); partial funding for Dr. McCue is provided by NIH NIAMS 1K08AR055713-01A2.

Acknowledgements: The authors thank the owners of the included horses for their participation in this and related studies, as well as the farm veterinarians and technicians involved in daily care and sample collection. Thanks to Dr. Aaron Rendahl for advice regarding statistical analysis.

In press as:

McCoy AM, Ralston SL, McCue ME. Short- and long-term racing performance of Standardbred pacers and trotters after early surgical intervention for tarsal osteochondrosis. (2014) *Equine Veterinary Journal* doi: 10.1111/evj.12297 [Published online 12 May 2014]..

Summary

Reasons for Performing Study: Osteochondrosis (OC) is commonly diagnosed in young Standardbred racehorses, but its effect on performance when surgically treated at a young age is still incompletely understood. This is especially true for Standardbred pacers, which are underrepresented in the existing literature.

Objective: To characterize the short- (2-year-old) and long-term (through 5-year-old) racing performance in Standardbred pacers and trotters after early surgical intervention (<17 mo. of age) for tarsal OC.

Study Design: Retrospective clinical study.

Methods: The study population consisted of related, age-matched Standardbred racehorses (n=278; 151 pacers, 127 trotters) with (n=133) or without (n=145) one or more tarsal OC lesions. All OC-affected horses were treated surgically prior to being sold as yearlings. Data obtained from publically available race records for each horse included starts, wins, finishes in the top 3 (win, place, or show), earnings, and fastest time. Comparisons between OC-affected and unaffected horses were made for the entire population and within gaits. A smaller related population (n=94) had the aforementioned performance measures evaluated for their 2- through 5-year-old racing seasons.

Results: OC status was associated with few performance measures. Trotters were at higher risk for lesions of the medial malleolus, but lower risk for lesions of the distal intermediate ridge of the tibia compared to pacers. Horses with bilateral OC lesions and lateral trochlear ridge (LTR) lesions started fewer races at 2 years of age than those with unilateral lesions or without LTR lesions.

Conclusions: OC seemed to have minimal effect on racing performance in this cohort, although horses with bilateral and LTR lesions started fewer races at 2. There was evidence for different distribution of OC lesions among pacers and trotters, which should be explored further.

Potential Relevance: Standardbreds undergoing early removal of tarsal OC lesions can be expected to perform equivalently to their unaffected counterparts.

Introduction

Osteochondrosis (OC) is a widely recognized manifestation of developmental orthopedic disease characterized by disruption of normal endochondral ossification at the ends of long bones. Certain joints in the horse are considered predilection sites for OC, including the stifle, tarsus, and metacarpo-/metatarsophalangeal joints, and the relative importance of each predilection site varies by breed.[181] Osteochondrosis of the tarsocrural joint is commonly diagnosed in young Standardbred racehorses, either on routine radiographic studies prior to yearling sales or with the onset of clinical signs (i.e. joint effusion and/or mild lameness) shortly after being put into race training. Prevalence of tarsal OC in Standardbreds has been reported to range from 10.1-26.2% based on a number of radiographic surveys, and this breed is considered predisposed to lesions in this joint.[32-34;37;38]

While the high prevalence of tarsal OC is well-documented in Standardbreds, the effect of this disease - and its treatment - on racing performance has been debated in the literature. Successful racing careers have been reported for 73.5%[182] and 84%[183] of surgically treated horses, while a "good" outcome was reported in only 23% of conservatively-treated Standardbreds[184], but these reports did not include a control group for comparison. Studies that do include matched controls variably report impaired performance in OC-affected Standardbreds[24-26] or no significant differences in performance parameters between groups.[22;24;27-29;33] However, the majority of these previous studies have the limitation of either a control group made up of horses with unknown history (i.e. some could have been OC-affected), or an affected group with unknown treatment status (i.e. performance could be affected by treatment). To our

knowledge, no study has reported on performance outcome in surgically-treated OC-affected horses compared to related, age-matched OC-unaffected horses for which the entire early history was known.

Standardbred racing in North America is somewhat unique in that horses compete at either the trot (2-beat, diagonal, symmetrical) or the pace (2-beat, lateral, symmetrical), whereas in Scandinavia and continental Europe only trotters race. However, while this phenotype is well-characterized, and modern breeding distinctly separates pacing lines from trotting, limited prospective information about relative performance of horses with these biomechanically distinct gaits is available in the literature.[8;185;186] To our knowledge, there are no reports of how OC affects race performance in pacers, nor any comparisons of the relative impact of the disease on pacers versus trotters.

The purpose of this study was: 1) to report early (2-year-old season) race performance in a cohort of Standardbred yearlings that underwent surgical treatment for removal of tarsal OC fragments prior to being sold as yearlings and to compare them to related age-matched horses that were unaffected with OC; 2) to report long-term (2- through 5-year-old seasons) race performance in a subset of these horses; and 3) to compare the effect of OC on early race performance within and between pacers and trotters.

Materials and Methods

Horses: The study population for evaluating short-term race performance was comprised of 278 Standardbred horses raised on a single breeding farm in the Eastern United States that were born in 2007 (n=59 horses), 2009 (n=114), or 2010 (n=105) and

identified for inclusion in the study as yearlings being prepared for sale. Osteochondrosis (OC) lesions were radiographically identified in one or both tarsi of 133 individuals and were surgically removed prior to the sale preparation period (<17 months of age). One hundred forty-five related age-matched controls were identified, all radiographically confirmed to be free of OC. The study population for evaluating long-term race performance was comprised of 94 Standardbred horses born and raised on the same breeding farm described above (and 59 of which are included in the short-term performance cohort). Thirty-two of these individuals had surgically removed OC lesions of one or both tarsi. Of the 62 age-matched related controls, 28 were radiographically confirmed to be free of OC and 34 were presumed unaffected because of lack of clinical signs including effusion and lameness. Although complete radiographic examinations (i.e. "repository films") were not available for all individuals, horses with musculoskeletal lesions other than tarsal OC were not knowingly included in either study cohort. Information regarding foaling date, gender, sire, gait, sale, and sale price for all yearlings was obtained from the farm. Distribution of OC-affected limbs, lesion location, and surgery date for affected horses was obtained from the written veterinary records of the attending surgeon or the farm veterinarian.

Performance Records: All 278 horses in the short-term performance cohort had completed their 2-year-old season as of 31 December 2012. Ninety-four horses, born in 2007, had completed their 5-year-old season by this date. All available race records were obtained from the United States Trotting Association. Thirteen horses had been exported to Europe and records for these horses were collected from the appropriate country's trotting association. Collected data included number of starts, wins, and top 3 finishes

(win, place, or show), earnings, and fastest qualifying time over a mile for each season. The summation of starts, wins, top 3 finishes, and earnings across 4 seasons were calculated as career performance results, if applicable. Horses that started a race but did not win any money were assigned a nominal earnings value of $1 to differentiate them from horses that never started a race.

Additionally, a randomly chosen subset of all 2010 offspring from nineteen Standardbred stallions with progeny included in the study population were selected (n = 288, representing 25% of all offspring from these stallions in the 2010 foaling season) for comparison to the horses in the present study. Freely available records for these individuals were obtained from the United States Trotting Association. Collected data included yearling sale price, number of starts at 2 years, earnings at 2 years, and fastest qualifying time at 2 years. This was done to determine if the individuals in the study cohort were representative of their contemporaries.

Statistical Analysis: Two-year-old performance ("short-term performance") was evaluated in all horses (n=278) in the short-term performance study cohort. Additionally, models were fit separately for two-year-old performance in all pacers (n=151), all trotters (n=127), and all horses affected with OC (n=133). Performance for the two- through five-year-old seasons ("long-term performance") was evaluated in all horses born in 2007 (n=94). For each of these groups, outcome variables of interest were 1) whether the horse started in a given season (yes/no); 2) number of starts; 3) number of wins; 4) number of top 3 finishes (win, place, or show); 5) earnings; 6) earnings per start, 7) fastest recorded time over a mile, and 8) yearling sales price. Earnings and earnings per start were evaluated only in horses that started a race. In the long-term performance group, outcome

variables were examined by season as well as cumulatively over 4 seasons as appropriate. OC status (affected vs. unaffected) was evaluated as an outcome in the entire short-term performance cohort as well as separately in pacers and trotters. In the OC-affected cohort, lesion location (DIRT, MM, LTR) and lesion distribution (bilateral vs. unilateral) were also examined as outcome variables.

Multiple regression was performed for all outcome variables. OC status (categorical variable) was the primary predictor variable of interest in these models, but other covariates included gender, gait (pace vs. trot), number of starts, fastest recorded time, yearling sale location, and sire, as appropriate (see **Supplemental Methods** for a detailed description of multiple regression model construction). When OC status (or specific lesion location or distribution) was the outcome of interest, predictor variables included gender, gait, and sire. Analysis of 2-year-old performance in the OC-affected group was performed using multiple regression models that included lesion location (DIRT, MM, LTR) and distribution (bilateral vs. unilateral) as predictor variables. In all cases, multiple regression was performed using generalized linear regression models for binomial (quasibinomial model) and count (negative binomial model) outcome variables and ordinary linear regression for continuous outcome variables (see **Supplemental Methods**). Selected findings of interest are reported in the main text, but full results of all regression models are reported in **Supplemental Results**. Proportions (i.e. proportion of horses starting in a given season) were compared between groups using a two-sided test. Comparison between the short-term performance study cohort (n=278) and the randomly chosen 2010 offspring (n=288) was performed using a Kolmogorov-Smirnov test, which tests the entire distribution of data rather than only the population average. Examined

variables included yearling sale price, number of starts at 2 years, earnings at 2 years, and fastest qualifying time at 2 years. In all analyses, outcome variables reported as a dollar amount were log-transformed to normalize the data. No other data transformations were performed. All statistical tests were performed in the R statistical computing environment[187] using the packages 'car'[188] and 'MASS'.[189] Significance was set at $p<0.05$.

Results

Short-term performance: All horses at 2 years of age: Of the 278 horses in this group, 156 (56.1%) were colts and 122 (43.9%) were fillies. Among the 133 OC-affected horses, colts (n=81) appeared to be overrepresented compared to fillies (n=52), but the proportion of affected individuals was not significantly different between genders (Δ proportion = 0.093, 95% CI -0.032-0.218, p=0.156). There were 151 (54.3%) pacers and 127 (45.7%) trotters. Yearlings were sold at one of five breed-recognized sales held between September and November of each year; two OC-affected horses were not sold and were not included in the analysis of sale price. OC status did not significantly affect yearling sale price (OC-affected median $20,000; range $1,500-260,000; OC-unaffected median $25,000; range $1,500-260,000) in multiple regression analysis ($p = 0.266$). The remaining predictor variables in this model, gender, sale and sire, were all significantly associated with sale price ($p < 0.002$) (**Supplemental Results Tables 1 and 2**).

Performance data for the 278 horses by OC status are summarized in **Table 1**. The proportion of horses starting at least one race did not differ between OC-affected (75/133 [56.4%]) and OC-unaffected (96/145 [66.2%]) groups (Δ proportion = 0.098,

95% CI -0.023-0.22, p = 0.12). In multiple regression analysis, OC status was not significantly associated with the number of starts, wins, or top 3 finishes, fastest time, or earnings and earnings per start. Full regression analysis results are reported in **Supplemental Results Table 1**.

To determine if the individuals in the study cohort were representative of their contemporaries, yearling sale price, proportion of horses starting at 2 years, number of starts at 2 years, earnings at 2 years, and fastest qualifying time at 2 years were compared between the short-term performance group and the randomly chosen subset of 2010 foals (n=288). There were no significant differences between the two groups for any of these outcome measures (**Supplemental Results Table 3**).

Short-term performance: Pacers and trotters at 2 years of age: Irrespective of OC status, trotters were significantly less likely to start a race at 2 years than were pacers (OR 0.56, 95% CI 0.34-0.93, p = 0.026) and started significantly fewer number of races (Incident Rate Ratio [IRR] 0.54 , 95% CI 0.41-0.71, p < 0.001). On average, trotters were 2.9 sec slower than pacers (95% CI 1.71-4.15, p < 0.001). Earnings were significantly different between pacers and trotters, but this effect disappeared once earnings were adjusted for the number of starts (that is, Earnings Per Start). Gait did not significantly affect the number of wins or top 3 finishes. The proportion of OC-affected individuals was significantly higher in trotters (71/127 [55.9%]) than pacers (62/151 [41.1%]) (Δ proportion = 0.148, 95% CI 0.024-0.272, p = 0.019). When OC was examined as an outcome variable in a model that included the predictor variables gender, gait, and sire, only gait was significantly associated with affectation status (p = 0.026) (**Supplemental**

Results Table 4). Full regression analysis results are reported in **Supplemental Results Table 1**.

Performance data for the 151 pacers are summarized by OC status in **Table 2**. The proportion of horses starting at least one race did not differ between OC-affected (41/62 [69.7%]) and OC-unaffected (62/89 [66.1%]) groups (p = 0.78, 95% CI -0.13-0.2). In multiple regression analysis, OC status was not significantly associated with the number of starts, wins, or top 3 finishes, fastest time, or earnings and earnings per start (**Supplemental Results Tables 5 and 6)**.

Performance data for the 127 trotters are summarized by OC status in **Table 2**. The proportion of horses starting at least one race did not differ between OC-affected (34/71 [47.9%]) and OC-unaffected (34/56 [60.7%]) groups (Δ proportion = 0.128, 95% CI -0.051-0.317, p = 0.208). Similarly to the pacers, in multiple regression analysis, OC status was not significantly associated with the number of starts, wins, or top 3 finishes, fastest time, or earnings and earnings per start (**Supplemental Results Tables 7 and 8)**.

Short-term performance: OC-affected horses at 2 years of age: Average age at the time of surgery for the 133 OC-affected horses was 11.8 months (median 12 mo.; range 7.5-17 mo.). Lesion distribution in 132 horses (264 joints) with complete records was as follows: 134 joints (50.8%) in 93 horses (69.9%) had lesions of the distal intermediate ridge of the tibia (DIRT), 73 joints (27.7%) in 48 horses (36.4%) had lesions of the medial malleolus (MM), 49 joints (18.6%) in 37 horses (28%) had lesions of the lateral trochlear ridge (LTR), and 5 joints (1.9%) in 5 horses (3.8%) had a lesion of the medial trochlear ridge (MTR). Thirty-nine horses had two different lesions (DIRT+MM, n=15; DIRT+LTR, n=16; MM+LTR, n=4; DIRT + MTR, n=2; LTR+MTR, n=1; MM + MTR,

n=1). Six horses had three different lesions (DIRT + MM + LTR, n=5; DIRT + MM + MTR, n=1). Eight horses were noted in the surgical record to have extensive lesions (2 DIRT, 2 MM, 4 LTR). In total, there were 261 lesions in 207 joints of these 132 horses. Seventy-six of the 133 affected horses (57.1%) were affected bilaterally with one or more types of lesion, while 57 (42.9%) were unilaterally affected. When gender and gait were considered as predictors for individual lesion locations and distribution (bilateral vs. unilateral) in a regression model, trotters had significantly increased odds of being affected with a MM lesion (OR 5.01, 95% CI 2.27-11.82, $p < 0.001$) and significantly decreased odds of having a DIRT lesion (OR 0.27, 95% CI 0.11-0.62, $p = 0.003$) (**Supplemental Results Table 9**). When sire was added to the model (i.e. gender, gait, and sire as predictor variables), gait remained significantly associated with MM and DIRT lesions ($p < 0.001$ for both), but sire was significantly associated only with DIRT lesions ($p = 0.016$) (**Supplemental Results Tables 10-13**). Lesion location and distribution were not significantly associated with yearling sale price in this group (**Supplemental Results Tables 14 and 15**).

Performance data are summarized in **Table 3** by OC lesion location and distribution. Horses with a bilateral lesion (any location) started significantly fewer number of races (IRR 0.6, 95% CI 0.36-1.00, $p = 0.03$) than horses with unilateral lesions. Similarly, horses with LTR lesions started 0.56 the number of races as horses without LTR lesions (95% CI 0.32-0.98, $p = 0.033$) Other factors significantly affecting the number of starts at 2 among OC-affected horses were gender and gait, with mares and stallions starting fewer races than geldings (IRR 0.58, 95% CI 0.34-0.97, $p = 0.031$; and IRR 0.47, 95% CI 0.25-0.86, $p=0.011$, respectively), and trotters starting fewer races than

pacers (IRR 0.45, 95% CI 0.27-0.73, p = 0.001) (**Table 4**). Only three of the eight horses with "extensive" lesions started a race at 2 years (37.5%). OC lesion location and distribution were not significantly associated with number of wins or top 3 finishes, nor, when only starters were considered, with earnings or earnings per start (**Supplemental Results Table 14**).

Long-term performance: 2007 horses, 2- through 5-year-old seasons: Of the 94 horses in this group, 62 (66%) were colts and 32 (34%) were fillies. Colts appeared to be overrepresented in the OC-affected group (n=25) compared to fillies (n=7), but the proportion of affected individuals was not significantly different between genders (Δ proportion = 0.184, CI -0.008-0.403, p = 0.08, 95%). Similarly to the larger cohort reported above, OC status did not significantly affect sale price in this group. Sales price was significantly affected by sale location and sire ($p < 0.001$), but, in contrast to the larger cohort, not by gender (**Supplemental Results Tables 16 and 17**).

Performance data for the 2-year-old through 5-year-old seasons, as well as cumulative data over all four race seasons, of 94 horses are summarized by OC status in **Supplemental Table 1**. The proportion of horses starting at least one race in any season, individually or cumulatively, did not differ between OC-affected and OC-unaffected groups. When seasons were examined individually, OC status was not significantly associated with the number of starts, or wins, fastest time, or earnings and earnings per start. OC status was significantly associated with the number of top 3 finishes at 3 years - horses with OC had 1.32 times the number of top 3 finishes than horses without OC in that year (95% CI 1.10-1.59, p = 0.004) (**Supplemental Results Table 20**). When cumulative performance over four seasons was considered, however, OC status was

significantly associated with fewer wins – horses with OC won 0.76 times the number of races as those without OC (95% CI 0.6-0.97, p = 0.028) (**Supplemental Results Table 16**). Other factors that were significant in these two multiple regression models were gait (IRR 1.28, 95% CI 1.03-1.59, p=0.035 for top 3 finishes at 3 years; IRR 1.58, 95% CI 1.22-2.04, p=0.001 for cumulative wins), number of starts (IRR 1.06, 95% CI 1.04-1.07, p<0.001 for top 3 finishes at 3 years; IRR 1.016, 95% CI 1.011-1.021, p < 0.001 for cumulative wins) and fastest time (0.94, 95% CI 0.91-0.97, p < 0.001 for top 3 finishes at 3 years; 0.89, 95% CI 0.85-0.93, p < 0.001 for cumulative wins). Full regression analysis results are reported in **Supplemental Results Tables 16-21**.

Discussion

Osteochondrosis (OC) of the hock is highly prevalent in the Standardbred horse[32-34;37;38], but its effects on performance are debated in the literature.[21;22;24-29;33;182-184] Here, we report early (2-year-old) and long-term (2- through 5-year-old) race performance in a cohort of age-matched related horses raised on a single breeding farm where surgical removal of OC lesions prior to yearling sales is standard of care. To our knowledge, this is the first report of performance in a cohort of OC-affected individuals with early surgical intervention and similar breeding to matched controls with known early medical history. In the short-term performance cohort, OC status was not significantly associated with any performance measure. In the smaller long-term performance cohort, OC status was significantly associated with only two performance measures, but with opposite effects – OC-affected individuals had a higher number of top 3 finishes at 3 years, but lower cumulative wins over 4 race seasons. It is difficult to explain why OC would have

opposite effects on two measures that would be expected to be somewhat correlated. For both models, gait, number of starts, and fastest time were also significantly associated with the outcome and likely explained the largest amount of variation between individuals. It is possible that if other unmeasured factors (i.e. related to inherent racing ability) could be included in a regression model that the statistical association with OC status would disappear in this group. Since similar effects were not seen in the larger study cohort (albeit looking at a more limited time frame), it is also possible that these are spurious associations related to the relatively small sample size or were significant by chance due to the multiple testing conducted within this group. Overall, OC status seemingly had minimal effect on performance in these study cohorts.

Among OC-affected individuals, those with bilateral lesions started significantly fewer races during their 2-year-old season when compared to those with unilateral lesions. OC-affected horses with LTR lesions also started significantly fewer races at 2 years than did affected horses without lesions in this location. It has previously been reported that horses with LTR lesions were not as successful after surgery as horses with other lesions.[182] We did not detect any other effects of any specific lesion location on performance, but it is possible that the number of lesions in this group of horses (especially LTR and MM) was not large enough to detect such effects. It is of note that a smaller proportion of horses with unilateral or bilateral lesions noted to be "extensive" by the attending surgeon started a race at 2 (3 out of 8, 37.5%) when compared to the OC-affected group as a whole (75/133, 56.4%), although the number of such individuals was too small for a meaningful statistical comparison.

Reports in the literature of performance in Standardbreds affected with OC are conflicting and somewhat difficult to compare directly due to differences in cohort selection and disease definition. In horses for which the treatment status was unknown, OC-affected horses have been reported both to perform as well as their unaffected contemporaries[22;27;29;33] and to have impaired performance.[26;183] Conservative treatment of hock OC has been advocated by some based on the seemingly minimal effect of the disease on performance. Indeed, Brehm et al. reported no significant difference in number of starts, wins, or places, amount of earnings, or fastest time in a group of 147 horses with conservatively-treated tarsal OC when compared to a randomly chosen group of known unaffected controls over 3 racing seasons, although the proportion of horses starting a race was not reported for either group.[28] The proportion of horses in our long-term performance cohort with OC that started at least one race over multiple seasons (27/32; 84.4%) was higher than reported in previous survey studies[26;27], and similar to the proportion reported to race after surgical treatment of OC lesions.[24;182;183] It is impossible to say whether the affected individuals in the present cohort would have performed as well without surgery. However, arthroscopic removal of osteochondral fragments has become the standard of care for treatment of most tarsal OC in young horses because of concern over the risk for long-term degenerative changes in the joint if fragments remain in place.[23] Progressive osteoarthritis with associated pain and dysfunction has been reported in humans[190] and dogs[191] with conservatively treated OC of the ankle/tarsus, and it is logical to conclude that similar sequelae could occur in equine patients with lesions in this location, although the onset of signs may be delayed until after the end of a typical racing career.

The optimal timing of surgical intervention is a question that to our knowledge has not been definitively addressed in the literature and cannot be fully addressed by our study since all of the horses in our cohort underwent surgery prior to being sold as yearlings. Based on radiographic changes of lesion appearance between 6 and 18 months of age, an argument has been made recently for delaying surgical intervention, especially for mild to moderate lesions, to allow for spontaneous healing to occur.[192] Previous work would suggest, however, that spontaneous healing of hock OC lesions is unlikely to occur after 5 months of age, and that lesions are permanent after 11 months of age.[193] In cases where clinical signs, including effusion, are present, delayed surgical intervention decreases the chances that these signs will resolve.[194;195] In our cohort, the majority of individuals underwent surgery at 11 months of age or older; those who were treated earlier typically had moderate to severe effusion of one or both joints. Examining this question from a performance perspective, Beard et al. reported that horses undergoing arthroscopy for tarsal OC were less likely to start as 2-year-olds compared to their unaffected counterparts and suggested that this could have been due to an interrupted training schedule.[25] Certainly, it has been reported that young horses with "planned training failure" related to arthroscopy lose more training days and have a lower financial return than those not requiring such intervention.[196] The proportion of OC-affected horses in the present cohort starting as 2-year-olds (75/133; 56.4%) was markedly higher than reported by Beard et al. (22%) for surgically-treated horses.[25] This difference may be due to the fact that early surgical intervention in the present study eliminated the treatment-related training disruption that would have otherwise occurred. Although the pathophysiology and natural progression of OC should be taken into account when

making a decision about surgical intervention, these data would suggest that early removal of OC lesions – i.e. prior to yearling sales – may be desirable for Standardbred racehorses.

Our data largely support previous reports regarding performance differences between pacers and trotters. Trotters were slower than pacers, were less likely to start a race, and started fewer races in their two year old season. However, contrary to previous reports[8;9;185;186], once these factors were accounted for in multiple regression models, there were no differences in racing success as measured by wins, top 3 finishes, and earnings per start between gaits. To our knowledge, the effect of OC on performance has not previously been compared between pacers and trotters. In our population, trotters were significantly more likely to be affected with OC than were pacers. Also, among OC-affected individuals, trotters were significantly more likely to be affected with MM lesions, while pacers were significantly more likely to be affected with DIRT lesions. Since pacing is naturally exhibited by young pace-bred Standardbreds prior to the onset of training, it is possible that the biomechanical differences between pacing and trotting could affect the manifestation of OC, as well as impact its effect on performance. There are at least three reported differences in the biomechanics of the trot and the pace that may have biological significance.[197-200] An alternative explanation could be that genetic risk factors vary between pacers and trotters, leading to the differences in disease prevalence and lesion distribution. The effect of sire on OC status was evaluated in our entire short-term performance population and was not found to be significant. When only OC-affected horses were considered, sire was not significantly associated with MM lesions, but was significantly associated with DIRT lesions. We hypothesize that pacers

and trotters likely share genetic risk factors for disease and develop the same early lesions, but that biomechanical differences in their natural gait patterns may determine which lesions go on to heal and which develop into permanent OC lesions, resulting in the different lesion locations between gaits. Ideally, a prospective evaluation of a large cohort of Standardbred pacer and trotter foals would be carried out to evaluate this hypothesis. However, as OC status did not affect performance outcomes in either pacers or trotters, differences in lesion prevalence and distribution may not be clinically significant.

There are several limitations of the present study design, including the relatively small sample size, especially for the long-term performance cohort; it is possible that some of the differences between groups that did not reach significance would have done so in a larger population. Another limitation, inherent in the retrospective design, is that although we have one to four years of race data available, direct follow-up with new owners after the yearlings were sold was not possible, so the subsequent medical history of these horses is unknown. Thus, there is no way to determine if performance failure in affected horses was related to OC or not. Similarly, it is possible that horses classified as unaffected as yearlings could have gone on to develop signs related to existing, but previously undiagnosed OC lesions after being put into training, although it is unlikely that this would have happened in a large number of cases. This is of greatest concern in the long-term performance cohort, in which half of the controls did not have radiographs and were instead considered “clinically free” of OC. It is widely accepted that some horses with OC do not show any clinical signs, although the prevalence of this has not been reported. Presence or absence of effusion was not recorded for all of the OC-

affected horses in the current cohort, so it is difficult to estimate how many clinically unaffected horses may have had a lesion. The decision was made to retain the non-radiographed controls because it was felt that the larger population would have more power to detect differences between groups, but we acknowledge the possible misclassification bias in our long-term performance results. However, we did carry out the same statistical analyses reported here in the long-term performance cohort using only radiographed controls and still found that OC was significantly associated with few performance measures (fewer wins at 5 years, slower time at 4 years, more top 3 finishes at 3 years; **Supplemental Table 22**). Thus, we feel that it is unlikely that misclassification of controls affected our overall conclusions.

The overall proportion of horses in this study starting at least one race during their 2-year-old season (171/278; 61.5%) as well as the overall proportion of horses starting at least one race over multiple seasons (77/94; 81.9%) was somewhat higher than that previously reported in randomly chosen control populations[25] or general radiographic surveys of Standardbred yearlings.[26;27] While this is unlikely to have been affected by selection bias, as the horses were chosen for inclusion prior to being sold as yearlings, it does raise the question of whether the conclusions drawn from this study are specific to horses raised on this single breeding farm. To help address this, a randomly chosen subset of all 2010 offspring from nineteen Standardbred stallions with progeny also included in the short-term performance study population were selected for comparison to the horses in the present study. There was no significant difference in any examined outcome measure between the two groups. This suggests that our study population is similar to the

larger population of racing Standardbreds in North America and that it is reasonable to expect that our findings can be extrapolated to the wider population.

In summary, consistent with our hypothesis, we found that Standardbreds which underwent early removal of tarsal OC lesions performed equivalently to their unaffected counterparts during their 2-year-old race season as well as over 4 consecutive race seasons (2- through their 5-year-old). Among OC-affected individuals however, those with bilateral lesions started fewer numbers of races at 2 years than those with unilateral lesions. Similarly, horses with LTR lesions started fewer numbers of races at 2 years than those without lesions at this location. This suggests that even if bilateral lesions or LTR lesions are removed at an early age (i.e. before yearling sales, as in this group of horses), they can still negatively impact early race performance. Following a larger group of horses over several race seasons will help to determine if this effect is maintained over the long term. While pacers and trotters exhibited differences in race performance (as expected), including slower record times and fewer starts in trotters when compared to pacers, OC status did not seem to affect performance outcomes in horses of either gait. Unexpectedly, trotters were significantly more likely to be affected with OC than were pacers. When specific lesion locations were considered, trotters were more likely to exhibit MM lesions than pacers, while pacers were more likely to have DIRT lesions than trotters. Biomechanical or genetic differences between gaits may be involved in this seemingly differing manifestation of disease between groups. Further research should be conducted to validate this finding in a larger population of Standardbred racehorses.

Table 1: Summary of 2-year-old performance measures for foals in the short-term performance cohort (n=278) with (OC+) and without (OC-) surgically-treated tarsal OC lesions.

		OC+	OC-
Starting at 2		75/133 (56.4%)	96/145 (66.2%)
Sale Price	Mean	$32,010	$34,870
	Median	$20,000	$25,000
	Range	$1,500-260,000	$1,500-260,000
Starts	Mean	4.2	4.5
	Median	1	3
	Range	0-21	0-16
Wins	Mean	0.5	0.7
	Median	0	0
	Range	0-7	0-8
Top 3 Finish	Mean	5.4	5.8
	Median	1	4
	Range	0-30	0-23
Earnings (starters only)	Mean	$26,830	$28,650
	Median	$10,110	$7,742
	Range	$1-194,000	$1-531,900
Earnings per Start (starters only)	Mean	$2,935	$3,128
	Median	$1,070	$1,216
	Range	$0.25-18,450	$0.33-48,360
Fastest Time (secs)	Mean	118.6	118.2
	Median	118.0	118.2
	Range	112.6-129.6 (n=49)	111.2-125.4 (n=62)

Table 2: Summary of 2-year-old performance measures for pacers (n=151) and trotters (n=127) with (OC+) and without (OC-) surgically-treated tarsal OC lesions.

		Pacers			Trotters	
		OC+	**OC-**		**OC+**	**OC-**
Starting at 2		41/62 (66.1%)	62/89 (69.7%)		34/71 (47.9%)	34/56 (60.7%)
Sale Price	Mean	$28,700	$34,390	Mean	$34,980	$35,630
	Median	$17,000	$22,000	Median	$22,000	$25,000
	Range	$1,500-115,000	$1,500-260,000	Range	$1,500-260,000	$2,500-170,000
Starts	Mean	5.9	5.3	Mean	2.7	3.3
	Median	4.5	5	Median	0	2
	Range	0-21	0-16	Range	0-15	0-14
Wins	Mean	0.6	0.9	Mean	0.5	0.4
	Median	0	0	Median	0	0
	Range	0-6	0-8	Range	0-7	0-5
Top 3 Finish	Mean	7.5	6.8	Mean	3.5	4.2
	Median	6.5	7	Median	0	2
	Range	0-30	0-23	Range	0-21	0-21
Earnings (starters only)	Mean	$26,800	$27,720	Mean	$26,860	$30,340
	Median	$10,990	$8,901	Median	$9,330	$3,522
	Range	$1-171,100	$1-531,900	Range	$1-194,000	$1-419,000
Earnings per Start (starters only)	Mean	$2,790	$3,062	Mean	$3,109	$3,249
	Median	$905	$1,216	Median	$1,248	$1,202
	Range	$0.25-17,110	$0.5-48,360	Range	$0.50-18,450	$0.33-29,930
Fastest Time (secs)	Mean	117.9	116.8	Mean	119.6	121
	Median	117.6	116.8	Median	119.4	121.5
	Range	112.6-126.4 (n=28)	111.2-125.4 (n=42)	Range	115.8-129.6 (n=21)	115.4-125.2 (n=20)

Table 3: Summary of 2-year-old performance measures for OC-affected horses (n=133). DIRT = distal intermediate ridge of the tibia. MM = medial malleolus of the tibia. LTR = lateral trochlear ridge of the talus.

		DIRT	MM	LTR	Bilateral Lesion (any location)
Starting at 2		56/93 (60.2%)	26/48 (54.2%)	17/37 (45.9%)	39/76 (51.3%)
Sale Price	Mean	$29,520	$31,890	$33,510	$33,580
	Median	$17,000	$20,000	$24,000	$17,000
	Range	$1,500-115,000	$1,500-260,000	$2,500-115,000	$1,500-260,000
Starts	Mean	4.4	3.2	3.1	3.6
	Median	2	1	0	1
	Range	0-21	0-14	0-13	0-21
Wins	Mean	0.5	0.5	0.4	0.5
	Median	0	0	0	0
	Range	0-7	0-7	0-4	0-7
Top 3 Finish	Mean	5.7	4.1	4	4.8
	Median	2	1	0	1
	Range	0-30	0-20	0-17	0-30
Earnings (starters only)	Mean	$26,300	$27,140	$19,940	$26,730
	Median	$8,549	$4,548	$4,750	$9,825
	Range	$1-194,000	$1-194,000	$1-171,100	$1-194,000
Earnings per Start (starters only)	Mean	$2,839	$3,039	$2,324	$3,094
	Median	$909	$944	$1,228	$943
	Range	$0.25-17,640	$0.25-18,450	$1-17,110	$0.50-18,450
Fastest Time (secs)	Mean	118.9	118.3	117.8	118.7
	Median	118.0	117.8	118	117.8
	Range	112.6-129.6	113.2-129.6	114.2-121.6	112.6-129.6

Table 4: Multiple regression results for number of starts at 2 years for OC-affected horses (n=133). DIRT = distal intermediate ridge of the tibia. MM = medial malleolus of the tibia. LTR = lateral trochlear ridge of the talus; bilat = bilateral; G = gelding; M = mare; S = stallion; P = pace; T = trot. Reference states for these binomial predictor variables were unaffected (no) for lesion location and unilateral (bilat [no]) for lesion distribution.

Outcome Variable	Predictor Variable	Estimate (IRR)	2.5%	97.5%	p-value
Starts (number)	DIRT (yes)	0.922376	0.507378	1.674533	0.782
	MM (yes)	0.865829	0.462288	1.667035	0.617
	LTR (yes)	0.556934	0.321267	0.980599	**0.033**
	bilat (yes)	0.599865	0.356129	1.004136	**0.03**
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.575455	0.340365	0.969076	**0.031**
	gender (S)	0.470114	0.25938	0.860978	**0.011**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.450336	0.275834	0.728408	**0.001**

Supplemental Table 1: Summary of performance measures for the 2-year-old through 5-year-old race seasons for horses born in 2007 that had surgically-removed OC lesions (OC+) or were classified as unaffected with OC (OC-). Cumulative performance is the summation of the given measure across the 2- through 5-year-old seasons.

Race Season			**OC+**	**OC-**
Yearling	Sale Price	Mean	$20,048	$34,202
		Median	$14,000	$21,000
		Range	$1,500-87,000	$500-260,000
2-year-old	Starting at 2		21/32 (65.6%)	38/62 (61.3%)
	Starts	Mean	6.4	4.5
		Median	4.5	2.5
		Range	0-21	0-16
	Wins	Mean	0.5	0.6
		Median	0	0
		Range	0-6	0-5
	Top 3 Finish	Mean	3	2
		Median	1	1
		Range	0-17	0-11
	Earnings (starters only)	Mean	$17,835	$27,036
		Median	$10,987	$5,606
		Range	$1-81,443	$1-245,280
	Fastest Time (secs)	Mean	120.6	119.1
		Median	121.4	119.2
		Range	114.6-126.4 (n=12)	113.8-125.4 (n=27)
3-year-old	Starting at 3		24/32 (75%)	52/62 (83.9%)
	Starts	Mean	13.2	13.8
		Median	14.5	15
		Range	0-28	0-33
	Wins	Mean	2	2.1
		Median	2	1
		Range	0-6	0-10
	Top 3 Finish	Mean	6.8	5.4
		Median	6	5
		Range	0-18	0-20
	Earnings (starters only)	Mean	$41,888	$34,675
		Median	$16,946	$16,938
		Range	$1-302,340	$1-369,586
	Fastest Time	Mean	116.8	117.1

	(secs)	Median	117.3	116.5
		Range	109.2-124.2 (n=20)	110.8-124.6 (n=40)
4-year-old	Starting at 4		20/32 (62.5%)	37/62 (59.7%)
	Starts	Mean	14.5	12.3
		Median	16.5	8.5
		Range	0-35	0-39
	Wins	Mean	1.9	1.9
		Median	1.5	0
		Range	0-7	0-13
	Top 3 Finish	Mean	6.1	5.1
		Median	5	2
		Range	0-17	0-22
	Earnings (starters only)	Mean	$35,103	$27,800
		Median	$28,123	$17,738
		Range	$950-111,617	$1-148,974
	Fastest Time (secs)	Mean	116.1	117.5
		Median	115.2	117.6
		Range	111.6-124.6 (n=19)	111.8-125.2 (n=30)
5-year-old	Starting		18/32 (56.3%)	32/62 (51.6%)
	Starts	Mean	12.8	12
		Median	11	1
		Range	0-40	0-43
	Wins	Mean	1.2	1.7
		Median	0	0
		Range	0-9	0-10
	Top 3 Finish	Mean	4.2	4.5
		Median	1	0
		Range	0-18	0-18
	Earnings (starters only)	Mean	$25,112	$30,153
		Median	$12,269	$20,517
		Range	$96-127,570	$1-212,767
	Fastest Time (secs)	Mean	115.4	116
		Median	114.2	115.9
		Range	111.6-123.2 (n=13)	111.4-121 (n=26)
Cumulative	Starting		27/32 (84.4%)	54/62 (87.1%)
	Starts	Mean	46.9	42.7
		Median	41.5	36.5
		Range	0-104	0-118
	Wins	Mean	5.7	6.4

	Median	4.9	5.5
	Range	0-21	0-23
Top 3 Finish	Mean	20	17.1
	Median	16.5	14
	Range	0-51	0-52
Earnings (starters only)	Mean	$93,849	$89,332
	Median	$56,352	$55,425
	Range	$1,070-427,380	$1-570,720
Earnings per Start (starters only)	Mean	$1,805	$2,020
	Median	$909	$856
	Range	$221-$14,086	$1-19,024
Fastest Time (secs)	Mean	115.6	116.5
	Median	114.6	116.2
	Range	109.2-124.2 (n=25)	110.8-124.6 (n=46)

Supplemental Methods

Multiple regression models were fit separately for five subgroups of the study cohort:

1) All horses in the short-term performance cohort at 2 years of age (n=278);

2) Pacers at 2 years of age (n=151);

3) Trotters at 2 years of age (n=127);

4) OC-affected horses at 2 years of age (n=133);

5) All horses in the long-term performance cohort (n=94). For the long-term performance cohort, performance was evaluated for each individual race season between 2 and 5 years of age, as well as cumulatively over the four seasons.

Generalized linear regression (GLM) models were used for all binary and count outcome variables. Specifically, a quasibinomal model (logit link function) was used for binary outcomes and a negative binomial model (log link function) was used for count outcomes. These were selected rather than binomial and poisson models, respectively, to address the problem of overdispersion in the data. For negative binomial models, the parameter theta (θ) was estimated to minimize the AIC. Above $\theta = 4$, the AIC changed minimally, so this was the maximum value used in any model. Continuous outcomes were assessed for normality and examined using an ordinary linear regression model (identity link function). Variables reported as dollar amounts (earnings, earnings per start, yearling sale price) were log-transformed to normalize the data. Earnings and earnings per start were only considered in horses that started at least one race (nominal winnings of $1 were assigned to horses that started a race but did not win any money, to distinguish

them from horses that never started a race). Horses that did not sell as yearlings were excluded from evaluation of yearling sales price.

Selection of outcome and predictor variables and model assessment: Each reported model follows the general regression equation $y = \mu + \beta x + \varepsilon$, where y is the outcome variable of interest, μ is the population average, x is a predictor variable of interest (several may be included in a model), and ε is the error. The variables that were included in the various models, as either outcomes or predictors, are listed in **Supplemental Methods Table 1**.

There is an ongoing debate regarding the "best" measure of performance in racehorses, and a variety of different performance indices have been proposed. Since none of these have been widely accepted, we chose to report a variety of performance outcomes, all of which have been reported in the literature, allowing for potential comparison with other studies as well as ease of interpretation by readers. We do follow the recommendations of Cheetham et al. (2010)[9] by reporting both starts and earnings as performance outcomes for the horses in our cohort.

For all models, OC status was the primary predictor variable of interest, except in the OC-affected group. For this group, lesion location (DIRT, MM, LTR) and distribution (bilateral vs. unilateral) were the primary predictor variables of interest. Lesion location and distribution were considered as independent binary variables (yes/no) because many individuals had more than one lesion. Additional predictor variables for each model were chosen based on based on a combination of previously published literature[8;9] and clinical judgment. We also had to take into account which information was readily available from the United States Trotting Association. Gender was included in every model as a

predictor variable, and gait was included whenever pacers and trotters were assessed together. Track conditions were not included (reported by Cheetham et al., 2010[9]), as all of our horses raced on dirt tracks. Sire was used as a proxy for inherited genetic factors, and was considered especially important when assessing OC status (or specific lesion location or distribution) as the outcome of interest. We recognize that our models do not account for every possible variable that may play a role in performance (see **Discussion**). In situations where more than one model was considered for a particular outcome variable, model fit was assessed by using AIC or adjusted R-squared, and the model with the best fit was reported. In some cases, two models fit the data nearly equally well, and in these cases, the simplest model was chosen. ANOVA analysis was performed to look at the overall significance of sale location and sire when these were included as predictors in the model. This was done both to preserve anonymity and because some individual sale/sire groups were too small for meaningful independent statistical analysis.

1. *Models assessed in the short-term performance cohort at 2 years of age (n=278)*
    a. *Models for binary outcome variables (family=quasibinomial):*
        i. Started (yes/no) = μ + OC + Gender + Gait + ε
        ii. OC (yes/no) = μ + Gender + Gait + Sire + ε
    b. *Models for count outcomes (family=negative binomial):*
        i. Starts (number) = μ + OC + Gender + Gait + ε
        ii. Wins (number) = μ + OC + Gender + Gait + Starts + Fastest Time + ε

iii. Top 3 Finishes (number) = μ + OC + Gender + Gait + Starts + Fastest Time + ε

c. *Models for continuous outcome variables (family=linear):*

i. log(Earnings) = μ + OC + Gender + Gait + Starts + Fastest Time + ε

ii. log(Earnings per Start) = μ + OC + Gender + Gait + Fastest Time + ε

iii. Fastest Time = μ + OC + Gender + Gait + ε

iv. log(Yearling Sales Price) = μ + OC + Gender + Sale + Sire + ε

2. *Models assessed in pacers at 2 years of age (n=151) OR in trotters at 2 years of age (n=127)*

a. *Models for binary outcome variables (family=quasibinomial):*

i. Started (yes/no) = μ + OC + Gender + ε

b. *Models for count outcomes (family=negative binomial):*

i. Starts (number) = μ + OC + Gender + ε

ii. Wins (number) = μ + OC + Gender + Starts + Fastest Time + ε

iii. Top 3 Finishes (number) = μ + OC + Gender + Starts + Fastest Time + ε

c. *Models for continuous outcome variables (family=linear):*

i. log(Earnings) = μ + OC + Gender + Starts + Fastest Time + ε

ii. log(Earnings per Start) = μ + OC + Gender + Fastest Time + ε

iii. Fastest Time = μ + OC + Gender + ε

iv. log(Yearling Sales Price) = μ + OC + Gender + Sale + Sire + ε

3. *Models assessed in OC-affected horses at 2 years of age (n=151)*

    a. *Models for binary outcome variables (family=quasibinomial):*

        i. Started (yes/no) = μ + OC + Gender + ε

        ii. DIRT (yes/no) = μ + Gender + Gait (+ Sire) + ε

        iii. MM (yes/no) = μ + Gender + Gait (+ Sire) + ε

        iv. LTR (yes/no) = μ + Gender + Gait (+ Sire) + ε

        v. Bilateral (yes/no) = μ + Gender + Gait (+ Sire) + ε

            - Lesion-specific models were run both with and without sire included

    b. *Models for count outcomes (family=negative binomial):*

        i. Starts (number) = μ + DIRT + MM + LTR + Bilat + Gender + ε

        ii. Wins (number) = μ + DIRT + MM + LTR + Bilat + Gender + Starts + Fastest Time + ε

        iii. Top 3 Finishes (number) = μ + DIRT + MM + LTR + Bilat + Gender + Starts + Fastest Time + ε

    c. *Models for continuous outcome variables (family=linear):*

        i. log(Earnings) = μ + DIRT + MM + LTR + Bilat + Gender + Gait + Starts + Fastest Time + ε

        ii. log(Earnings per Start) = μ + DIRT + MM + LTR + Bilat + Gender + Gait + Fastest Time + ε

        iii. Fastest Time = μ + DIRT + MM + LTR + Bilat + Gender + Gait + ε

iv. log(Yearling Sales Price) = μ + DIRT + MM + LTR + Bilat + Gender + Sale + Sire + ε

4. *Models assessed in the long-term performance cohort for the 2-, 3-, 4-, and 5-year-old race seasons individually, as well as across all 4 seasons cumulatively (n=94)*

   a. *Models for binary outcome variables (family=quasibinomial):*

      i. Started (yes/no) = μ + OC + Gender + Gait + ε

      ii. OC (yes/no) = μ + Gender + Gait + Sire + ε

   b. *Models for count outcomes (family=negative binomial):*

      i. Starts (number) = μ + OC + Gender + Gait + ε

      - for number of starts at 2 years only, Sire was also included in the model

      ii. Wins (number) = μ + OC + Gender + Gait + Starts + Fastest Time + ε

      iii. Top 3 Finishes (number) = μ + OC + Gender + Gait + Starts + Fastest Time + ε

   c. *Models for continuous outcome variables (family=linear):*

      i. log(Earnings) = μ + OC + Gender + Gait + Starts + Fastest Time + ε

      ii. log(Earnings per Start) = μ + OC + Gender + Gait + Fastest Time + ε

      iii. Fastest Time = μ + OC + Gender + Gait + ε

      iv. log(Yearling Sales Price) = μ + OC + Gender + Sale + Sire + ε

Supplemental Methods Table 1: Outcome and predictor variables included in the multiple regression models. OC = osteochondrosis; DIRT = distal intermediate ridge of the tibia; MM = medial malleolus; LTR = lateral trochlear ridge. Reference states for sale location and sire are not listed to preserve anonymity.

	Reference State	Use
Binary Variables (yes/no)		
OC	unaffected (no)	outcome and predictor
Lesion location		outcome and predictor
DIRT	unaffected (no)	
MM	unaffected (no)	
LTR	unaffected (no)	
Lesion distribution		outcome and predictor
bilateral	unilateral (no)	
Started at least 1 race in a given season	did not start (no)	outcome
Categorical Variables		
Gender	G	predictor
mare (M)		
stallion (S)		
gelding (G)		
Gait	P	predictor
pace (P)		
trot (T)		
Sale location	anonymous	predictor
Sire	anonymous	predictor
Count Variables		
Starts		outcome and predictor
Wins		outcome
Top 3 finishes (win, place, or show)		outcome
Continuous variables		
Earnings (log transformed)		outcome
Earnings per start (log transformed)		outcome
Yearling sale price (log transformed)		outcome
Fastest recorded time (sec)		outcome and predictor

Supplemental Results

Multiple regression models were constructed for the following eight outcome variables: 1) whether the horse started in a given season (yes/no); 2) number of starts; 3) number of wins; 4) number of top 3 finishes (win, place, or show); 5) earnings; 6) earnings per start, 7) fastest recorded time, and 8) yearling sales price (see **Materials and Methods** and **Supplemental Methods**). Findings related to the major predictor variable of interest, osteochondrosis (OC) status (affected vs. unaffected), are reported in the main text. Selected variables other than OC status found to be significant predictors of performance outcome variables are also reported in the main text. However, for the sake of completeness, the full results of each regression model are reported in the supplemental tables below. For each model, the estimate (odds ratio [OR], incident risk ratio [IRR], or linear estimate, as appropriate), 95% confidence interval, and p-value of the included predictor variables are reported. Predictors meeting the significance threshold of $p < 0.05$ are marked in **bold**. ANOVA analysis was performed to look at the overall significance of sale location and sire when appropriate.

Supplemental Results Table 1: Multiple regression results for *Short-term performance: All horses at 2 years of age*. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.72	0.43	1.19	0.20
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.75	0.43	1.32	0.33
	gender (S)	0.64	0.32	1.28	0.21
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.56	0.34	0.93	**0.03**
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.95	0.72	1.25	0.72
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.86	0.63	1.16	0.32
	gender (S)	0.82	0.59	1.25	0.41
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.54	0.41	0.71	**<0.001**
Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.82	0.61	1.10	0.19
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.91	0.65	1.28	0.60
	gender (S)	0.99	0.66	1.47	0.95
	starts	1.15	1.11	1.20	**<0.001**
	time	0.86	0.82	0.90	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.70	1.21	2.40	**0.003**
Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.05	0.96	1.15	0.33
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.07	0.96	1.19	0.21
	gender (S)	1.04	0.92	1.18	0.53
	starts	1.15	1.13	1.16	**<0.001**
	time	0.97	0.95	0.98	**<0.001**

	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.09	0.98	1.21	0.13
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.48	0.76	2.89	0.24
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.94	0.91	4.12	0.09
	gender (S)	1.49	0.59	3.76	0.39
	starts	1.38	1.27	1.50	**<0.001**
	time	0.70	0.63	0.78	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	3.52	1.63	7.59	**0.002**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.73	0.88	3.38	0.11
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.67	0.77	3.60	0.19
	gender (S)	1.45	0.56	3.72	0.44
	time	0.72	0.65	0.81	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.13	1.00	4.55	0.05
Fastest Time (sec)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.12	-1.07	1.31	0.84
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-1.42	-2.76	-0.09	**0.04**
	gender (S)	-1.77	-3.41	-0.14	**0.03**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.93	1.71	4.15	**<0.001**
Yearling Sales Price (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.89	0.71	1.10	0.27
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.67	0.53	0.85	**0.001**
	gender (S)	0.93	0.69	1.26	0.66
	sale	see ANOVA			
	sire				

Supplemental Results Table 2: Yearling Sales Price ANOVA, all horses (n = 278).

Predictor Variable	Sum of Squares	Degrees of Freedom	F value	Pr(>F)
OC	0.774	1	1.2415	0.27
gender	7.874	2	6.3169	**0.002**
sale	39.548	4	15.8629	**<0.001**
sire	49.128	45	1.7516	**0.004**

Supplemental Results Table 3: Summary of 2-year-old performance measures for foals in the short-term performance cohort (n=278) and randomly selected 2010 foals (n=288). P-value for the outcome "starting at 2" was determined by comparing proportions using a two-sided test. P-value for the remaining outcomes was determined using a Kolmogorov-Smirnov test, which tests the entire distribution of data rather than only the population average.

		Short-Term Performance Cohort	Randomly Selected 2010 Foals	p-value
Starting at 2		171/278 (61.5%)	163/288 (56.6%)	0.27
Sale Price	Mean	$33,510	$33,250	0.90
	Median	$22,000	$20,000	
	Range	$1,500-260,000	$700-450,000	
Starts	Mean	4.3	3.5	0.36
	Median	3	1.5	
	Range	0-21	0-18	
Earnings (starters only)	Mean	$27,850	$31,740	0.23
	Median	$9,102	$7,355	
	Range	$1-531,900	$1-918,300	
Earnings per Start (starters only)	Mean	$3,043	$3,814	0.39
	Median	$1,178	$1,319	
	Range	$0.25-48,360	$0.20-91,830	
Fastest Time (secs)	Mean	118.4	117.9	0.24
	Median	118.0	117.0	
	Range	111.2-129.6 (n=111)	109.4-130.8 (n=77)	

Supplemental Results Table 4: OC Status ANOVA, all horses (n=278).

Predictor Variable	Degrees of Freedom	Deviance	Residual Degrees of Freedom	Residual Deviance	F value	Pr(>F)
NULL			277	384.87		
gender	2	3.716	275	381.16	1.6342	0.20
gait	1	5.685	274	375.47	5.0004	**0.03**
sire	45	61.968	229	313.50	1.2113	0.18

Supplemental Results Table 5: Multiple regression results for pacers (*Short-term performance: Pacers and trotters at 2 years of age*). OC = osteochondrosis; G = gelding; M = mare; S = stallion; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.84	0.41	1.71	0.62
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.80	0.36	1.75	0.58
	gender (S)	0.83	0.31	2.32	0.72
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.10	0.80	1.52	0.56
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.88	0.62	1.25	0.48
	gender (S)	0.77	0.50	1.21	0.25
Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.73	0.50	1.05	0.11
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.87	0.57	1.32	0.50
	gender (S)	0.98	0.59	1.63	0.93
	starts	1.15	1.10	1.21	**<0.001**
	time	0.86	0.81	0.91	**<0.001**

Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.01	0.92	1.11	0.90
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.05	0.95	1.16	0.38
	gender (S)	1.02	0.89	1.16	0.78
	starts	1.13	1.12	1.15	**<0.001**
	time	0.97	0.95	0.98	**<0.001**
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.45	0.72	2.91	0.30
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	2.21	1.02	4.78	**0.04**
	gender (S)	1.01	0.39	2.63	0.99
	starts	1.31	1.20	1.42	**<0.001**
	time	0.68	0.61	0.76	**<0.001**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.70	0.84	3.43	0.14
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.83	0.84	3.99	0.13
	gender (S)	0.81	0.31	2.11	0.66
	time	0.72	0.65	0.81	**<0.001**
Fastest Time (secs)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.15	-0.41	2.70	0.15
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-0.91	-2.64	0.82	0.30
	gender (S)	-2.16	-4.25	-0.07	**0.04**
Yearling Sales Price (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.97	0.71	1.33	0.86
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.63	0.45	0.89	**0.01**
	gender (S)	0.88	0.57	1.37	0.58
	sale	see ANOVA			
	sire				

Supplemental Results Table 6: Yearling Sales Price ANOVA, all pacers (n = 151).

Predictor Variable	Sum of Squares	Degrees of Freedom	F value	Pr(>F)
OC	0.02	1	0.0296	0.86
gender	5.167	2	3.8556	**0.02**
sale	21.975	4	8.1995	**<0.001**
sire	21.584	26	1.239	0.22

Supplemental Results Table 7: Multiple regression results for trotters (*Short-term performance: Pacers and trotters at 2 years of age*). OC = osteochondrosis; G = gelding; M = mare; S = stallion; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.62	0.30	1.27	0.20
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.73	0.32	1.64	0.44
	gender (S)	0.51	0.19	1.34	0.18
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.82	0.51	1.31	0.41
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.87	0.51	1.49	0.62
	gender (S)	0.93	0.50	1.78	0.83
Wins	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.03	0.85	1.83	0.92
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.98	0.52	1.85	0.95
	gender (S)	1.09	0.48	2.41	0.83
	starts	1.15	1.05	1.27	**0.01**
	time	0.86	0.77	0.96	**0.01**
Top 3 Finishes	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.19	0.99	1.44	0.08
	gender (G)	REF	n/a	n/a	n/a

	gender (M)	1.10	0.89	1.37	0.39
	gender (S)	0.91	0.69	1.19	0.50
	starts	1.21	1.17	1.25	**<0.001**
	time	1.00	0.97	1.07	0.89
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	2.33	0.57	9.49	0.23
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.51	0.30	7.52	0.61
	gender (S)	1.42	0.18	11.14	0.73
	starts	1.65	1.31	2.08	**<0.001**
	time	0.82	0.63	1.06	0.12
Earnings Per Start	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	2.07	0.47	9.05	0.32
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.28	0.24	6.97	0.77
	gender (S)	4.11	0.54	31.35	0.17
	time	0.71	0.55	0.91	**0.01**
Fastest Time	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	-1.28	-3.17	0.60	0.18
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-1.92	-4.03	0.19	0.07
	gender (S)	-1.11	-3.74	1.53	0.40
Yearling Sales Price	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.84	0.62	1.15	0.28
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.74	0.53	1.03	0.08
	gender (S)	1.02	0.67	1.56	0.92
	sale	see ANOVA			
	sire				

Supplemental Results Table 8: Yearling Sales Price ANOVA, all trotters (n = 127).

Predictor Variable	Sum of Squares	Degrees of Freedom	F value	Pr(>F)
OC	0.703	1	1.1994	0.28
gender	2.404	2	2.0511	0.13
sale	18.34	4	7.8243	**<0.001**
sire	22.756	19	2.0438	**0.01**

Supplemental Results Table 9: Horses in the OC-affected group had models fit with lesion location and distribution as outcome variables (see **Materials and Methods** and **Supplemental Methods**). DIRT = distal intermediate ridge of the tibia; MM = medial malleolus of the tibia; LTR = lateral trochlear ridge of the talus; bilat = bilateral; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
DIRT	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.75	0.31	1.79	0.52
	gender (S)	2.01	0.68	6.55	0.22
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.27	0.11	0.62	**0.003**
MM	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.15	0.48	2.82	0.75
	gender (S)	0.73	0.25	2.03	0.55
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	5.01	2.27	11.82	**<0.001**
LTR	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.08	0.46	2.56	0.87
	gender (S)	0.41	0.12	1.23	0.13
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.88	0.40	1.95	0.76

Bilat	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.49	0.22	1.11	0.09
	gender (S)	0.78	0.30	2.00	0.60
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.96	0.47	1.96	0.92

Lesion location and distribution were also examined with sire included in the model.

Supplemental Results Table 10: DIRT ANOVA, all OC-affected horses (n=133).

Predictor Variable	Degrees of Freedom	Deviance	Residual Degrees of Freedom	Residual Deviance	F value	Pr(>F)
NULL			131	160.24		
gender	2	3.133	129	157.10	1.8911	0.16
gait	1	10.053	128	147.05	12.1365	**0.001**
sire	35	51.112	93	95.94	1.7629	**0.016**

Supplemental Results Table 11: MM ANOVA, all OC-affected horses (n=133).

Predictor Variable	Degrees of Freedom	Deviance	Residual Degrees of Freedom	Residual Deviance	F value	Pr(>F)
NULL			131	173.05		
gender	2	0.747	129	172.30	0.3575	0.70
gait	1	17.280	128	155.02	16.5456	**<0.001**
sire	35	46.314	93	108.71	1.2670	0.18

Supplemental Results Table 12: LTR ANOVA, all OC-affected horses (n=133).

Predictor Variable	Degrees of Freedom	Deviance	Residual Degrees of Freedom	Residual Deviance	F value	Pr(>F)
NULL			131	156.62		
gender	2	3.528	129	153.09	1.7357	0.18
gait	1	0.098	128	152.99	0.0961	0.76
sire	35	49.914	93	103.08	1.4034	0.10

Supplemental Results Table 13: Bilateral lesion ANOVA, all OC-affected horses (n=133).

Predictor Variable	Degrees of Freedom	Deviance	Residual Degrees of Freedom	Residual Deviance	F value	Pr(>F)
NULL			132	181.65		
gender	2	3.151	130	178.50	1.3773	0.26
gait	1	0.012	129	178.49	0.0102	0.92
sire	35	42.542	94	135.95	1.0624	0.40

Supplemental Results Table 14: Multiple regression results for *Short-term performance: OC-affected horses at 2 years of age*. As all horses in this group were affected with OC, primary predictor variables of interest in this group were specific lesion location (DIRT, MM, LTR) and distribution (bilateral vs. unilateral). Reference states for these binomial predictor variables were unaffected (no) for lesion location and unilateral (bilat [no]) for lesion distribution. The reference group for effect estimates for all other categorical variables is denoted by REF. DIRT = distal intermediate ridge of the tibia; MM = medial malleolus of the tibia; LTR = lateral trochlear ridge of the talus; bilat = bilateral; time = fastest recorded time over a mile; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	DIRT (yes)	1.68	0.64	4.53	0.30
	MM (yes)	1.60	0.61	4.45	0.35
	LTR (yes)	0.68	0.27	1.67	0.40
	bilat (yes)	0.50	0.22	1.12	0.10
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.50	0.20	1.19	0.12
	gender (S)	0.55	0.20	1.50	0.25

	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.45	0.20	1.02	0.06
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	DIRT (yes)	0.92	0.51	1.67	0.78
	MM (yes)	0.87	0.46	1.67	0.62
	LTR (yes)	0.56	0.32	0.98	**0.03**
	bilat (yes)	0.60	0.36	1.00	**0.03**
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.58	0.34	0.97	**0.03**
	gender (S)	0.47	0.26	0.86	**0.01**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.45	0.28	0.73	**0.001**
Wins (number)	DIRT (yes)	0.81	0.38	1.69	0.58
	MM (yes)	0.96	0.40	2.25	0.93
	LTR (yes)	0.78	0.36	1.66	0.52
	bilat (yes)	1.49	0.82	2.74	0.19
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.11	0.57	2.17	0.76
	gender (S)	0.70	0.30	1.61	0.40
	starts	1.10	1.03	1.19	**0.01**
	time	0.85	0.78	0.93	**0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.72	0.96	3.11	0.08
Top 3 Finishes (number)	DIRT (yes)	1.11	0.94	1.31	0.22
	MM (yes)	1.07	0.89	0.28	0.48
	LTR (yes)	1.06	0.90	1.23	0.51
	bilat (yes)	1.06	0.93	1.20	0.41
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.08	0.94	1.25	0.27
	gender (S)	0.91	0.77	1.08	0.29
	starts	1.12	1.11	1.14	**<0.001**
	time	0.97	0.95	0.98	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.13	0.99	1.28	0.08
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	DIRT (yes)	0.58	-0.68	1.85	0.36
	MM (yes)	0.09	-1.30	1.48	0.89
	LTR (yes)	-0.10	-1.29	1.09	0.86
	bilat (yes)	0.51	-0.46	1.48	0.29

	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.62	-0.47	1.70	0.26
	gender (S)	-0.46	-1.77	0.85	0.48
	starts	0.27	0.16	0.39	**<0.001**
	time	-0.33	-0.46	-0.21	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.53	0.54	2.52	**0.003**
Earnings Per Start (log[$])	DIRT (yes)	0.41	-0.91	1.73	0.53
	MM (yes)	-0.28	-1.71	1.15	0.70
	LTR (yes)	-0.64	-1.83	0.55	0.28
	bilat (yes)	0.22	-0.77	1.21	0.65
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.18	-0.90	1.26	0.74
	gender (S)	-0.95	-2.24	0.34	0.14
	time	0.93	-0.03	1.89	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	-0.31	-0.44	-0.18	0.06
Fastest Time (secs)	DIRT (yes)	-0.34	-3.57	2.89	0.83
	MM (yes)	-3.00	-6.38	0.37	0.08
	LTR (yes)	-2.57	-5.37	0.23	0.07
	bilat (yes)	0.68	-1.74	3.09	0.58
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-2.51	-5.04	0.03	0.05
	gender (S)	-3.81	-6.72	-0.90	**0.01**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.02	-0.24	4.29	0.08
Yearling Sales Price (log[$])	DIRT (yes)	0.72	0.45	1.14	0.16
	MM (yes)	0.91	0.57	1.45	0.69
	LTR (yes)	1.13	0.70	1.83	0.60
	bilat (yes)	1.21	0.84	1.74	0.31
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.87	0.59	1.30	0.50
	gender (S)	1.12	0.67	1.88	0.65
	sale	see ANOVA			
	sire				

Supplemental Results Table 15: Yearling Sales Price ANOVA, all OC-affected horses (n = 133).

Predictor Variable	Sum of Squares	Degrees of Freedom	F value	Pr(>F)
DIRT (yes)	1.385	1	2.0421	0.16
MM (yes)	0.111	1	0.1644	0.69
LTR (yes)	0.186	1	0.2744	0.60
bilat (yes)	0.709	1	1.045	0.31
gender	0.769	2	0.5665	0.57
sale	13.454	4	4.9586	**0.001**
sire	30.224	36	1.2377	0.21

Long-term performance: 2007 horses, 2- through 5-year-old seasons: Performance outcome variables were examined by race season as well as cumulatively in this group (see **Materials and Methods** and **Supplemental Methods**). The results of the multiple regression model for yearling sales price is reported with the cumulative performance models below. Yearlings in this cohort were sold at one of four breed-recognized sales between September and November 2008; one OC-affected horse was not sold and was not included in the analysis of sale price.

Supplemental Results Table 16: Multiple regression results for cumulative performance, 2- through 5-year-old seasons. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.56	0.16	2.02	0.36
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.09	0.006	0.53	**0.03**
	gender (S)	0.05	0.003	0.30	**0.01**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.34	0.09	1.13	0.09
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.96	0.67	1.40	0.82
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.67	0.46	1.00	0.05
	gender (S)	0.66	0.42	1.06	0.08
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.56	0.40	0.80	**0.001**
Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.76	0.60	0.97	**0.03**
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.00	0.77	1.29	0.98
	gender (S)	1.18	0.87	1.60	0.28
	starts	1.02	1.01	1.02	**<0.001**
	time	0.89	0.85	0.93	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.58	1.22	2.04	**0.001**
Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.93	0.77	1.12	0.45
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.95	0.77	1.17	0.59
	gender (S)	1.11	0.87	1.41	0.41

	starts	1.02	1.02	1.02	**<0.001**
	time	0.91	0.88	0.94	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.35	1.11	1.65	**0.01**
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.82	0.56	1.19	0.29
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.61	1.06	2.43	**0.03**
	gender (S)	1.10	0.67	1.79	0.71
	starts	1.01	1.00	1.02	**0.002**
	time	0.73	0.69	0.78	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.83	1.89	4.22	**<0.001**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.87	0.61	1.24	0.43
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.76	1.19	2.59	**0.01**
	gender (S)	1.18	0.74	1.87	0.48
	time	0.78	0.74	0.82	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.69	1.83	3.95	**<0.001**
Fastest Time (secs)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	-0.71	-2.37	0.96	0.40
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-0.97	-2.76	0.82	0.29
	gender (S)	-1.06	-3.20	1.09	0.33
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	3.55	1.97	5.13	**<0.001**
Yearling Sales Price (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.91	0.67	1.23	0.55
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.89	0.64	1.25	0.51
	gender (S)	0.84	0.57	1.23	0.36
	sale	see ANOVA			
	sire				

Supplemental Results Table 17: Yearling Sales Price ANOVA, horses born in 2007 (n = 94).

Predictor Variable	Sum of Squares	Degrees of Freedom	F value	Pr(>F)
OC	0.1375	1	0.3658	0.55
gender	0.3709	2	0.4933	0.61
sale	20.7153	3	18.3694	**<0.001**
sire	29.5974	24	3.2807	**<0.001**

Supplemental Results Table 18: Multiple regression results for 2-year-old season. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.13	0.44	2.99	0.81
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.75	0.27	2.10	0.59
	gender (S)	0.76	0.24	2.50	0.65
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.41	0.16	0.98	0.05
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.16	0.69	1.97	0.55
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.82	0.47	1.43	0.47
	gender (S)	0.73	0.40	1.38	0.31
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.55	0.34	0.89	**0.01**

Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.91	0.45	1.76	0.78
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.96	0.47	1.94	0.90
	gender (S)	1.00	0.46	2.15	1.00
	starts	1.11	1.03	1.20	**0.01**
	time	0.88	0.81	0.96	**0.01**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.32	0.70	2.49	0.39
Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.92	0.63	1.33	0.67
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.07	0.73	1.56	0.73
	gender (S)	1.00	0.65	1.54	0.99
	starts	1.16	1.11	1.20	**<0.001**
	time	0.96	0.91	1.00	0.08
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.02	0.72	1.45	0.91
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.40	0.63	3.10	0.39
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.71	0.75	3.88	0.19
	gender (S)	1.15	0.45	2.91	0.76
	starts	1.18	1.09	1.27	**<0.001**
	time	0.76	0.69	0.84	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.24	1.09	4.60	**0.03**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.47	0.68	3.17	0.32
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.72	0.78	3.78	0.17
	gender (S)	1.09	0.44	2.71	0.85
	time	0.79	0.72	0.87	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.93	0.95	3.91	0.07
Fastest Time (secs)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.49	-1.16	4.15	0.26
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-0.42	-3.17	2.34	0.76

	gender (S)	-0.66	-3.85	2.53	0.68
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.93	-1.54	3.39	0.45

Supplemental Results Table 19: Multiple regression results for 3-year-old season. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.44	0.13	1.41	0.17
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.31	0.07	1.21	0.11
	gender (S)	0.17	0.04	0.69	**0.02**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.38	0.11	1.14	0.10
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.86	0.62	1.22	0.40
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.83	0.58	1.20	0.33
	gender (S)	0.65	0.43	1.00	0.05
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.57	0.41	0.78	**0.001**
Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.00	0.72	1.37	0.98
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.79	0.55	1.15	0.22
	gender (S)	1.05	0.68	1.30	0.83
	starts	1.05	1.02	1.07	**0.001**
	time	0.90	0.85	0.95	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.76	1.19	2.61	**0.01**

Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.32	1.10	1.59	**0.004**
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.96	0.77	1.19	0.72
	gender (S)	1.24	0.98	1.59	0.09
	starts	1.06	1.04	1.07	**<0.001**
	time	0.94	0.91	0.97	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.28	1.03	1.59	**0.04**
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.31	0.93	1.86	0.12
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.60	1.06	2.40	**0.03**
	gender (S)	1.04	0.66	1.64	0.86
	starts	1.04	1.01	1.06	**0.01**
	time	0.76	0.72	0.80	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.63	1.74	3.96	**<0.001**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.26	0.91	1.74	0.17
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.55	1.06	2.28	**0.03**
	gender (S)	1.01	0.72	1.67	0.65
	time	0.78	0.74	0.82	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.96	2.04	4.29	**<0.001**
Fastest Time (secs)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	-0.43	-2.15	1.30	0.62
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-2.66	-4.55	-0.77	**0.01**
	gender (S)	-1.36	-3.54	0.82	0.22
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	3.97	2.33	5.62	**<0.001**

Supplemental Results Table 20: Multiple regression results for 4-year-old season. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.81	0.29	2.26	0.69
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.19	0.06	0.58	**0.01**
	gender (S)	0.34	0.10	1.16	0.09
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.20	0.07	0.50	**0.001**
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	1.05	0.66	1.69	0.85
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.58	0.35	0.97	**0.04**
	gender (S)	0.73	0.41	1.33	0.29
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.56	0.36	0.88	**0.01**
Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.77	0.53	1.12	0.18
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.07	0.72	1.61	0.73
	gender (S)	0.95	0.58	1.56	0.84
	starts	1.05	1.03	1.08	**<0.001**
	time	0.93	0.87	0.93	**0.02**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.49	1.00	2.23	0.05
Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.93	0.73	1.18	0.54
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.91	0.69	1.21	0.51
	gender (S)	1.01	0.73	1.40	0.97

	starts	1.07	1.05	1.09	**<0.001**
	time	0.97	0.93	1.00	0.08
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.13	0.87	1.47	0.38
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.82	0.37	1.84	0.62
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.18	0.47	2.98	0.72
	gender (S)	0.47	0.16	1.38	0.17
	starts	1.10	1.05	1.15	**<0.001**
	time	0.78	0.69	0.88	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	3.29	1.40	7.77	**0.01**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.76	0.35	1.66	0.49
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.11	0.45	2.73	0.82
	gender (S)	0.42	0.15	1.17	0.10
	time	0.76	0.68	0.85	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	3.02	1.32	6.95	**0.01**
Fastest Time (secs)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	-1.56	-3.57	0.46	0.13
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	-1.40	-3.74	0.94	0.24
	gender (S)	-1.99	-4.63	0.66	0.14
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	3.11	1.12	5.10	**0.003**

Supplemental Results Table 21: Multiple regression results for 5-year-old season. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF.

Outcome Variable	Predictor Variable	Estimate (OR)	2.5%	97.5%	p-value
Started (yes/no)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.90	0.33	2.37	0.83
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.21	0.07	0.60	**0.01**
	gender (S)	0.43	0.13	1.38	0.16
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.36	0.14	0.87	**0.03**
Outcome Variable	**Predictor Variable**	**Estimate (IRR)**	**2.5%**	**97.5%**	**p-value**
Starts (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.94	0.56	1.61	0.83
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.58	0.33	1.02	0.07
	gender (S)	0.57	0.30	1.17	0.10
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	0.55	0.33	0.90	**0.02**
Wins (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.67	0.45	1.00	0.06
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	1.01	0.63	1.61	0.97
	gender (S)	1.34	0.81	2.24	0.27
	starts	1.05	1.02	1.08	**0.001**
	time	0.90	0.82	0.98	**0.03**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.14	0.90	2.22	0.14
Top 3 Finishes (number)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.92	0.70	1.21	0.54
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.85	0.61	1.18	0.33
	gender (S)	1.07	0.75	1.54	0.71

	starts	1.04	1.02	1.06	**<0.001**
	time	0.94	0.88	1.00	0.06
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	1.15	0.84	1.58	0.38
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**
Earnings (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.77	0.50	1.18	0.23
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.77	0.46	1.27	0.29
	gender (S)	0.94	0.54	1.65	0.82
	starts	1.05	1.02	1.08	**0.003**
	time	0.74	0.67	0.81	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.66	1.65	4.29	**<0.001**
Earnings Per Start (log[$])	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	0.78	0.52	1.18	0.23
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.81	0.50	1.30	0.36
	gender (S)	0.93	0.57	1.51	0.76
	time	0.74	0.68	0.80	**<0.001**
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.59	1.63	4.10	**<0.001**
Fastest Time (secs)	OC (no)	REF	n/a	n/a	n/a
	OC (yes)	-0.35	-2.15	1.45	0.69
	gender (G)	REF	n/a	n/a	n/a
	gender (M)	0.48	-1.63	2.58	0.65
	gender (S)	-0.17	-.233	1.99	0.87
	gait (P)	REF	n/a	n/a	n/a
	gait (T)	2.93	1.18	4.69	**0.002**

Supplemental Results Table 22: Multiple regression results for only radiographed individuals in the 2007 study cohort (n = 32 OC-affected; n= 28 OC-unaffected) for which OC was a significant predictor variable. OC = osteochondrosis; G = gelding; M = mare; S = stallion; P = pace; T = trot; OR = odds ratio; IRR = incident rate ratio; time = fastest recorded time over a mile. The reference group for effect estimates is denoted by REF. The comparable results in the full long-term performance cohort can be found in the tables indicated in the last column.

Outcome Variable	Predictor Variable	Estimate (IRR)	2.5%	97.5%	p-value	Results comparison
Wins at 5 years (number)	OC (no)	REF	n/a	n/a	n/a	Suppl. Results Table 21
	OC (yes)	0.59	0.37	0.93	**0.04**	
	gender (G)	REF	n/a	n/a	n/a	
	gender (M)	0.73	0.39	1.38	0.35	
	gender (S)	1.36	0.74	2.50	0.32	
	starts	1.06	1.03	1.10	**0.004**	
	time	0.88	0.78	1.00	0.07	
	gait (P)	REF	n/a	n/a	n/a	
	gait (T)	1.33	0.69	2.52	0.38	
Top 3 Finishes at 3 years (number)	OC (no)	REF	n/a	n/a	n/a	Suppl. Results Table 19
	OC (yes)	1.36	1.11	1.68	**0.01**	
	gender (G)	REF	n/a	n/a	n/a	
	gender (M)	1.15	0.87	1.53	0.33	
	gender (S)	1.34	1.03	1.75	**0.04**	
	starts	1.06	1.04	1.08	**<0.001**	
	time	0.96	0.92	0.99	**0.03**	
	gait (P)	REF	n/a	n/a	n/a	
	gait (T)	1.04	0.79	1.36	0.79	
Outcome Variable	**Predictor Variable**	**Estimate**	**2.5%**	**97.5%**	**p-value**	**Results comparison**
Fastest Time at 4 years (secs)	OC (no)	REF	n/a	n/a	n/a	Suppl. Results Table 20
	OC (yes)	-2.73	-5.06	-0.40	**0.02**	
	gender (G)	REF	n/a	n/a	n/a	
	gender (M)	-2.88	-5.45	-0.31	**0.03**	
	gender (S)	-4.60	-7.78	-1.42	**0.01**	
	gait (P)	REF	n/a	n/a	n/a	
	gait (T)	4.32	2.06	6.58	**0.001**	

Chapter 4

A Genome-Wide Association Study of Tarsal Osteochondrosis in North American Standardbreds

A Genome-Wide Association Study of Tarsal Osteochondrosis in North American Standardbreds

Annette M. McCoy

From the Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA

Acknowledgements: Thanks to Dr. Sarah Ralston (Rutgers, The State University of New Jersey) for assistance with sample collection.

Sources of Funding: Funding provided by the University of Minnesota Equine Center/Minnesota Racing Commission and the United States Equestrian Federation, Inc. Dr. McCoy was funded by an institutional NIH T32 Comparative Medicine and Pathology Training Grant (University of Minnesota) and a Doctoral Dissertation Fellowship (University of Minnesota); partial funding for Dr. McCue was provided by NIH NIAMS 1K08AR055713-01A2.

Summary

Osteochondrosis (OC) is a commonly diagnosed developmental orthopedic disease in the horse, as well as other domestic animal species and humans, which is characterized by abnormal cartilage within a joint that occurs secondary to focal failure of endochondral ossification. This disease is commonly diagnosed on prepurchase radiographs in yearling racehorses, although clinical signs including joint effusion and mild lameness are often not seen until an affected individual is put into work. The condition frequently requires surgical therapy and is of major economic importance to the equine industry due to the cost of treatment and lost training days. While OC is recognized across many breeds, there are differences in prevalence at various predilection sites (including the fetlock, tarsus, and stifle) between breeds. Heritability studies in Standardbreds and Warmbloods, both considered particularly prone to developing OC, suggest that as much as 50% of disease risk is inherited. However, to date, specific genes and alleles underlying risk are unknown.

Several genome-wide association studies (GWAS) for OC have been published; however, the regions of disease association reported in these studies rarely overlap. Reasons for these discrepancies may include differences in disease definition, presence of confounding environmental risk factors, failure to account for population structure, or true differences in risk alleles between breeds and/or predilection sites. To address some of the potential weaknesses of previous reports, a GWAS was performed on a cohort of individuals that were born and raised on a single breeding farm in the eastern United States. Horses with OC were identified and treated surgically prior to being sold as yearlings. The final cohort consisted of 182 horses, which were genotyped on either the

Illumina Equine SNP50 or SNP70 beadchips. These platforms share only ~45,000 single nucleotide polymorphism (SNP) markers, so BEAGLE software was used to impute missing genotypes, resulting in a final pool of ~73,000 SNPs for analysis. Genome-wide association analysis was performed using the program GEMMA, which accounts for relatedness between individuals by incorporating a marker-based relationship matrix into a mixed model. This analysis yielded five SNP markers that were moderately associated with OC status ($p \leq 5.1 \times 10^{-5}$). These markers were located within two distinct loci on ECA14, from ~16.4-18.3Mb and from ~33.6-36.2Mb. Several potential candidate genes are located within these regions, including fibroblast growth factor 1 (*FGF1*) and histone deacetylase 3 (*HDAC3*). These putative risk loci should be followed up by validation in an independent population and regional fine mapping to identify specific putative functional risk alleles.

Introduction

Osteochondrosis (OC) is most simply defined as a failure of endochondral ossification, the process by which a cartilage template becomes bone in the limbs of a growing animal. It is characterized by the presence of abnormal cartilage within a joint that may be thickened, soft or collapsed, or separated entirely from the underlying bone. In the last case, the condition is commonly referred to as osteochondrosis dissecans (OCD).[76] While OC can affect nearly any joint, certain areas of predilection are known, including the stifle (lateral trochlear ridge of the femur), hock (distal intermediate ridge of the tibia, lateral trochlear ridge of the talus, medial malleolus), and fetlock (distal dorsal mid-sagittal ridge of the third metacarpus/metatarsus).[181] It has been postulated that OC could be caused by either abnormal forces on normal cartilage or by normal forces on abnormal cartilage[201], but the exact pathophysiology is not yet completely understood. Evidence from experimental models suggests that abnormalities in vascular supply to the articular cartilage and subchondral bone at predilection sites underlie the condition[94;163] (see **Chapter 2** for a more thorough discussion of the pathophysiology of disease). Contributing factors that have been suggested include nutrition, exercise, genetics, conformation and other biomechanical factors, trauma, stress response, *in utero* environment, and hormonal interactions.[137;138]

OC is widely recognized in young horses across breeds and is of particular interest because of its potential to cause joint effusion and/or lameness in yearling horses preparing for sales and entering training. Radiographic surveys reflect a range of disease prevalence, and it is recognized that lesions are more common at one predilection site than another in different breeds (see **Chapter 1, Table 1**). In Warmbloods, for example,

average reported prevalence of OC lesions in the fetlock is 22.3%, while average prevalence in the hock and stifle are reported to be 11.5% and 7.0%, respectively.[38;82-84] By comparison, in Standardbreds, OC of the hock is most common, with an average reported prevalence of 14.7% as compared to 3.3% and 6.3% in the fetlock and stifle, respectively.[32-35;37;38;42] The reason for differences in predilection sites between breeds is unknown, but could be due to biomechanical factors related to gait and use, modifying genetic factors that vary between breeds, or a combination of the two.

Young horses affected with OC may improve with conservative therapy alone, but in many cases surgical intervention is required. Further, severe manifestations of this disease, or inadequate treatment of mild to moderate forms, can lead to long-term debilitative consequences. In these cases, OC can be career- or even life-threatening. In some reports, treated horses perform as well as their unaffected cohorts[19;24;27], but the cost of surgery and loss of training days still represent a significant economic burden to the horse industry.[196] Additionally, Thoroughbreds with OC identified on pre-purchase radiographs have been shown to command a lower price at yearling sales[81] and may be less likely to start a race as 2-year-olds.[25] In a cohort of Standardbred yearlings who were treated surgically prior to yearling sales (if affected with OC), we found that sale price did not differ between affected and unaffected individuals. Overall, short- (2-year-old season) and long-term (over five race seasons) racing performance was similar between affected and unaffected individuals. However, within the OC-affected group, horses with bilateral lesions were less likely to start a race at 2 years than those with unilateral lesions, as were individuals with lateral trochlear ridge lesions when compared to those with lesions in other locations (see **Chapter 3**).

It is generally accepted that OC is a complex disease, with both environment and genetics playing a major role in the development of lesions. Of the many environmental risk factors that have been suggested for OC[137;138], diet and exercise have been the subject of the most research. Dietary factors that have been implicated include copper deficiency[139;140], excess phosphorus[141], and excess dietary energy.[142] However, while manipulation of diet reportedly reduced the incidence of OC in a prospective study of 17 breeding farms, these adjustments alone did not result in elimination of the condition.[144] The role of exercise is even less clear-cut. In one large study, exercise affected the distribution of OC lesions within joints, but did not affect the total number of lesions.[132] Another study found that regular, but limited, exercise seemed to reduce the risk of OC development.[133] More recently, a large multi-breed field study reported that restricted and/or irregular exercise at a young age (< 2mo) were associated with more severe lesions, but that turnout in large and/or rough paddocks also increased the overall risk of being affected with OC.[202;203] Given these equivocal findings, it is perhaps unsurprising that despite widespread awareness of disease and efforts to reduce its impact via management changes, the prevalence of OC has remained nearly unchanged over the past 30 years.[35;38] The limited response to environmental management alone highlights the importance of genetics in disease development.

The genetic contribution to OC risk has been quantified in a limited number of breeds considered to be particularly prone to the condition. Heritability estimated from pedigree analysis has been reported to range from 0.19 to 0.52 in Standardbreds[32;37] and French Trotters.[42] Similarly, heritability of 0.15 to 0.46 has been reported for Warmbloods[84;86] and South German Coldbloods[204], depending on disease definition,

although slightly lower estimates were reported from a hospital population of Swedish Warmbloods.[205] Based on these reports, it can be estimated that between 15% and 52% of the global risk for developing OC can be attributed to genetic factors. Within individual progeny groups (offspring of the same sire), up to 70% of foals have reportedly been affected by OC.[32] Philipsson et al (1993) reported significantly higher incidence of hock OC in progeny of Standardbred sires known to be affected themselves with OC.[37] The variation in heritability estimates reflects the fact that OC is a complex disease, with known environmental interactions, and likely has multiple genetic alleles conferring susceptibility. Identification of genetic risk factors, in addition to environmental manipulation, will be key in efforts to reduce disease prevalence.

The presence of OC across domestic horse populations, including a feral horse population[77], as well as shared major predilection sites and lesion morphology suggests a unified underlying pathophysiology and shared genetic risk across breeds. The concept of shared genetic risk alleles, even in very divergent horse breeds, has previously been demonstrated in type I polysaccharide storage myopathy (PSSM1), in which a single mutation in *GYS1* results in glycogen storage disease in more than 30 different horse breeds.[206;207] Unlike PSSM1, however, it is unlikely that a single mutation is responsible for all OC genetic risk. Instead, genetic risk in OC is most likely due to the summation of multiple alleles at different genes (polygenic disease). In a polygenic disease model, differences in the heritability and prevalence of OC between breeds can be explained by the combination of risk alleles, along with the relative frequency of those alleles, in each breed (**Figure 1**). Within a single breed, not all individuals will have all or even any major risk alleles or share identical modifying alleles, but if the breed has a high

prevalence and high heritability of disease, it is assumed that alleles of major effect are present at high frequency.

Despite strong evidence demonstrating the heritable nature of OC, the specific genes and alleles underlying OC risk in the horse are completely unknown. Previous attempts have been made to use genome-wide association studies (GWAS) to identify the chromosomal regions harboring OC risk alleles. Three large GWAS for OC have been reported in European Warmblood breeds.[53;64;208] Additionally, one GWAS in Norwegian Standardbred trotters[65], one in Thoroughbreds[66], and one in French Trotters[63] have been recently published. These studies and follow-up fine mapping efforts[209;210] have identified multiple chromosomal loci that could potentially contribute to heritability of disease (**Table 1**). However, the findings have not been consistent across studies and investigation of only a single candidate gene has been reported based on the GWAS findings.[156] While a statistically significant association with disease was found, physiological justification for this gene's role in OC was not established and a functional allele conferring risk was not identified. The lack of agreement in these previous mapping studies may reflect confounding due to environmental risk factors and variability in phenotypic criteria for OC.

The limitations of previous GWAS may be overcome, in part, by selecting a study cohort made up of individuals with a shared early environment. This should minimize the effect of environmental confounders on disease association. Thus, the purpose of the present study was to utilize genome-wide association to establish chromosomal regions associated with tarsal osteochondrosis in a cohort of Standardbred yearlings born and raised on a single breeding farm.

Materials and Methods

Horses (**Table 2**)*:* The initial study cohort was comprised of 94 Standardbred yearlings born in 2007 and raised on a single breeding farm in the eastern United States. Individuals on this farm are bred from one of two well-defined lines, one trotting and one pacing, though all are related within thirteen generations to a single breed foundation sire. Management practices, including diet and exercise regimen, are the same for all foals at this facility during their first year of life. Prevalence of OC on this farm ranges from 10-20%, and is fairly consistent from year to year. Most individuals with radiographically-diagnosed OC lesions undergo surgical correction prior to being sold as yearlings, although horses with very mild lesions are treated conservatively. Yearlings were identified for inclusion in the study during preparation for one of three breed-recognized sales events. Thirty-two horses had surgically-confirmed OC lesions in one or both tarsi, while 62 horses were identified as related age-matched controls. Twenty-eight of these controls were radiographically confirmed to be free of OC and 34 were presumed unaffected because of lack of clinical signs including effusion and lameness.

To address concerns about the small sample size of the original group, additional horses from the same breeding farm were added to the study cohort over ensuing seasons. The final study cohort included 182 individuals, nearly double the size of the original group. In addition to the 94 yearlings born in 2007, individuals were included from the 2009 (n=16), 2010 (n=52), and 2012 (n=20) foal crops. Thus, a total of 70 OC-affected horses were included in this group, with 112 related age-matched controls. All controls collected after 2007 were radiographically confirmed to be free of OC.

DNA Isolation and Whole-Genome Genotyping: Blood (2007 and 2012 foals) or hair roots (2009 and 2010 foals) were collected for the purpose of DNA extraction while the yearlings were housed at a single sale preparation facility. DNA was isolated from collected samples using the Gentra® Puregene® Blood Kit (Qiagen, Valencia, CA) per manufacturer recommendations. Briefly, for blood samples, RBC lysis solution was added to samples at a 3:1 ratio, incubated, and centrifuged. After discarding the supernatant, Cell lysis solution was added to the white blood cell pellet and the cells were re-suspended, after which protein was precipitated and discarded. DNA was precipitated in isopropanol and subsequently washed in ethanol prior to final hydration. A similar protocol was followed for hair root samples, omitting the RBC lysis step. Quantity and purity of extracted DNA were assessed using spectrophotometric readings at 260 and 280nm (NanoDrop 1000, Thermo Scientific, Wilmington, DE).

Genome-wide genotyping of single nucleotide polymorphism (SNP) markers was performed by Neogen GeneSeek (Lincoln, NE) using the Illumina Custom Infinum SNP genotyping platform. Samples from the 2007 foal crop were genotyped at 54,602 SNPs using the first generation Illumina Equine SNP50 chip, while the remaining samples were genotyped at 65,157 SNP markers using the second generation Illumina Equine SNP70 chip.

Genotype Imputation: For analyses which combined horses genotyped on the Equine SNP50 and the Equine SNP70 chips, genotype imputation was performed (see **Chapter 5**). The two equine genotyping platforms share 45,703 SNPs. This shared set of markers can be extracted and the files merged into a single data set, but data from tens of thousands of markers is lost. Genotype imputation is a technique that statistically

estimates genotypes from non-assayed SNPs by comparing haplotype blocks in the study population with haplotype blocks in a more densely genotyped reference population. A pipeline for imputation of equine genotyping data was established and validated utilizing BEAGLE[211] software for imputation (see **Chapter 5**). Using this pipeline, imputation was performed in the 2007 cohort for the ~18,000 markers unique to the Equine SNP70 chip, while imputation was performed in the remaining samples for the ~9,000 markers unique to the Equine SNP50 chip. Resulting imputed files were merged with the original data files using the --merge command in PLINK.[212]

Genome-Wide Association (GWA) Analysis: Two different GWA studies were performed; an initial analysis utilizing only the 2007 cohort, and a second GWA using the final cohort of horses from 2007, 2009, 2010, and 2012. The initial GWA in the cohort of 94 horses genotyped on the Equine SNP50 chip was performed utilizing two approaches. Initially, a logistic regression model was performed in PLINK (--logistic). To control for multiple testing, 10,000 t-max label-swapping permutations were applied (--mperm 10000). SNPs were pruned for minor allele frequency (MAF) below 1% and genotyping success below 90% (--maf 0.01, --geno 0.1). To account for relatedness among the individuals in the cohort, the analysis was repeated using sire and gender as covariates in the model (--covar, --sex). However, because accounting for sire alone did not reflect the entirety of the relationships among horses, GWA was also performed using ROADTRIPS[213], which incorporates a pedigree-based relatedness matrix into a mixed model of association. A four-generation pedigree was constructed for each member of the study cohort and the KinInbcoef utility (freely available at http://www.stat.uchicago.edu/~mcpeek/software/ KinInbcoef/index.html) was used to

generate the relatedness matrix. Horses were identified as a member of one of two families, depending on whether they were bred from the trotter or pacer line. A population-level disease prevalence of 0.12 was included in the model based on the weighted average of reports in the literature available at the time.

The second GWA in the final cohort of 182 horses from all four foaling seasons was carried out after imputation using GEMMA (Genome-wide Efficient Mixed Model Analysis) software.[214] GEMMA offers several advantages over the previous approaches. It accounts for population structure through the use of a marker-based relatedness matrix, can incorporate covariates into the model (not possible using ROADTRIPS), and is highly efficient for large data sets. Additionally, it estimates the variance for each SNP rather than estimating average variance across all SNPs (i.e. in the similar program EMMAX[215]). This approach is more appropriate, and improves power, when looking for a few SNPs of moderate to major effect, as opposed to many SNPs of small effect. The GWA was performed using the options to create a centered relatedness matrix (-gk 2) and perform all three possible frequentist tests: Wald, likelihood ratio, and score (-fa 4). The analysis was performed both with and without incorporation into the mixed model (-c) of a covariate file including gender and gait (pacer or trotter). The relatedness matrix was constructed using a linkage-disequilibrium (LD)-pruned set of markers (100 SNP windows, sliding by 25 SNPs along the genome, pruned at $r^2 > 0.2$; PLINK command --indep-pairwise 100 25 0.2).[59] SNPs were pruned prior to GWA using the default GEMMA parameters of MAF <1% and missingness <95%.

Association plots were generated using the base graphics package in the R statistical computing environment.[187] Based on previously published guidelines,

uncorrected p-values of less than 5×10^{-7} were considered to indicate genome-wide significant association, while uncorrected p-values between 5×10^{-5} and 5×10^{-7} were considered to indicate moderate association.[67] When permutations were applied, a p-value of <0.05 was considered to be genome-wide significant.

Results

GWA Results for Horses Born in 2007: After pruning, 46,624 SNPs were available for analysis in PLINK. After simple logistic regression analysis, there were no SNPs reaching genome-wide significance (**Figure 2A**). The most significantly associated SNP was located on ECA6 (chr6.24462083; $p = 2.14 \times 10^{-4}$). The minor allele had an odds ratio (OR) of 5.12 (95%CI 2.16-12.15), with a frequency of 0.42 and 0.17 in cases and controls, respectively. The next eight most significant SNPs were adjacent to this marker, delineating an approximately 400kb region of interest on ECA6. This region contained two named genes (*TRAF3IP1* [TNF receptor-associated factor 3 interacting protein 1] and *ASB1* [ankyrin repeat and SOCS box containing 1]), with an additional twelve genes within 500kb on either side. When 10,000 label swapping permutations were applied, the corrected p-values for the nine top hits on ECA6 ranged from 0.458-0.996. **Table 3** gives the position, OR, and uncorrected p-value for the top 50 SNPs in the unpermuted logistic regression analysis.

When sire and gender were included as covariates in the logistic regression model in PLINK, there were, again, no markers reaching genome-wide significance (**Figure 2B**). The SNP most significantly associated with disease in this model was located on ECA21 (chr21.55458111; $p = 1.02 \times 10^{-4}$). Six SNPs from a 30kb region on ECA2,

followed by four of the SNPs from the previously identified region on ECA6 were the next most significantly associated markers. The ECA2 SNPs were located adjacent to the gene *INTS9* (integrator complex subunit 9). **Table 4** gives the position, OR, and uncorrected p-value for the top 50 SNPs in this logistic regression analysis with covariates.

The mixed model analysis in ROADTRIPS revealed ten SNPs that showed moderate evidence of association with OC status ($p \leq 5 \times 10^{-5}$ as determined by RW, the ROADTRIPS version of the W_{QLS} test) (**Figure 3**). Six of these SNPs were on ECAX, while the other four were located within the previously identified region on ECA6. Five of the six ECAX SNPs were within 5kb of each other (~79.98Mb), with a single SNP located 3Mb away (chrX:76454458). The single SNP was within the gene *DIAPH2* (diaphanous-related formin 2) while the clustered SNPs were ~10kb from *ARMCX2* (armadillo repeat containing, X-linked 2). **Table 5** gives the position and uncorrected p-values using three different test statistics for the top 50 SNPs in this mixed model analysis.

GWA Results for Final Study Cohort: Horses born after 2007 were genotyped on the second generation equine SNP chip (Illumina Equine SNP70). In order to combine these data with data from the original cohort (obtained on the Illumina Equine SNP50) without the loss of marker information, genotype imputation was performed (see Materials and Methods). Imputation was successfully carried out in the final study cohort (n=182) using the pipeline described above and in **Chapter 5**. After imputation, there were 74,595 markers in the complete data set, an increase of nearly 29,000 markers over the shared set. SNP pruning for MAF and genotyping success was subsequently

performed during the course of the GEMMA mixed model analysis as described above. After pruning, 61,046 SNPs were available for GWA analysis in GEMMA. One individual (unaffected with OC) was removed due to missing data, resulting in a final study cohort of 181 horses. Inclusion of gender and gait as covariates did not alter the analysis, so the results of the simpler model are presented here. The mixed model analysis in GEMMA revealed five SNPs on ECA14 that showed moderate evidence of association with OC status ($p \leq 5.1 \times 10^{-5}$ as determined by the likelihood ratio test) (**Figure 4**). Several of these SNPs were found within the top 50 hits from the analyses in the original study cohort (**Table 7**). An additional 24 SNPs on 12 chromosomes were within one order of magnitude of this level of significance. **Table 6** gives the uncorrected p-values using three different test statistics for the top 50 SNPs in this mixed model analysis. All five of the top SNPs were located on ECA14, but at two distinct loci. Four SNPs were loosely clustered between ~16.4-17.8Mb (with a slightly less significant hit at ~18.3Mb), while a single SNP was located at 34.2Mb. This single SNP, located within the gene *ARHGAP26* (Rho GTPase activating protein 26), was flanked by three less significantly associated SNPs, suggesting a second region of interest between ~33.6-36.2Mb. Forty-two named genes, 13 predicted pseudogenes, and 3 non-coding RNAs were located within the two regions of interest on ECA14 (**Table 8**).

Discussion

GWA analysis in the final cohort of 181 individuals identified five SNP markers within two loci on ECA14 that were moderately associated ($p \leq 5.1 \times 10^{-5}$) with OC status. These regions have not been identified as significantly associated with OC in any

previously published GWAS (the previously reported association on ECA14 in French Trotters spanned a region from 67-79Mb[63]). Although the most highly associated SNPs in the final cohort could be found within the top 50 hits in previous analyses within a smaller cohort (n=94), their statistical significance was markedly lower. The progression of results from the various approaches to GWA in this study population demonstrate the profound effect that population size and structure, as well as the number of included markers, can have on association analysis and underscores the importance of appropriate study design.

Population structure is a concern in association analysis because of the risk of false positives.[216] Alleles shared by affected related individuals are as likely to be due simply to the fact that they share a common ancestor (identity-by-descent) as to their common disease state. Although the study population described here had the advantage of a shared early environment between cases and controls, thus reducing the confounding effects of management factors such as diet and exercise, horses within trotter and pacer lines were highly related to each other. A variety of approaches were used to try to account for population structure in this study cohort, including label-swapping permutations in PLINK, the inclusion of sire as a covariate in PLINK, use of a pedigree-based relationship matrix in ROADTRIPS, and finally the use of a marker-based relationship matrix in GEMMA. Although pedigree information tracing back to a single common ancestor was available for all individuals in this study cohort, information from only four generations of individuals could be included in the ROADTRIPS relationship matrix. Additionally, some individuals in the third and fourth antecedent generations were unknown and could not be included in the pedigree. It is possible that these missing

data could have affected the association analysis. The creation of a marker-based relationship matrix does not rely on knowing any information about individual pedigrees and is therefore thought to be the better choice for populations in which the pedigree is only a few generations deep, or when there are missing individuals.[217] Furthermore, marker-based relationship matrices have been demonstrated to control for false positives better than pedigree-based ones in association studies of complex traits.[217]

A major potential limitation of this study cohort is the incomplete phenotyping of a subset of the horses identified as controls from the 2007 foaling season. Of sixty-two controls in this group, 34 were assumed unaffected based on lack of clinical signs alone, while the remaining 28 were confirmed to be free of OC using radiographic examination. As individuals with OC may not demonstrate clinical signs (e.g. joint effusion, mild lameness) prior to being put into training, it is possible that some of these horses may have gone on to be diagnosed with OC. This misclassification could either result in spurious associations with disease or could mask true associations with disease. The inclusion of additional, radiographed controls (n=50) in the final study cohort should reduce the impact of any misclassification bias in the original cohort; however, ideally, enough additional controls would be added to this group that the non-radiographed controls could be removed without dramatically reducing overall study population size.

Markers identified as being associated with disease in a GWAS are unlikely to be the variants truly conferring genetic risk. This is due in large part to the fact these SNPs were chosen for inclusion in the genotyping panel based on their frequency within the population and their distribution across the genome rather than on their location within protein-coding genes. Instead, genotyped variants are likely "tagging" true risk variants

with which they are in linkage disequilibrium (LD).[54;56;58] Horses exhibit extensive LD, and Standardbreds in particular have the greatest long-range LD (> 1,200kb) among horse breeds.[59] Thus, it is reasonable that a SNP demonstrated to be associated with disease in a GWAS could be reflecting the effects of a risk variant up to 1Mb distant (or farther) from that SNP marker.

Within the region of interest from ~16.4-18.3Mb on ECA14, there are three genes that could be considered potential candidates for playing a role in OC risk (**Table 8**). Methionine adenosyltransferase II beta (*MAT2B*) is located 0.3Mb from the most highly associated SNP in the GWAS (chr14.16401778). *MAT2B* catalyzes the synthesis of S-adenosylmethionine (SAMe). Methionine is an essential amino acid in normal skeletogenesis[218], and exogenous SAMe is utilized therapeutically for osteoarthritis because of its beneficial effects on cartilage, including increased proteoglycan synthesis.[219] Hyaluronan-mediated motility receptor (*HMMR*) and cyclin G1 (*CCNG1*) are located just downstream of *MAT2B*. *HMMR* is a hyaluronan-binding protein that has been identified in epiphyseal cartilage, articular cartilage, and interzone cells (located in what will become the joint space) in the developing joints of embryonic chicks, and is believed to play a major role in synovial joint formation.[220] Although the role of *CCNG1* in cartilage has not been reported, members of the cyclin family have been reported to regulate chondrocyte proliferation[221;222], and cyclin-dependent kinase inhibitors have been shown to mediate growth arrest in chondrocytes.[223]

The region of interest on ECA14 from ~ 33.6-36.2Mb is gene-dense, and there are four genes that could be considered potential biologic candidates for OC risk (**Table 8**). These include nuclear receptor subfamily 3, group C, member 1 (*NR3C1*, a

glucocorticoid receptor), rho GTPase activating protein 26 (*ARHGAP26*), fibroblast growth factor 1 (*FGF1*), and histone deacetylase 3 (*HDAC3*). An additional potential candidate gene, solute carrier family 35, member A4 (*SLC35A4*), is located just outside of this region at ~ 36.3Mb. *NR3C1*, which is known to interact with relaxin, a hormone important in bone remodeling, was demonstrated to be expressed in the developing maxilla and mandible of mice, although its role in long bones has not been reported.[224] Rho GTPases play an important role in chondrocyte differentiation and normal long bone development, and GTPase activating proteins, such as *ARHCAP26* are crucial mediators of their activity.[225] Although *ARHCAP26* is an interesting potential candidate gene for OC risk, the highly associated SNP within the gene (chr14.3284113) is unlikely to be of any functional significance because it is located within a large intron. *FGF1*, while best known for its role in regulating bone growth, is also one of the primary growth factors present in developing growth plate cartilage.[226] Conditional knockout of *HDAC3* in osteochondral progenitor cells in mice resulted in impaired endochondral ossification and a "runted" phenotype[227], while knockout specifically within chondrocytes resulted in cells that were smaller than their wild-type counterparts and produced less extracellular matrix.[228] *SLC35A4* is thought to be a member of the UDP-galactose transporter family and thus may play an important role in the formation of normal keratin sulfate chains (an important component of articular cartilage).[229] Mutations in the related gene *SLC35A3* have been linked to arthrogryposis in a human family and complex vertebral malformation in cattle.[230;231]

Although the genes described above are located within and near the chromosomal regions on ECA14 most highly associated with OC status, and have known functions that

make them potential candidates for involvement in disease risk, it is possible that genes located near less significantly associated markers on other chromosomes may play a role in disease risk as well. Increasing the marker density in the GWA may help to resolve this question. A new commercial SNP chip with ~670,000 markers is currently under development, and imputation using a population genotyped on this chip, or even to whole-genome sequencing data, could be performed with the data from this study cohort in the future. It will be important to validate the findings of the reported GWAS in one or more independent populations before expending a significant amount of resources on following up potential candidate genes. An appropriate second population in which to follow up these results might be similar to the one reported by Lykkjen et al. (2010)[65]; that is, a group of Standardbreds phenotyped for tarsal OC. However, eventually, validation of results should be attempted in Standardbreds with OC lesions in joints other than the tarsus, and in another breed (i.e. Warmblood) affected by tarsal OC. The goal of this would be to determine if identified putative risk alleles are specific to the Standardbred breed, or to tarsal OC, or are universal risk alleles for OC (i.e. across all predilection sites and breeds).

Table 1: Other loci associated with OC in published reports. FT = French Trotter; STB = Standardbred; TB = Thoroughbred; WB = Warmblood

Chromosome	OC location	Breed	Reference(s)
1	hock	STB	Lykkjen et al., 2010
2	fetlock, hock	WB	Dierks et al., 2010
3	fetlock, hock, stifle	STB, TB, FT, WB	Lykkjen et al., 2010; Teyssèdre et al., 2012; Corbin et al., 2012; Orr et al., 2013
4	fetlock, hock, stifle	STB, TB	Lykkjen et al., 2010; Corbin et al., 2012
5	fetlock, hock	STB, WB	Lampe et al., 2009; Lykkjen et al., 2010
9	hock	STB	Lykkjen et al., 2010
10	hock	STB, WB	Lykkjen et al., 2010; Orr et al., 2013
13	fetlock, global	FT	Teyssèdre et al., 2012
14	hock, global	FT	Teyssèdre et al., 2012
15	fetlock, global	FT	Teyssèdre et al., 2012
18	fetlock, hock, stifle	STB, TB, WB	Lampe et al., 2009; Lykkjen et al., 2010; Corbin et al., 2012
27	hock	STB	Lykkjen et al., 2010
28	hock	STB	Lykkjen et al., 2010

Table 2: Summary of included horses by foaling year, OC status, and genotyping array. SNP50 = Illumina Equine SNP50 chip (54,602 markers); SNP70 = Illumina Equine SNP70 chip (65,157 markers).

Year Foaled	Cases	Controls (radiographs)	Controls (no radiographs)	Genotyping Array
2007	32	28	34	SNP50
2009	8	8	n/a	SNP70
2010	20	32	n/a	SNP70
2012	10	10	n/a	SNP70
TOTAL	70	78	34	182

Table 3: Top 50 SNPs from PLINK logistic regression in 94 individuals (not permuted, no covariates). After pruning, analysis included 46,624 SNPs. CHR = chromosome; BP = base pair; OR = odds ratio; SE = standard error; L95 = lower boundary of 95% confidence interval; U95 = upper boundary of 95% confidence interval; P = p-value.

RANK	CHR	BP	OR	SE	L95	U95	P
1	6	24462083	5.12	0.4411	2.157	12.15	0.000214
2	6	24528853	5.12	0.4411	2.157	12.15	0.000214
3	6	24535638	5.12	0.4411	2.157	12.15	0.000214
4	6	24537029	5.12	0.4411	2.157	12.15	0.000214
5	6	24515086	0.2719	0.3693	0.1319	0.5609	0.000422
6	6	24491977	4.008	0.3987	1.835	8.755	0.000498
7	6	24703959	4.681	0.4486	1.943	11.28	0.00058
8	6	24719195	4.681	0.4486	1.943	11.28	0.00058
9	6	24335046	4.124	0.443	1.731	9.827	0.001381
10	16	53524828	3.165	0.3633	1.553	6.45	0.001516
11	24	38034532	7.181	0.6271	2.101	24.55	0.001669
12	15	87906913	0.1719	0.5651	0.0568	0.5205	0.001836
13	15	27839435	0.3256	0.3642	0.1594	0.6648	0.002064
14	10	73629937	3.133	0.3718	1.512	6.491	0.00213
15	16	73940069	0.2535	0.4496	0.105	0.6119	0.002269
16	8	67438493	3.352	0.3968	1.54	7.296	0.002305
17	16	50579676	4.045	0.4586	1.646	9.937	0.00231
18	16	50839833	4.045	0.4586	1.646	9.937	0.00231
19	16	51098759	4.045	0.4586	1.646	9.937	0.00231
20	1	1.23E+08	2.645	0.3213	1.409	4.965	0.002468
21	7	9337110	2.972	0.3598	1.468	6.016	0.002468
22	6	26582849	2.701	0.3292	1.417	5.15	0.00254
23	14	17626659	0.2349	0.4819	0.09136	0.604	0.002645
24	16	63504296	3.622	0.4291	1.562	8.398	0.002704
25	10	73000129	0.3691	0.3326	0.1923	0.7083	0.002728
26	24	41552201	2.878	0.3542	1.438	5.763	0.002838
27	24	41563627	2.878	0.3542	1.438	5.763	0.002838
28	16	58842993	3.924	0.4631	1.583	9.726	0.003159
29	10	73096930	0.3429	0.3644	0.1679	0.7005	0.003315
30	10	73172265	0.3429	0.3644	0.1679	0.7005	0.003315
31	10	73509922	0.3429	0.3644	0.1679	0.7005	0.003315
32	16	74211968	2.837	0.3556	1.413	5.695	0.003368
33	19	1116053	4.618	0.5226	1.658	12.86	0.003412
34	19	1121542	4.618	0.5226	1.658	12.86	0.003412

35	16	74441109	3.001	0.3757	1.437	6.267	0.003446
36	6	26929498	2.784	0.3536	1.392	5.567	0.003792
37	28	15461258	0.3525	0.3603	0.174	0.7143	0.003808
38	16	30303982	0.3607	0.3529	0.1806	0.7204	0.00386
39	20	15154550	3.953	0.4778	1.55	10.08	0.004017
40	32	76454485	4.859	0.5511	1.65	14.31	0.004125
41	6	26302299	2.752	0.3548	1.373	5.515	0.004328
42	15	28682409	0.3408	0.3782	0.1624	0.7152	0.004425
43	28	25786535	0.2205	0.5325	0.07764	0.6261	0.004523
44	6	23408249	2.795	0.3631	1.372	5.694	0.004647
45	10	72307543	0.3692	0.3523	0.1851	0.7364	0.004674
46	22	35045474	4.74	0.552	1.607	13.98	0.004821
47	14	16839269	3.769	0.4715	1.496	9.498	0.004892
48	9	3280613	0.3829	0.3412	0.1962	0.7474	0.0049
49	6	23454143	2.967	0.3898	1.382	6.37	0.005273
50	15	28105809	0.3182	0.4105	0.1423	0.7115	0.005282

Table 4: Top 50 SNPs from PLINK logistic regression in 94 individuals including sire and gender as covariates (not permuted). After pruning, analysis included 46,624 SNPs. See Table 3 for complete legend.

RANK	CHR	BP	OR	SE	L95	U95	P
1	21	55458111	0.0657	0.7006	0.01664	0.2594	0.000102
2	2	57424190	0.05795	0.7552	0.01319	0.2546	0.000162
3	2	57427278	0.05795	0.7552	0.01319	0.2546	0.000162
4	2	57430837	0.07514	0.6931	0.01931	0.2923	0.000188
5	2	57432244	0.07514	0.6931	0.01931	0.2923	0.000188
6	2	57444117	0.07514	0.6931	0.01931	0.2923	0.000188
7	2	57444354	0.07514	0.6931	0.01931	0.2923	0.000188
8	6	24462083	9.396	0.6176	2.801	31.53	0.000286
9	6	24528853	9.396	0.6176	2.801	31.53	0.000286
10	6	24535638	9.396	0.6176	2.801	31.53	0.000286
11	6	24537029	9.396	0.6176	2.801	31.53	0.000286
12	28	25961150	0.0877	0.6739	0.02341	0.3286	0.000305
13	28	15461258	0.134	0.5639	0.04438	0.4047	0.000365
14	10	73629937	9.559	0.6353	2.752	33.2	0.00038
15	24	41552201	8.736	0.6116	2.635	28.96	0.000394
16	24	41563627	8.736	0.6116	2.635	28.96	0.000394
17	28	26449798	0.1101	0.6237	0.03243	0.3739	0.000404
18	1	1.19E+08	10.55	0.6669	2.855	38.99	0.000411
19	6	24515086	0.1556	0.5319	0.05487	0.4414	0.00047
20	1	1.08E+08	0.0714	0.7576	0.01617	0.3152	0.000494
21	28	25672835	0.1566	0.5341	0.05496	0.446	0.000518
22	15	36771362	0.08285	0.7179	0.02029	0.3383	0.000521
23	15	37308319	0.08285	0.7179	0.02029	0.3383	0.000521
24	28	25961303	0.09746	0.6713	0.02615	0.3633	0.000524
25	10	81200428	8.279	0.6098	2.506	27.35	0.000527
26	1	1.19E+08	16.24	0.8054	3.349	78.72	0.000539
27	1	1.19E+08	16.24	0.8054	3.349	78.72	0.000539
28	12	10689357	0.1464	0.5565	0.04921	0.4359	0.000556
29	16	50579676	30.58	0.9923	4.373	213.8	0.000567
30	16	50839833	30.58	0.9923	4.373	213.8	0.000567
31	16	51098759	30.58	0.9923	4.373	213.8	0.000567
32	21	55460684	0.07191	0.7663	0.01601	0.3229	0.000593
33	8	24137705	0.1592	0.5352	0.05578	0.4546	0.000597
34	22	42508442	11.23	0.7046	2.823	44.7	0.000598
35	1	1.08E+08	0.07407	0.7618	0.01664	0.3297	0.000635
36	1	1.08E+08	0.1036	0.664	0.02821	0.3809	0.000641

37	1	1.08E+08	0.1015	0.6708	0.02725	0.378	0.000649
38	1	1.08E+08	0.1015	0.6708	0.02725	0.378	0.000649
39	16	73940069	0.0773	0.7546	0.01761	0.3392	0.000693
40	15	87906913	0.05158	0.8749	0.009285	0.2866	0.000703
41	6	24703959	7.522	0.5959	2.339	24.18	0.000708
42	6	24719195	7.522	0.5959	2.339	24.18	0.000708
43	10	13745244	6.603	0.5579	2.212	19.71	0.000717
44	12	10689242	0.1573	0.5489	0.05366	0.4614	0.000754
45	6	24491977	6.532	0.5571	2.192	19.47	0.000755
46	28	26150205	0.03849	0.9679	0.005773	0.2566	0.000764
47	28	26155894	0.03849	0.9679	0.005773	0.2566	0.000764
48	14	17626659	0.05214	0.8784	0.009322	0.2916	0.000772
49	1	1.08E+08	0.1015	0.6815	0.0267	0.386	0.000788
50	24	42096866	5.418	0.5076	2.003	14.65	0.000873

Table 5: Top 50 SNPs from ROADTRIPS mixed model analysis in 94 individuals using a 4-generation pedigree split by gait. Analysis included 54,602 SNPs. CHR = chromosome; BP = base pair; RM = p-value for ROADTRIPS version of M_{QLS} test; RCHI = p-value for ROADTRIPS version of the corrected X_2 test; RW = p-value for ROADTRIPS version of the W_{QLS} test.

RANK	CHR	BP	RM	RCHI	RW
1	chrX	79979738	8.93E-06	1.56E-05	1.39E-05
2	chrX	76454485	2.99E-06	4.44E-06	1.63E-05
3	chrX	79979058	1.04E-05	1.50E-05	2.42E-05
4	chrX	79979395	1.04E-05	1.50E-05	2.42E-05
5	chrX	79983234	1.04E-05	1.50E-05	2.42E-05
6	chrX	79983620	1.04E-05	1.50E-05	2.42E-05
7	chr6	24462083	0.000458738	6.30E-05	4.12E-05
8	chr6	24528853	0.000458738	6.30E-05	4.12E-05
9	chr6	24535638	0.000458738	6.30E-05	4.12E-05
10	chr6	24537029	0.000458738	6.30E-05	4.12E-05
11	chrX	84859782	3.73E-05	3.53E-05	7.07E-05
12	chrX	66006952	0.00123319	3.08E-05	7.42E-05
13	chr24	38034532	0.0011476	0.000121002	9.31E-05
14	chr15	87906913	0.000105162	0.00011991	0.00011291
15	chr6	24515086	0.000889196	5.49E-05	0.000121304
16	chrX	16579194	0.00110409	0.0013979	0.000151411
17	chr8	67438493	0.000586249	0.000275008	0.000159886
18	chrX	67327021	0.0017205	0.000109417	0.00021431
19	chrX	67146611	0.000494866	7.25E-05	0.00021877
20	chrX	16979524	4.93E-05	0.000122651	0.000234181
21	chr6	24491977	0.0011008	0.000126378	0.00023817
22	chr14	90314622	0.000524895	0.000154027	0.000281945
23	chr6	24703959	0.00239662	0.000487405	0.000288255
24	chr6	24719195	0.00239662	0.000487405	0.000288255
25	chr20	15154550	0.00168439	0.000808136	0.000408051
26	chr6	26582849	0.00431685	0.000466756	0.000418638
27	chr20	31358687	0.0109585	0.000802162	0.000427655
28	chr28	15461258	0.032141	0.00155196	0.000455821
29	chrX	68799846	0.00441644	0.00036499	0.00046004
30	chr10	75235811	0.00191524	0.000524314	0.000474589
31	chr10	73000129	0.000216858	0.000490672	0.000558533
32	chr10	73629937	0.00555815	0.00102957	0.000571296

33	chr14	16401778	0.000152713	0.000395805	0.000586718
34	chr6	26302299	0.00241311	0.00113469	0.00059129
35	chr20	27710724	0.00478353	0.00176603	0.000611443
36	chr24	41552201	0.000756999	0.00101535	0.000617785
37	chr24	41563627	0.000756999	0.00101535	0.000617785
38	chrX	73500683	0.000211849	0.000626891	0.000626346
39	chr24	38284330	0.00306638	0.00110106	0.000655918
40	chr24	38483281	0.00306638	0.00110106	0.000655918
41	chr1	122968083	0.000856651	0.00036499	0.000662519
42	chr10	16474237	0.00906775	0.00331612	0.000666629
43	chr10	13745244	0.00343848	0.00139705	0.000757968
44	chr14	17626659	0.000132491	0.000350335	0.000784994
45	chrX	120790249	0.0244048	0.00176524	0.000794547
46	chr16	50579676	0.00582464	0.000595886	0.000796011
47	chr16	50839833	0.00582464	0.000595886	0.000796011
48	chr16	51098759	0.00582464	0.000595886	0.000796011
49	chr15	27839435	0.00109997	0.000734196	0.000807059
50	chr7	9337110	0.00126035	0.000632697	0.000809381

Table 6: Top 50 SNPs from GEMMA mixed model analysis in 181 individuals (no covariates). After pruning, analysis included 61,046 SNPs. Uncorrected p-values are presented for the Wald test, the Likelihood ratio test (lrt) and the Score test. CHR = chromosome; BP = base pair.

RANK	CHR	BP	p_wald	p_lrt	p_score
1	14	16401778	1.04E-05	7.99E-06	2.34E-05
2	14	34284113	1.83E-05	1.43E-05	3.77E-05
3	14	17858976	3.49E-05	2.77E-05	6.50E-05
4	14	17866794	3.49E-05	2.77E-05	6.50E-05
5	14	17626659	6.41E-05	5.17E-05	1.09E-04
6	10	56558910	1.02E-04	8.32E-05	1.63E-04
7	14	33630011	1.09E-04	8.91E-05	1.73E-04
8	21	54501469	1.83E-04	1.51E-04	2.72E-04
9	14	34366588	1.87E-04	1.55E-04	2.77E-04
10	14	17534553	1.98E-04	1.64E-04	2.91E-04
11	1	1.18E+08	2.17E-04	1.80E-04	3.16E-04
12	14	36214363	2.21E-04	1.83E-04	3.21E-04
13	4	28769871	2.52E-04	2.10E-04	3.61E-04
14	21	48322513	3.00E-04	2.51E-04	4.21E-04
15	32	1.08E+08	3.25E-04	2.88E-04	4.74E-04
16	10	58040174	3.65E-04	3.07E-04	5.02E-04
17	2	99965882	3.93E-04	3.32E-04	5.37E-04
18	10	72307543	4.59E-04	3.89E-04	6.16E-04
19	15	28682409	4.21E-04	4.01E-04	6.33E-04
20	24	38866310	4.87E-04	4.13E-04	6.49E-04
21	20	16930188	5.55E-04	4.72E-04	7.30E-04
22	2	61394335	5.63E-04	4.78E-04	7.39E-04
23	20	55342664	5.70E-04	4.85E-04	7.48E-04
24	16	50579676	5.80E-04	4.93E-04	7.59E-04
25	16	50839833	5.80E-04	4.93E-04	7.59E-04
26	16	51040831	5.80E-04	4.93E-04	7.59E-04
27	16	51098759	5.80E-04	4.93E-04	7.59E-04
28	28	26310499	5.66E-04	5.01E-04	7.70E-04
29	14	18305845	6.06E-04	5.16E-04	7.90E-04
30	7	11538216	6.22E-04	5.34E-04	8.15E-04
31	4	9497538	6.46E-04	5.51E-04	8.37E-04
32	6	76057683	6.52E-04	5.56E-04	8.44E-04
33	20	55142573	7.07E-04	6.04E-04	9.08E-04
34	2	10012882	7.31E-04	6.25E-04	9.36E-04

35	15	21562368	7.34E-04	6.27E-04	9.39E-04
36	2	35143572	3.04E-04	6.29E-04	9.47E-04
37	24	42096866	7.38E-04	6.31E-04	9.44E-04
38	2	25893580	8.03E-04	6.88E-04	1.02E-03
39	21	52679697	8.03E-04	6.89E-04	1.02E-03
40	32	1.19E+08	8.13E-04	6.97E-04	1.03E-03
41	6	74791020	8.15E-04	6.99E-04	1.03E-03
42	14	36525721	8.45E-04	7.26E-04	1.07E-03
43	4	1137614	8.69E-04	8.25E-04	1.20E-03
44	4	30108889	9.84E-04	8.47E-04	1.22E-03
45	24	41552201	1.04E-03	8.98E-04	1.29E-03
46	24	41563627	1.04E-03	8.98E-04	1.29E-03
47	15	27775658	1.05E-03	9.07E-04	1.30E-03
48	15	27839435	1.05E-03	9.07E-04	1.30E-03
49	14	26778965	1.07E-03	9.26E-04	1.33E-03
50	14	26781798	1.07E-03	9.26E-04	1.33E-03

Table 7: Comparison of hits shared across analyses in the original (n = 94) and final (n = 181) study cohorts. CHR = chromosome, BP = base pair; COV = covariates (sire and gender included in one PLINK analysis). P-value for GEMMA analysis based on likelihood ratio test, p-value for ROADTRIPS based on modified W_{QLS} test (RW).

CHR	BP	GEMMA (final)		ROADTRIPS (original)		PLINK NO COV (original)		PLINK + COV (original)	
		RANK	P	RANK	P	RANK	P	RANK	P
14	16401778	1	$8x10^{-6}$	33	0.00058				
	16839269					47	0.0049		
	17858976	3	$2.8x10^{-5}$						
	17866794	4	$2.8x10^{-5}$						
	17626659	5	$5.2x10^{-5}$	44	0.00078	23	0.0026	48	0.00077
1	118288481	11	0.00018						
	118823177							26	0.00054
	118829656							27	0.00054
	119086697							18	0.00041
	122968083			41	0.00066	20	0.0025		
10	72307543	18	0.00039			45	0.0047		
	73000129			31	0.00056	25	0.0027		
	73629937			32	0.00057	14	0.0021	14	0.00038
15	28682409	19	0.0004			42	0.0044		
	27839435	48	0.00091	49	0.00081	13	0.0021		
16	50579676	24	0.00049	46	0.0008	17	0.0023	29	0.00057
	50839833	25	0.00049	47	0.0008	18	0.0023	30	0.00057
	51098759	27	0.00049	48	0.0008	10	0.0023	31	0.00057
24	42096866	37	0.00063					50	0.00087
	41552201	45	0.0009	36	0.0006	26	0.0028	15	0.00039
	41563627	46	0.0009	37	0.0006	27	0.0028	16	0.00039
6	24462083			7	$4x10^{-5}$	1	0.00021	8	0.00029
	24528853			8	$4x10^{-5}$	2	0.00021	9	0.00029
	24535638			9	$4x10^{-5}$	3	0.00021	10	0.00029
	24537029			10	$4x10^{-5}$	4	0.00021	11	0.00029
	24515086			15	0.00012	5	0.00042	19	0.00047
	24491977			21	0.00024	6	0.0005	45	0.00076
	24703959			23	0.00029	7	0.00058	41	0.00071
	24719195			24	0.00029	8	0.00058	42	0.00071
	26582849			26	0.00042	22	0.0025		
	26302299			34	0.00059	41	0.0043		

Table 8: Named genes located within the top regions of association on ECA14 from the GWAS in the final study cohort (n=181). Markers in **bold** are associated with p-values ≤ 5 x 10^{-5}. Genes in **bold** are considered potential candidate genes based on annotated function. Only predicted protein-coding genes are listed.

Region on ECA14	Markers	Genes within region	Genes within 1Mb of region
~16.4-18.3Mb	**14.16401778**	***MAT2B***, ***HMMR***, *NUDCD2*, ***CCNG1***, *GABRG2*, *GABRA1*, *GABRA6*, *GABRB2*	*ATP10B*
	14.17534553		
	14.17626659		
	14.17858976		
	14.17866794		
	14.18305845		
~33.6-36.2Mb	14.33630011	***NR3C1***, ***ARHGAP26***, ***FGF1***, *SPRY4*, *GNPDA1*, *NDFIP1*, *KIAA0141*, *RNF14*, *PCDH12*, *PCDH1*, *ARAP3*, *FCHSD1*, *RELL2*, ***HDAC3***, *DIAPH1*, *PCDHGC5*, *PCDHGB4*, *PCDHGA2*, *PCDHGA1*, *SLC25A2*, *PCDHB10*, *PCDHB16*, *PCDHB7*, *PCDHB2*, *PCDHA12*, *PCDHB14*, *PCDHB15*, *PCDHB3*, *PCDHB1*, *PCDHA3*, *PCDHA1*, *PCDHAC2*, *PCDHB11*, *PCDHA8*	*NDUFA2*, *SRA1*, *ZMAT2*, *CD14*, *ANKHD1*, *HARS2*, *HARS*, *DND1*, *WDR55*, *TMCO6*, ***SLC35A4***, *APBB3*, *SLC4A9*, *HBEGF*, *CYSTM1*, *PFDN1*, *NRG2*, *PSD2*, *CXXC5*, *UBE2D2*, *KCTD16*, *YIPF5*
	14.34284113		
	14.34366588		
	14.36214363		

Figure 1: Schematic of a hypothetical polygenic genetic risk model for OC shared across breeds. In this model, there are 9 alleles that confer genetic risk for OC development. Two of these are major risk alleles (A and C), shared across breeds and responsible for a large portion of genetic risk. The other 7 alleles are termed "modifying risk alleles" (B, D, E, F, G, H, I). In this model, each breed has a unique combination of modifying risk alleles. Modifying alleles have a minor effect on disease risk on their own, but in combination with the major risk alleles may be responsible for the differences in disease manifestation between breeds. Closely related breeds (e.g. Quarter Horse and Thoroughbred) are more likely to share some modifying risk alleles than are more distantly related breeds (e.g. Quarter Horse and Warmblood).

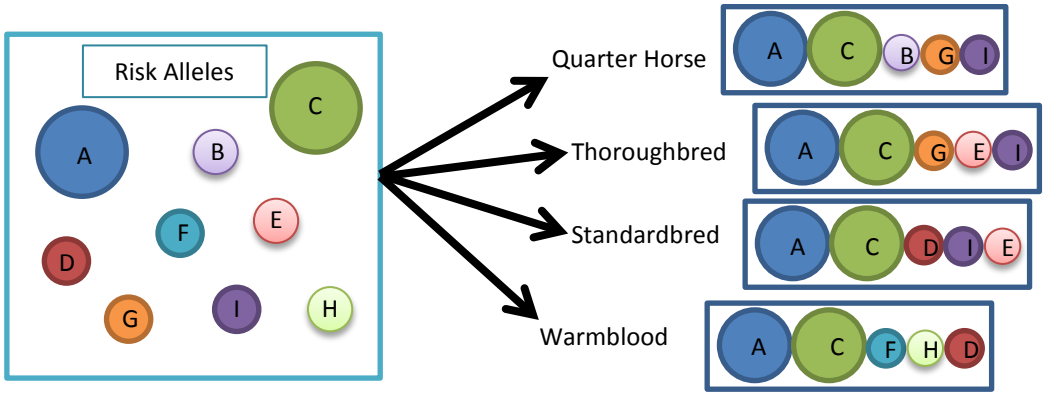
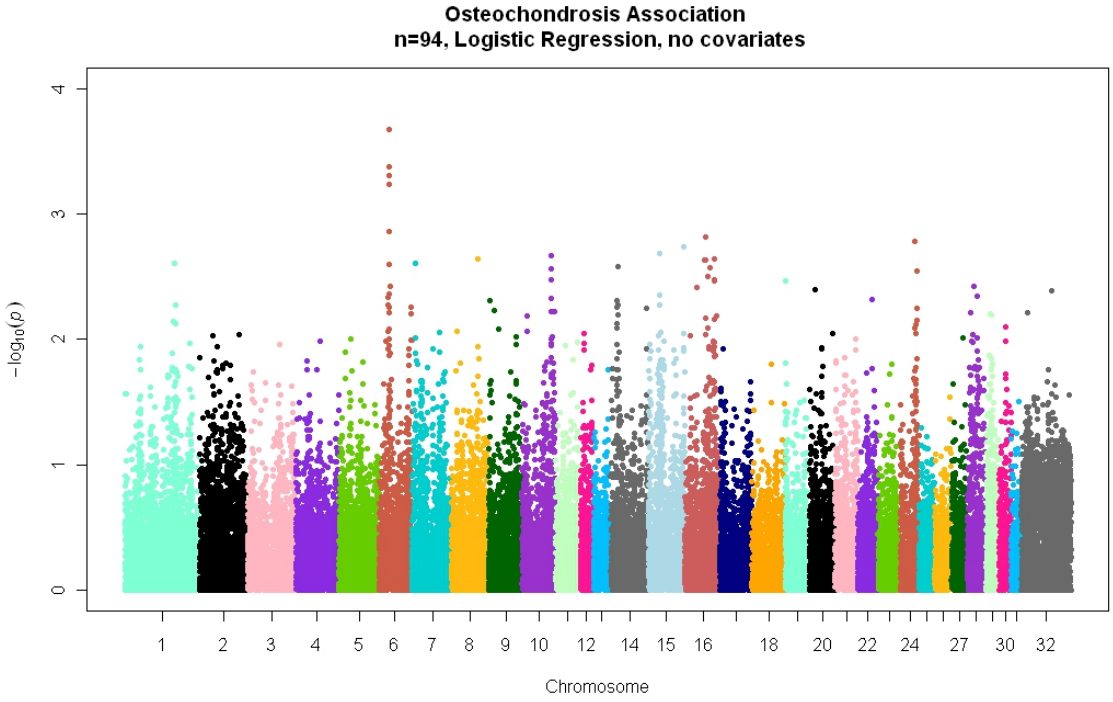


Figure 2: Manhattan plot of results from logistic regression analysis in PLINK (no permutations). A) no covariates; B) sire and gender covariates. The 31 autosomal and X chromosome (32) are represented in different colors along the x-axis and the –log(p-value) is on the y-axis. Each colored dot represents a SNP. Top hits vary between analyses and do not reach genome-wide significance. See Tables 2 and 3 for specific SNPs and p-values.

A)



B)

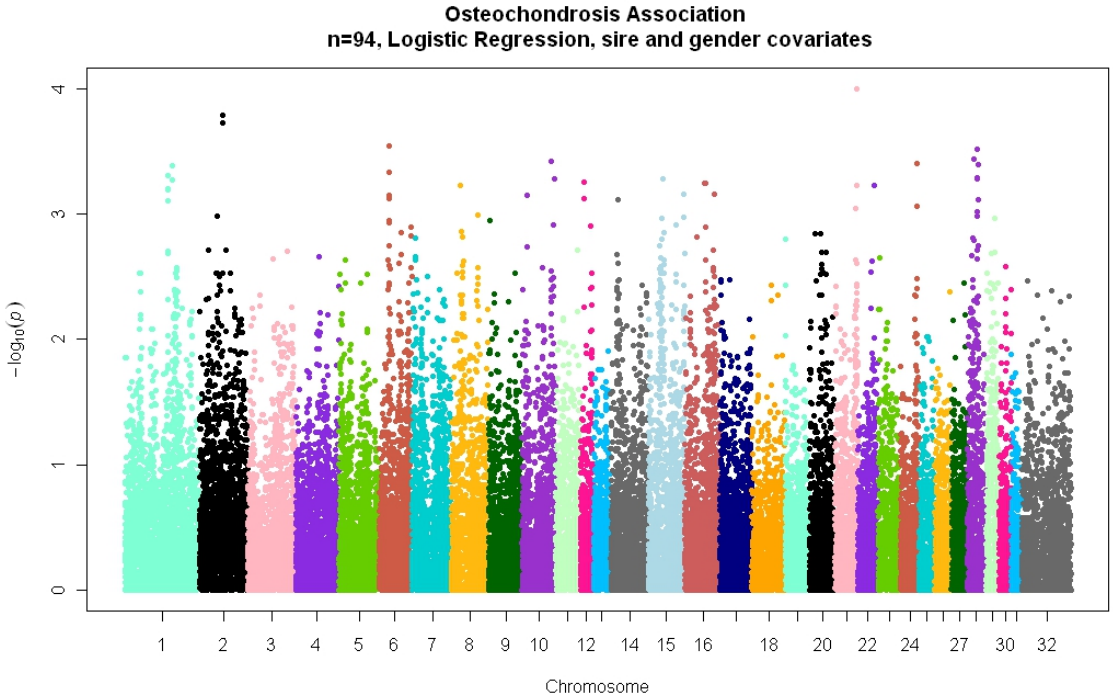

Osteochondrosis Association
n=94, Logistic Regression, sire and gender covariates
$-\log_{10}(p)$
0
1
2
3
4
1
2
3
4
5
6
7
8
9
10
12
14
15
16
18
20
22
24
27
30
32
Chromosome

Figure 3: Manhattan plot of results from mixed model analysis using ROADTRIPS. See Figure 2 for complete legend. Top hits are on ECA6 and X, but do not reach genome-wide significance. See Table 4 for specific SNPs and p-values.

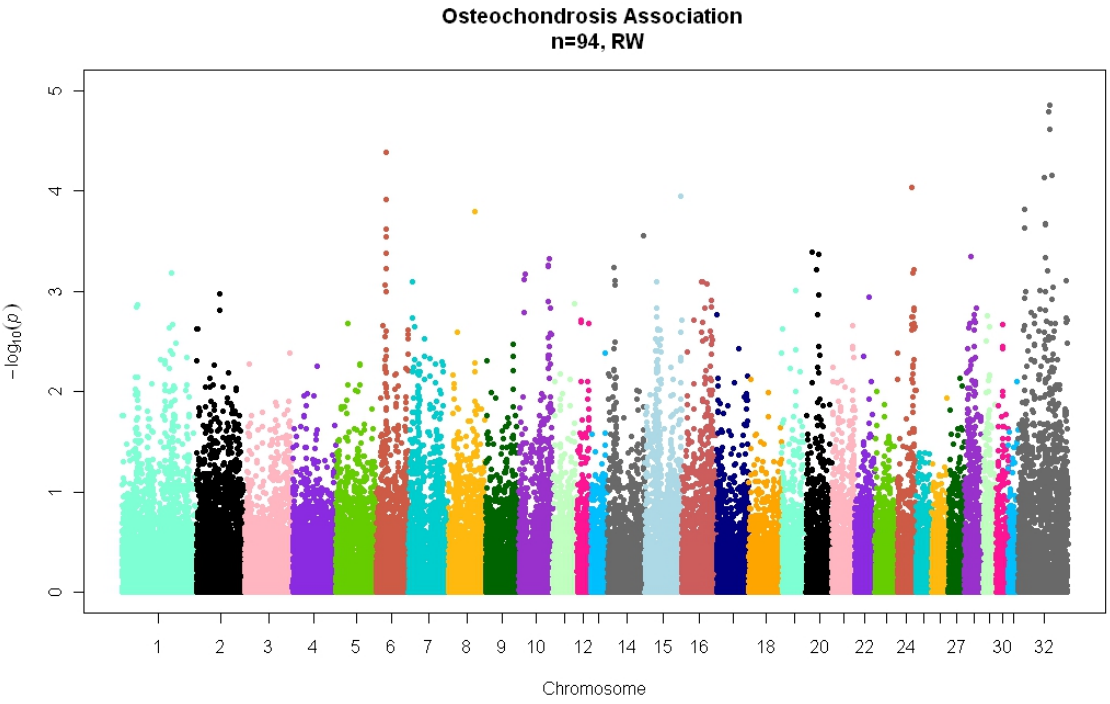
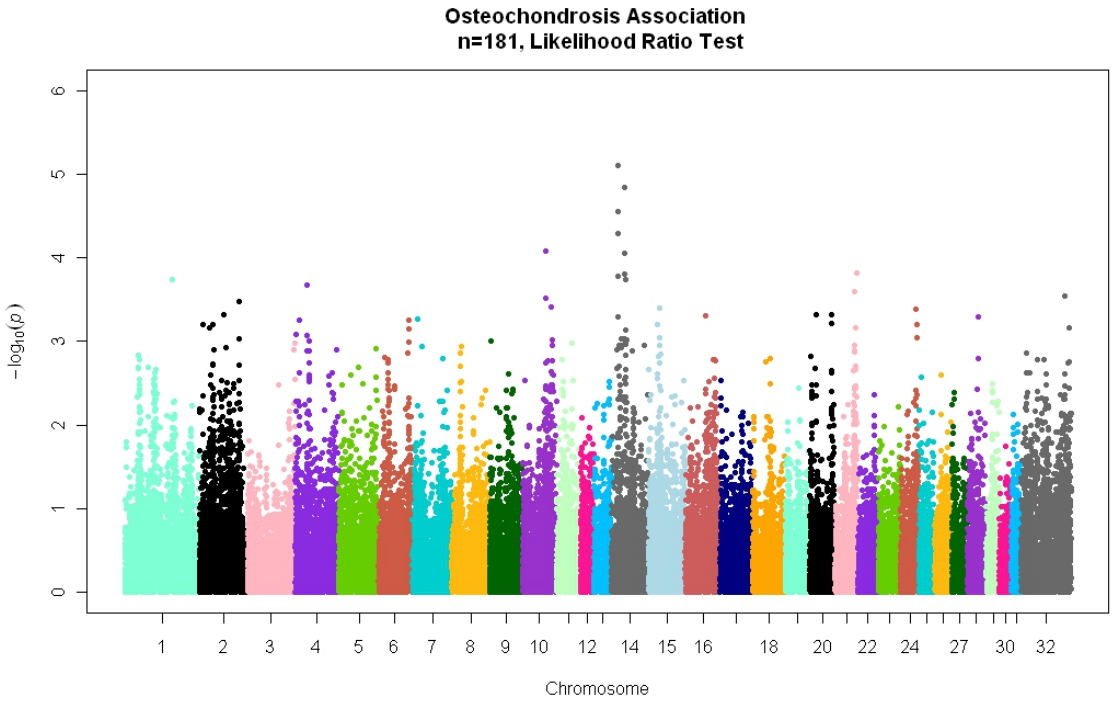


Figure 4: Manhattan plot of results from mixed model analysis using GEMMA. See Figure 2 for complete legend. Top hits are on ECA14, but do not reach genome-wide significance. See Table 5 for specific SNPs and p-values.



Chapter 5

Validation of Imputation Between Equine Genotyping Arrays

Validation of Imputation Between Equine Genotyping Arrays

Annette M. McCoy and Molly E. McCue

From the Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA

Acknowledgements: Thanks to Dr. James MacLeod for TB data and Robert Schaefer for custom shell script. AMM was supported by a NIH institutional training grant (T32OD10993).

Published as:

McCoy AM, McCue ME. Validation of imputation between equine genotyping arrays. (2014) *Anim Genet* 45:153. (Brief Note)

Background

Two genotyping arrays are available for the horse, containing ~54,000 and ~65,000 markers, of which only ~45,000 are shared. This leads to a loss of information when combining datasets generated on separate arrays. Genotype imputation offers a potential solution to this problem. Our objective was to assess the accuracy of genotype imputation for the two equine genotyping arrays across scenarios constructed to examine factors previously reported to affect imputation success in domestic animals and humans, including imputed population size, reference population size, reference population makeup (similar or different from the imputed population), and length of shared haplotype blocks (linkage disequilibrium; LD).[232;233]

Methods

Genotypes from 248 horses of three breeds (Quarter Horse [QH], *n* = 143; Standardbred [STB], *n* = 72; Thoroughbred [TB], *n* = 33) genotyped on the Illumina Equine SNP70 beadchip were "masked" down to the 45,703 markers shared by the SNP70 and SNP50 chips, and subsequently imputed back to the complete marker set for five chromosomes (ECA 1, 6, 15, 26, and X) using BEAGLE[211] with default settings (**Supplemental Methods, Figure S1**). Additionally, thirty QH genotyped on the SNP50 had their genotypes masked and imputed, using a reference population of 280 horses from thirteen diverse breeds.

Results/Conclusions

Results for twenty SNP70 scenarios are summarized in **Table S1**. Overall mean imputation success was 94.8% (individual horse range 82.2-100%). Generally, ECA1, 15, and 26 performed better than ECA6 and X. For ECA6, this may be partly due to the fact that a large block of imputed markers are located at the end of the chromosome and thus do not have an ideal haplotype context for imputation. Contrary to previous reports[233], size of the imputed population did not impact imputation success. Imputation success increased with larger reference population sizes (**Figure S2**) and when imputed and reference populations were breed-matched. However, large mixed breed reference populations resulted in more accurate imputation than small breed-matched reference populations. Breeds with longer LD had higher imputation success than those with shorter LD (TB > STB > QH) (**Figure S2**). These results reflect findings reported in humans.[232;234] Allelic R^2, the estimated squared correlation between the imputed allele dosage and the true allele dosage for a marker, was used as a measure of confidence for imputed genotype calls. The overall mean R^2 was 0.771 (range 0.582-0.981). Imputation success and R^2 were highly linearly correlated (r^2 = 0.79). Results for the SNP50 were comparable to the SNP70 (**Supplemental Results**). The total number of markers available for analysis after imputation was 73,200, an increase of ~27,500 markers from the set shared by the two chips. In conclusion, imputation between the two arrays was highly accurate.

Supplemental Material

Background

Two genotyping arrays have been designed for the horse. The first (Illumina Equine SNP50) was designed with 60,000 markers, of which 54,602 successfully genotyped in the commercial array. The second array (Illumina Equine SNP70) was designed to increase genome-wide marker density and to fill in coverage gaps identified in the SNP50 chip. Although all of the successful markers from the SNP50 were incorporated into the 74,000 markers designed for the SNP70 chip (65,157 successful), only 45,703 are actually genotyped with this platform. Thus, while 19,454 markers were added in the SNP70 array when compared to the SNP50, 8,899 markers remain unique to the older platform (**Figure S3**). To maximize the number of markers available for analysis when combining data obtained from the two arrays, imputation must therefore be carried out in "both directions" (i.e. from the SNP50 to the SNP70 and also from the SNP70 to the SNP50).

Methods

Data were retrieved from 248 horses of three breeds (Quarter Horse [QH], n = 143; Standardbred [STB], n = 72; Thoroughbred [TB], n = 33) genotyped on the Illumina Equine SNP70 beadchip (65,157 markers). These data were "masked" down to the list of 45,703 markers shared by the SNP70 and SNP50 chips, and subsequently imputed back to the complete marker set using BEAGLE.[211] The imputed genotypes were then compared to the known genotypes at each location to determine imputation accuracy.

Genotypes that were missing in the original data were excluded from analysis. On the basis of preliminary data (**Table S2**), five chromosomes were chosen for validation of imputation accuracy: equine chromosomes (ECA) 1, 6, 15, 26, and X. These were considered representative of all of the chromosomes as they reflected a range of both imputation success and chromosome size. Between 29% and 35% of the markers on each chromosome were imputed. Twenty scenarios were constructed, varying the imputed population size (range 5-30 individuals), imputed population breed (QH, STB, or TB), reference population size (range 20-100), and/or reference population make-up (breed-matched to the imputed population, or made up of an equal mix of all three breeds ["mixed" population]) (**Table S1**).

To confirm imputation accuracy for the 8,899 markers that are present on the SNP50 and not the SNP70 array, genotype masking and subsequent imputation was carried out for five chromosomes, as above, in thirty QH genotyped on the Illumina Equine SNP50 beadchip (54,602 markers). Between 11% and 37% of the markers on each chromosome were imputed (**Table S3**). Based on the public availability of genotyping data from a large number of horses of diverse breeds (www.animalgenome.org/repository/pub/UMN2012.1130/[59;235]) and the success of a large mixed breed population in the SNP70 scenarios (above), imputation accuracy in this QH population was confirmed in a scenario using a reference population comprised of 280 horses of thirteen diverse breeds (Thoroughbred, n = 44; Andalusian, n = 19; Arabian, n = 23; Belgian, n = 22; Franches-Montagnes, n = 20; French Trotter, n = 17; Hanoverian, n = 19; Icelandic, n = 17; Mongolian, n = 21; Norwegian Fjord, n = 21; Saddlebred, n = 21; Standardbred, n = 19; Swiss Warmblood, n = 17).

BEAGLE requires three input files for each chromosome to be imputed: a genotypes file (.bgl) for both the test population and the reference population, and a marker file, which includes the marker name, chromosomal position, and list of possible alleles at each locus. The marker file was generated by modifying the allele frequency output (--freq) from PLINK.[212] The genotypes files were converted from PLINK .map/.ped format to .bgl format using the phasing pipeline utility associated with GERMLINE.[236] BEAGLE was implemented using the default settings for unphased unrelated data.

To maximize the impact of imputation in a real dataset, horses genotyped on each platform (SNP50 or SNP70) should be alternately used as the reference and imputed populations such that each individual has actual genotypes from one platform and imputed genotypes from the non-overlapping markers from the other platform. To complete the data analysis pipeline for these real data, BEAGLE phased output files are converted back to PLINK .ped format using custom shell script (available at https://github.com/schae234/Beagle2Ped) with the phasing pipeline utility (above). Accompanying .map files must then be generated from the ordered list of markers in the phased BEAGLE output for the imputed population. Converted imputed files are subsequently merged with the original genotype data using PLINK (--merge). Merged imputed files can then be utilized for any number of analyses. The complete pipeline is illustrated in **Figure S1**.

SNP50 imputation results and comments

Results for imputation for the SNP50 chip are presented in **Table S3**. The average imputation success across all chromosomes was 94.2% (range for individual horses 87.4%-98.8%). When compared to results for the same size imputed population (n = 30) in the SNP70 scenarios, imputation success in this SNP50 scenario was somewhat lower for ECA1, higher for ECA6, and about the same for ECA 15, 26 and X. The mean R^2 across all chromosomes, reflecting confidence in the imputed genotype calls, was 0.725 (range 0.680-0.795). This is lower than was found in the SNP70 scenario with a large mixed breed population (mean 0.76). However, in that scenario, one-third of the horses in the reference population were of the same breed as the imputed population (QH), while in the SNP50 scenario, there were no QH in the reference population. This supports findings reported in the main text that a reference population that is breed-matched to the imputed population gives better results than a mixed reference population. Although the results cannot truly be directly compared because they looked at performance of imputation in different arrays, it is of note that nearly tripling the size of the reference population (from 100 to 280 individuals) did not result in a marked increase in imputation success. This reflects findings reported in human data, in which increasing reference population sizes over a threshold gave diminishing returns for improvement in imputation, except for very low frequency polymorphisms.[234;237]

Table S1: Summary of SNP70 validation scenario results. QH, Quarter Horse; STB, Standardbred; TB, Thoroughbred; SNP, single nucleotide polymorphism (marker); r^2, estimated squared correlation between the imputed allele dosage and the true allele dosage for a marker; ECA, *Equus caballus*.

Imputed pop breed	QH					QH				
# in imputed pop	5					10				
Reference pop breed	QH					QH				
# in reference pop	40					40				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4835	2412	2533	964	3342	4835	2412	2533	964	3342
# SNPs imputed	1468	767	733	365	1173	1468	767	733	365	1173
Mean imputation success	0.936	0.914	0.939	0.926	0.937	0.945	0.922	0.933	0.934	0.914
Minimum individual imputation success	0.909	0.896	0.903	0.907	0.886	0.921	0.893	0.912	0.912	0.866
Maximum individual imputation success	0.945	0.924	0.971	0.961	0.965	0.978	0.938	0.955	0.954	0.952
Mean r^2 for imputed SNPs	0.686	0.697	0.696	0.690	0.834	0.721	0.643	0.692	0.677	0.703
% SNPs r^2 < 0.5	0.17	0.15	0.17	0.17	0.07	0.15	0.22	0.19	0.20	0.16

Imputed pop breed	QH					QH				
# in imputed pop	20					30				
Reference pop breed	QH					QH				
# in reference pop	40					40				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4835	2412	2533	964	3342	4835	2412	2533	964	3342
# SNPs imputed	1468	767	733	365	1173	1468	767	733	365	1173
Mean imputation success	0.944	0.925	0.936	0.930	0.921	0.941	0.918	0.932	0.928	0.925
Minimum individual imputation success	0.914	0.872	0.911	0.89	0.872	0.884	0.87	0.888	0.88	0.883
Maximum individual imputation success	0.978	0.958	0.955	0.967	0.978	0.962	0.977	0.969	0.975	0.974
Mean r^2 for imputed SNPs	0.712	0.655	0.657	0.669	0.724	0.693	0.655	0.698	0.670	0.751
% SNPs r^2 < 0.5	0.16	0.21	0.23	0.21	0.15	0.18	0.24	0.17	0.22	0.12

Imputed pop breed	QH					QH				
# in imputed pop	10					20				
Reference pop breed	QH					QH				
# in reference pop	60					60				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4835	2412	2533	964	3342	4835	2412	2533	964	3342
# SNPs imputed	1468	767	733	365	1173	1468	767	733	365	1173
Mean imputation success	0.959	0.929	0.957	0.948	0.923	0.955	0.936	0.954	0.947	0.927
Minimum individual imputation success	0.942	0.897	0.936	0.925	0.882	0.925	0.898	0.923	0.897	0.876
Maximum individual imputation success	0.975	0.952	0.971	0.98	0.955	0.981	0.969	0.979	0.991	0.97
Mean r^2 for imputed SNPs	0.768	0.7	0.765	0.737	0.764	0.758	0.717	0.741	0.738	0.73
% SNPs r^2 < 0.5	0.10	0.18	0.12	0.13	0.11	0.11	0.15	0.13	0.15	0.15

Test (imputed) pop breed	QH					QH				
# in test pop	30					10				
Reference pop breed	QH					QH				
# in reference pop	60					100				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4835	2412	2533	964	3342	4835	2412	2533	964	3342
# SNPs imputed	1468	767	733	365	1173	1468	767	733	365	1173
Mean imputation success	0.956	0.931	0.952	0.941	0.935	0.969	0.951	0.967	0.966	0.936
Minimum individual imputation success	0.897	0.884	0.915	0.892	0.876	0.938	0.923	0.947	0.951	0.874
Maximum individual imputation success	0.983	0.983	0.988	0.994	0.993	0.988	0.972	0.98	0.99	0.977
Mean r^2 for imputed SNPs	0.76	0.711	0.739	0.718	0.772	0.818	0.764	0.837	0.819	0.779
% SNPs r^2 < 0.5	0.12	0.17	0.14	0.18	0.10	0.06	0.11	0.05	0.06	0.10

Test (imputed) pop breed	QH					QH				
# in test pop	20					30				
Reference pop breed	QH					QH				
# in reference pop	100					100				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4835	2412	2533	964	3342	4835	2412	2533	964	3342
# SNPs imputed	1468	767	733	365	1173	1468	767	733	365	1173
Mean imputation success	0.964	0.949	0.964	0.959	0.937	0.969	0.945	0.968	0.953	0.953
Minimum individual imputation success	0.937	0.917	0.945	0.918	0.88	0.897	0.877	0.917	0.895	0.903
Maximum individual imputation success	0.987	0.976	0.985	0.994	0.973	0.991	0.984	0.996	0.996	0.993
Mean r^2 for imputed SNPs	0.805	0.771	0.806	0.796	0.757	0.815	0.748	0.82	0.779	0.816
% SNPs r^2 < 0.5	0.07	0.11	0.07	0.08	0.12	0.06	0.13	0.06	0.07	0.07

Test (imputed) pop breed	QH					QH				
# in test pop	10					20				
Reference pop breed	Mixed					Mixed				
# in reference pop	100					100				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4839	2413	2538	964	3351	4839	2413	2538	964	3351
# SNPs imputed	1471	768	736	365	1178	1471	768	736	365	1178
Mean imputation success	0.956	0.928	0.953	0.953	0.923	0.955	0.937	0.95	0.953	0.925
Minimum individual imputation success	0.93	0.907	0.936	0.925	0.878	0.927	0.883	0.927	0.904	0.879
Maximum individual imputation success	0.978	0.943	0.968	0.983	0.965	0.985	0.969	0.972	0.992	0.969
Mean r^2 for imputed SNPs	0.8	0.726	0.785	0.799	0.734	0.784	0.735	0.763	0.779	0.735
% SNPs r^2 < 0.5	0.07	0.13	0.09	0.08	0.14	0.09	0.15	0.12	0.11	0.15

Test (imputed) pop breed	QH					QH				
# in test pop	30					10				
Reference pop breed	Mixed					QH				
# in reference pop	100					20				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4839	2413	2358	964	3351	4839	2413	2538	964	3351
# SNPs imputed	1471	768	736	365	1178	1468	767	733	365	1173
Mean imputation success	0.954	0.928	0.952	0.945	0.935	0.901	0.879	0.897	0.884	0.907
Minimum individual imputation success	0.89	0.87	0.903	0.885	0.888	0.881	0.861	0.87	0.849	0.851
Maximum individual imputation success	0.986	0.977	0.986	0.99	0.99	0.938	0.904	0.924	0.912	0.968
Mean r^2 for imputed SNPs	0.766	0.726	0.768	0.763	0.78	0.624	0.649	0.582	0.598	0.711
% SNPs r^2 < 0.5	0.11	0.15	0.10	0.13	0.10	0.19	0.17	0.23	0.21	0.14

Test (imputed) pop breed	STB					STB				
# in test pop	10					10				
Reference pop breed	STB					STB				
# in reference pop	40					60				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4839	2413	2538	964	3352	4839	2413	2538	964	3352
# SNPs imputed	1468	767	733	365	1173	1468	767	733	365	1174
Mean imputation success	0.983	0.969	0.984	0.971	0.979	0.988	0.978	0.991	0.98	0.984
Minimum individual imputation success	0.971	0.946	0.971	0.914	0.953	0.975	0.956	0.983	0.937	0.972
Maximum individual imputation success	0.996	0.989	0.996	0.994	0.994	1	0.991	0.998	0.999	0.992
Mean r^2 for imputed SNPs	0.981	0.815	0.908	0.836	0.899	0.923	0.869	0.939	0.905	0.918
% SNPs r^2 < 0.5	0.02	0.08	0.02	0.05	0.03	0.01	0.07	0.01	0.02	0.03

Test (imputed) pop breed	STB					STB				
# in test pop	10					10				
Reference pop breed	Mixed					STB				
# in reference pop	100					20				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4839	2413	2538	964	3352	4835	2412	2533	964	3342
# SNPs imputed	1472	768	738	365	1182	1468	767	733	365	1174
Mean imputation success	0.986	0.965	0.986	0.971	0.976	0.945	0.915	0.956	0.898	0.941
Minimum individual imputation success	0.974	0.946	0.979	0.934	0.956	0.939	0.887	0.92	0.822	0.908
Maximum individual imputation success	0.996	0.995	0.993	0.994	0.988	0.972	0.936	0.978	0.944	0.974
Mean r^2 for imputed SNPs	0.903	0.842	0.906	0.818	0.872	0.756	0.645	0.751	0.601	0.789
% SNPs r^2 < 0.5	0.02	0.08	0.01	0.05	0.05	0.11	0.19	0.13	0.22	0.09

Test (imputed) pop breed	TB					TB				
# in test pop	10					10				
Reference pop breed	Mixed					TB				
# in reference pop	90					20				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4839	2413	2538	964	3351	4839	2413	2538	964	3351
# SNPs imputed	1471	768	736	365	1177	1471	768	736	365	1177
Mean imputation success	0.987	0.97	0.989	0.983	0.968	0.971	0.944	0.972	0.967	0.948
Minimum individual imputation success	0.98	0.948	0.977	0.961	0.945	0.951	0.904	0.955	0.949	0.855
Maximum individual imputation success	0.995	0.991	0.996	0.994	0.987	0.99	0.976	0.984	0.986	0.991
Mean r^2 for imputed SNPs	0.912	0.824	0.914	0.906	0.909	0.84	0.802	0.832	0.837	0.926
% SNPs r^2 < 0.5	0.01	0.08	0.01	0.02	0.04	0.07	0.09	0.08	0.09	0.03

Table S2: Summary of preliminary imputation data. Imputation was performed on four Standardbred horses genotyped on both the SNP50 and SNP70 platforms. The reference population for the SNP50 → SNP70 scenario was 72 Standardbreds genotyped on the SNP70 chip. The reference population for the SNP70 → SNP50 scenario was 94 Standardbreds genotyped on the SNP50 chip. Imputation success was calculated as: (total # genotypes imputed - # genotype errors) / total # genotypes imputed. Chr, chromosome; SNP, single nucleotide polymorphism (marker); # SNP errors, number of SNPs with one or more incorrectly imputed genotypes; # genotype errors, total number of incorrectly imputed genotypes across all SNPs in all individuals. Highlighted chromosomes are those chosen for further follow-up, as described in the text.

SNP50 → SNP70					
Chr	# SNPs	# SNPs imputed	# SNP errors	# genotype errors	imputation success
1	4835	1456	80	103	0.982
2	3430	1042	57	73	0.982
3	3019	860	56	70	0.980
4	2938	808	80	97	0.970
5	2737	772	49	55	0.982
6	2412	755	52	68	0.977
7	2607	745	51	66	0.978
8	2654	780	37	54	0.983
9	2336	675	35	48	0.982
10	2306	684	53	62	0.977
11	1790	554	24	35	0.984
12	876	285	27	39	0.966
13	1136	362	24	40	0.972
14	2670	710	37	50	0.982
15	2533	720	22	32	0.989
16	2416	686	65	71	0.974
17	2184	511	38	43	0.979
18	2146	554	40	60	0.973
19	1716	471	26	39	0.979
20	1778	537	41	55	0.974
21	1678	462	34	44	0.976
22	1416	383	29	36	0.977
23	1498	408	20	22	0.987
24	1401	404	24	28	0.983
25	1070	318	21	29	0.977
26	964	362	11	12	0.992
27	1066	325	18	24	0.982
28	1245	336	31	31	0.977
29	803	180	21	28	0.961
30	813	222	10	12	0.986
31	638	185	23	26	0.965
X	3342	1171	128	205	0.956

SNP70 → SNP50					
Chr	# SNPs	# SNPs imputed	# SNP errors	# genotype errors	imputation success
1	4357	978	69	87	0.978
2	2801	413	47	62	0.962
3	2809	650	36	53	0.980
4	2549	419	28	35	0.979
5	2261	296	27	33	0.972
6	1850	193	35	51	0.934
7	2229	367	22	33	0.978
8	2159	285	30	45	0.961
9	1958	297	36	54	0.955
10	1910	288	30	53	0.954
11	1451	215	17	20	0.977
12	707	116	18	31	0.933
13	955	181	13	21	0.971
14	2222	262	32	44	0.958
15	2150	337	21	26	0.981
16	2078	348	31	34	0.976
17	1869	196	26	30	0.962
18	1908	316	23	37	0.971
19	1414	169	15	20	0.970
20	1475	234	32	46	0.951
21	1349	133	18	20	0.962
22	1177	144	20	29	0.950
23	1241	151	13	18	0.970
24	1094	97	15	22	0.943
25	924	172	11	20	0.971
26	957	355	11	15	0.989
27	871	130	11	15	0.971
28	1090	181	5	9	0.988
29	769	146	8	13	0.978
30	713	122	7	9	0.982
31	595	142	12	14	0.975
X	2530	359	68	119	0.917

Table S3: Summary of SNP50 validation scenario results. QH, Quarter Horse; SNP, single nucleotide polymorphism (marker); r^2, estimated squared correlation between the imputed allele dosage and the true allele dosage for a marker; ECA, *Equus caballus*.

Test (imputed) pop breed	QH				
# in test pop	30				
Reference pop breed	Mixed				
# in reference pop	280				
Chromosome	**ECA1**	**ECA6**	**ECA15**	**ECA26**	**ECAX**
# SNPs in reference panel	4373	1858	2154	958	2541
# SNPs imputed	997	209	352	357	344
Mean imputation success	0.937	0.964	0.94	0.932	0.936
Minimum individual imputation success	0.884	0.924	0.908	0.874	0.895
Maximum individual imputation success	0.978	0.988	0.968	0.979	0.965
Mean r^2 for imputed SNPs	0.682	0.795	0.723	0.680	0.746
% SNPs r^2 < 0.5	0.25	0.08	0.17	0.23	0.14

Figure S1: Complete pipeline for imputation of equine genotyping data. The progression of file types is represented in ovals on the left side, while the program or utility associated with each step is shown in the boxes on the right.

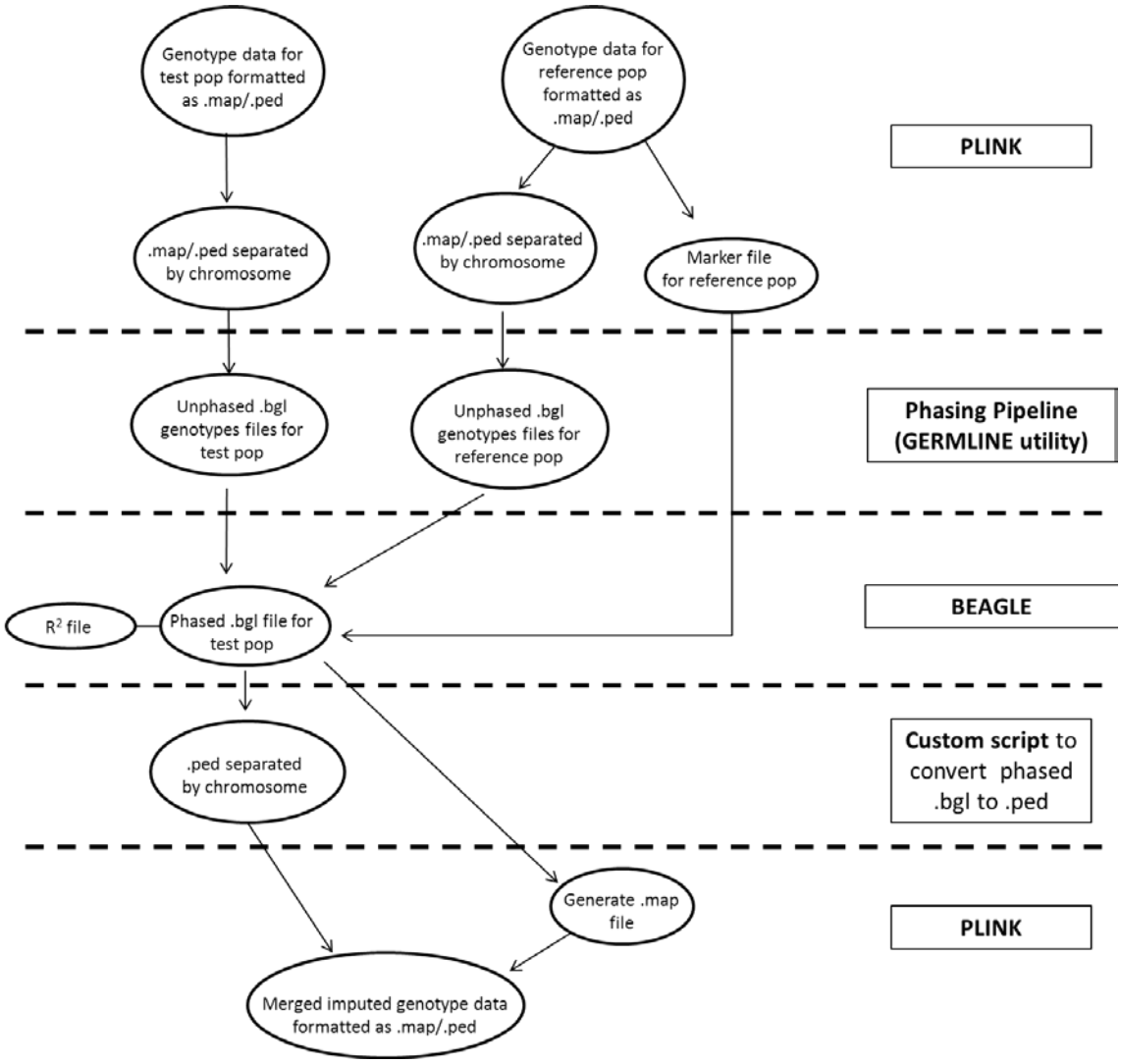


Figure S2: Mean imputation success with an imputed population *n* = 10 across a range of reference population sizes (*n* = 20-100) for each of three breeds (Quarter Horse [QH], red squares; Standardbred [STB], blue circles; Thoroughbred [TB], green triangles). It appears as though there may be diminishing returns for reference population sizes greater than 100 individuals.

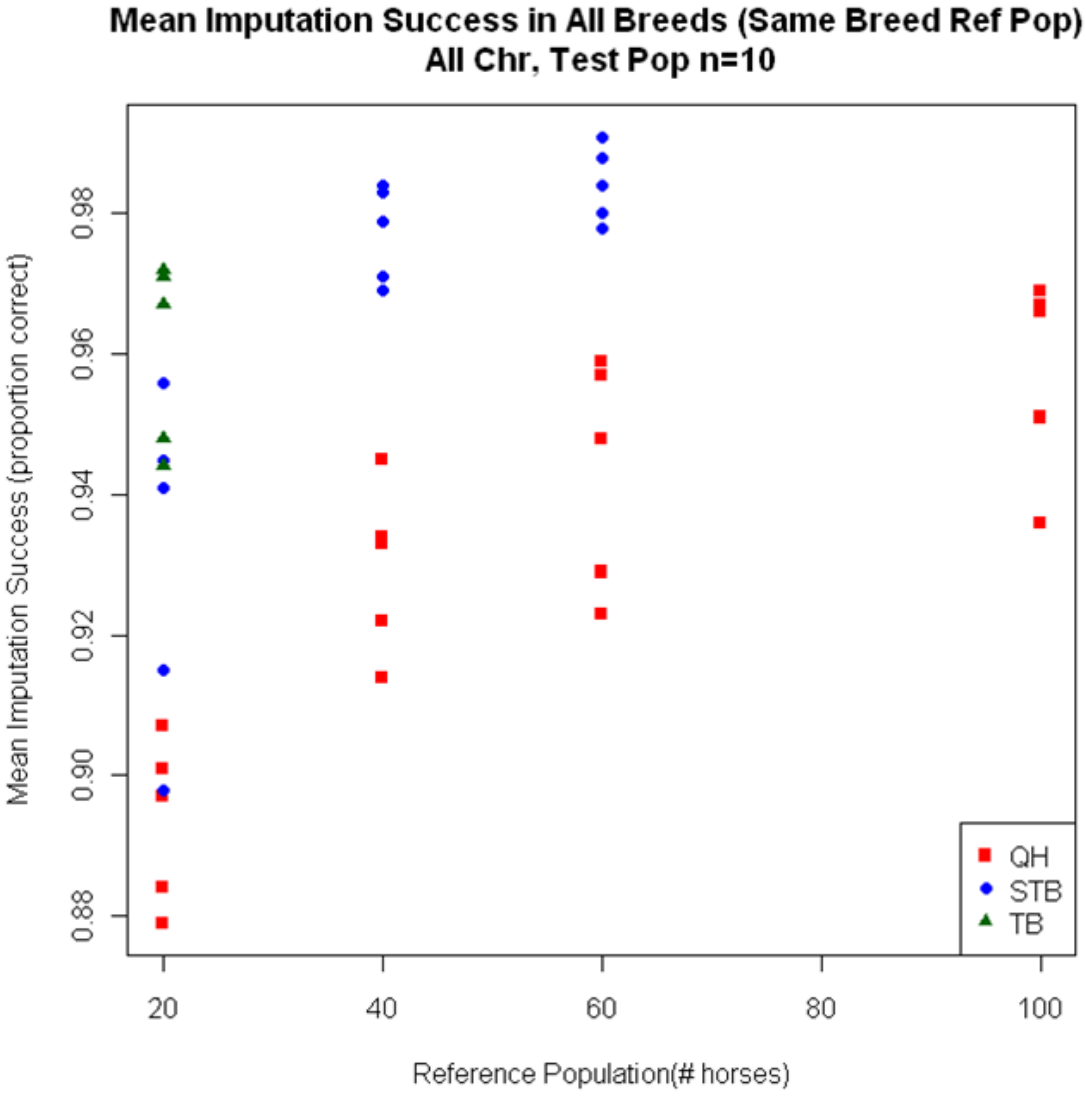
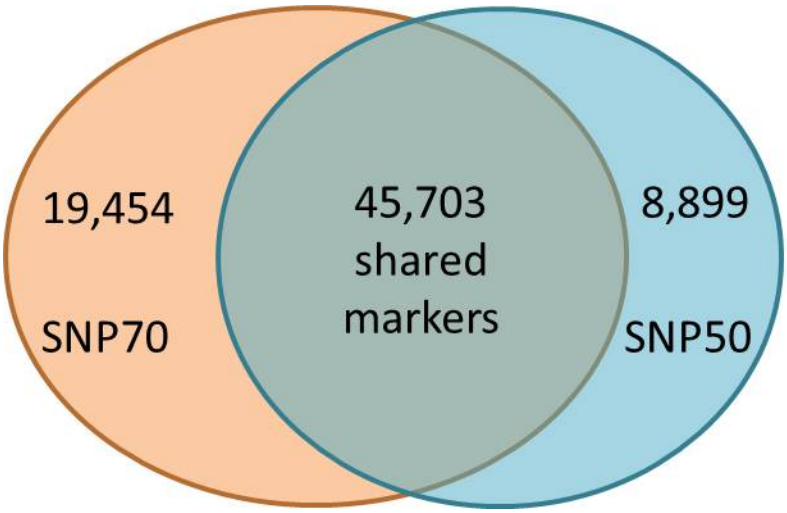


Figure S3: Venn diagram of marker overlap between the Illumina Equine SNP50 and SNP70 beadchips.



Chapter 6

Investigation of Putative Risk Alleles for Osteochondrosis in the Horse

Investigation of Putative Risk Alleles for Osteochondrosis in the Horse

Annette M. McCoy

From the Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA

Acknowledgements: A. Hauser performed genotyping of selected ECA6 variants under the guidance of the author as part of an independent study project.

Sources of Funding: Funding provided by the University of Minnesota Equine Center/Minnesota Racing Commission and the United States Equestrian Federation, Inc. Dr. McCoy was funded by an institutional NIH T32 Comparative Medicine and Pathology Training Grant (University of Minnesota) and a Doctoral Dissertation Fellowship (University of Minnesota); partial funding for Dr. McCue was provided by NIH NIAMS 1K08AR055713-01A2.

Summary

Osteochondrosis (OC), simply defined as a failure of endochondral ossification, is a complex disease with both genetic and environmental risk factors that is commonly diagnosed in young horses, as well as other domestic animal species and humans. Although up to 50% of the risk for developing OC is reportedly inherited, specific genes and alleles underlying risk are thus far completely unknown. Only a single candidate gene has been investigated, and its physiological relevance to OC is questionable.

In part, the lack of candidate gene investigation may be due to difficulties inherent to the candidate gene approach to variant discovery and validation. The increasing affordability of whole-genome sequencing offers an alternative approach to variant discovery that is both efficient and cost-effective. In this study, whole-genome sequencing was completed in 18 horses to an average depth of either 6x (n = 12) or 12x (n = 6) for the purpose of variant discovery. These horses were selected from a larger cohort in whom a genome-wide association study (GWAS) for OC had been performed. Variants within the chromosomal regions of association from this GWAS, as well as from 9 additional regions based on previously published GWAS for tarsal OC, were prioritized based on predicted functional effect and segregation with OC status. Two hundred forty variants were selected for follow-up genotyping in 180 individuals (the GWAS study cohort) using a Sequenom high-throughput genotyping assay. After correction for relatedness, three SNPs (one each on ECA10, 14, and 21) were highly associated with OC status. These SNPs were located within genes whose known physiologic function makes them feasible candidate genes for OC risk (*ARHGAP26*, *PREP*, and *SEMA5A*). These putative risk alleles should be validated in an independent population.

Introduction

Until recently, identification of variants underlying diseases or traits of interest was primarily accomplished by Sanger sequencing of candidate genes located within chromosomal regions of association from linkage mapping or genome-wide association studies (GWAS). This approach has been successful in a number of cases, notably, for diseases caused by mutations in single genes such as type I polysaccharide storage myopathy (PSSM1), and for traits of interest such as coat color.[206;238-241] However, candidate gene sequencing also has weaknesses. In addition to being time-consuming and expensive, especially for large genes, it is heavily reliant on the quality of the reference genome. Regions for which there is missing sequence in the reference are very difficult to sequence because of obstacles to primer design, and incomplete annotations may result in variants of importance being overlooked entirely.

As the cost of next-generation sequencing has decreased, an alternative for variant discovery has emerged, namely, whole-genome sequencing. There is still a trade-off between coverage depth and number of individuals sequenced, and the general consensus seems to be that for the purposes of variant discovery, sequencing a larger number of individuals at a shallow depth is preferable to sequencing a smaller number of individuals at greater depth.[242;243] This is the approach taken by a number of large human genome sequencing consortia, including the 1000 Genomes Project.[244] The first report of variant discovery using whole-genome sequencing in a single horse was published in 2012[245], and since then, there has been only a single report of using whole-genome sequencing to investigate variants in candidate genes for a disease.[246] In the latter report, this approach resulted in discovery of a nonsense mutation causing Incontinentia Pigmenti in mares,

supporting the feasibility of the approach, at least for Mendelian traits. The utility of whole-genome sequencing for variant discovery in complex traits has also been demonstrated, however. A nonsense mutation in *DMRT3* (doublesex and mab3-related transcription factor 3) thought to be permissive for "gaitedness" in the horse was discovered through deep sequencing of two individuals exhibiting opposite haplotypes within a chromosomal region of interest identified on GWAS.[11]

Osteochondrosis (OC) is a common manifestation of developmental orthopedic disease recognized in young horses across breeds, as well as many other species. Histologic studies in animals suggest that the failure of endochondral ossification that is the hallmark of OC is most likely due to vascular abnormalities at certain predilection sites at the ends of long bones.[94;163] Surgical transection of vessels within the epiphyseal-articular cartilage complex in young foals can certainly lead to OC-like lesions[130]; however, specific underlying risk factors for naturally-occurring disease are still incompletely understood.

Environmental factors, particularly diet and exercise, have been widely reported to play a role in the development of OC.[132;133;139-142;202] However, reduction in disease prevalence via management changes alone has been reported to be limited[144], highlighting the importance of genetic risk factors in the manifestation of disease. Heritability estimates in a variety of breeds range from 0.15 to 0.52[32;37;42;84;86;204], suggesting that up to 50% of disease risk may be inherited. A number of GWAS have been performed to identify chromosomal regions associated with OC. However, findings have not been consistent across studies, and putative quantitative trait loci (QTLs) on 13 different chromosomes have been suggested (see **Table 1, Chapter 4** for summary of

published GWAS). Further, although many potential candidate genes within these loci have been identified in the GWAS reports, investigation of variants within only a single gene has been published.[156]

The purpose of this study was to identify putative functional variants underlying genetic risk for OC in the horse, using Standardbreds with tarsal OC as a model population. The high prevalence and high heritability of OC within this breed suggest that risk factors of moderate to major effect should be common among affected individuals. We performed genome-wide association analyses in a cohort of individuals born and raised on a single breeding farm in the eastern United States, which was selected specifically to reduce the effect of environmental confounders on disease association. The initial cohort consisted of 94 horses born in 2007 and additional horses were added over subsequent foaling seasons. The final GWAS in 182 individuals identified two distinct loci on ECA14 that were moderately associated with OC status (see **Chapter 4**). Whole-genome sequencing was performed for a subset of the GWAS population for the purposes of variant discovery. Subsequently, a high-throughput genotyping assay was utilized to investigate variants in the larger population.

Materials and Methods

Horses: As described in **Chapter 4**, the initial study cohort was comprised of 94 Standardbred yearlings born in 2007. Over subsequent foaling seasons, 88 horses were added to the group (born in 2009 [n=16], 2010 [n=52], and 2012 [n=20]) for a final cohort of 182 individuals. Seventy horses were affected with OC in one or both tarsi, while 112 were identified as controls. Thirty-four of the controls, all from the original

cohort, were presumed unaffected because of lack of clinical signs including effusion and lameness, while the remainder (n=78) were radiographically examined and found to be free of disease (see **Table 2, Chapter 4** for summary of entire cohort).

Selection of Horses for Whole-Genome Sequencing by Haplotype Analysis: After difficulties were encountered during candidate gene sequencing (see **Appendix 1** for details), the decision was made to pursue whole-genome sequencing in a subset of the study cohort. By this time, the study cohort had increased in size to 162 individuals, and a GWAS in this new group suggested chromosomal regions of association with OC on ECA2 and ECA14 in addition to the region on ECA6 previously described. As a result, haplotypes were evaluated in these three regions to select individuals for whole-genome sequencing. Single nucleotide polymorphisms (SNPs) within the regions were pruned for minor allele frequency <1% (--maf 0.01) and genotyping success of >90% (--geno 0.9) using PLINK.[212] Genotype data was computationally phased and missing genotypes were imputed using fastPHASE 1.2.[247] Haplotype blocks were evaluated in Haploview[248] and by manual sorting of genotypes within Excel (Microsoft Corporation, Redmond, WA). Haplotypes were evaluated for both their absolute frequency within the OC-affected and OC-unaffected groups and for differences in frequency between groups. For each region, the most common haplotype within an affectation status that also exhibited a large difference between OC-affected and OC-unaffected groups was selected as the haplotype of interest. Individuals that exhibited these haplotypes in one or more of the regions of interest were eligible for selection for whole-genome sequencing. Horses were preferentially selected if they had the haplotype of interest in more than one region of

interest; however, consideration was also given to balancing the selected cohort by gender, gait (pace or trot), and sire.

Whole-Genome Sequencing: Genomic DNA (2-6μg) was submitted to the University of Minnesota Biomedical Genomics Center (UMGC) for quality control, library preparation, and sequencing. Five samples failed initial quality control; DNA was re-isolated from hair root samples (using the standard protocol as described in **Chapter 4**) prior to resubmission. Samples were subjected to standard library preparation including fragmentation, polishing, and adaptor ligation, and were prepared with an indexed barcode for a paired-end run on the Illumina HiSeq sequencer. Depth of coverage was determined by the calculation C = LN/G, where C is the coverage, L is the read length, N is the number of reads, and G is the genome length. For 100bp paired-end reads, given an approximate genome length of 3 billion base pairs, 160 million reads results in ~12x coverage of the genome, and 80 million reads results in ~6x coverage of the genome. Of the nine affected horses, 3 were sequenced at 12x coverage and 6 at 6x coverage; the same distribution was used for the nine unaffected horses (**Table 1**). Samples were split between all eight lanes on each of two flowcells (10 samples on one flowcell, 8 on the other). Raw sequence data was deposited within designated storage at the Minnesota Supercomputing Institute (MSI).

Data analysis, including quality control, alignment, and variant detection, was carried out following published best practices[249] within the Galaxy framework hosted by MSI. Sequence quality was assessed using the FastQC tool. Subsequently, bases with a quality score of ≤28.0 were trimmed from the 3' end of the sequence (FASTQ quality trimmer) and paired reads were re-synchronized (resynch). Quality of the trimmed and

synchronized reads was re-assessed. Reads were then mapped to the reference sequence (EquCab 2.0, Sept. 2007) using BWA for Illumina. Ambiguously mapped and low quality reads were removed (filter SAM), after which reads were sorted and mate-pair information updated (paired-read mate fixer). PCR duplicates were also removed (mark duplicate reads) and reads were realigned around indels (realigner target creator, indel realigner), followed by re-assessment for duplicate reads. Base quality recalibration was performed to remove systematic bias (count covariates, analyze covariates, table recalibration). This process was completed for the reads from each of the eight lanes for every individual before merging the mapped and recalibrated "lane-level" BAM files into a single "sample-level" file. Removal of duplicates and realignment around indels was repeated on the merged file. The eighteen sample-level files were merged into three groups of six, evenly divided between affected and unaffected individuals, for the purposes of variant calling using the UnifiedGenotyper utility of the Broad Institute's Genome Analysis ToolKit (GATK) with a threshold phred-scale score of 20.0. Variants were filtered using the following thresholds: Quality Depth (QD) < 2.0 (assesses variant quality score taking into account depth of coverage at that variant), Read Position Rank Sum < -20.0 (Mann-Whitney Rank Sum test on the distance of the variant from the end of each read covering it), Fisher Strand (FS) > 200.0 (phred-scaled p-value to detect strand bias). Filtered variant lists from the three groups were combined into a single variant calling file (VCF) for subsequent analysis. Predicted functional effect for each called variant was determined based on the current equine reference genome annotation using the SnpEff tool in GATK. Frequency of variants within cases and controls, and the significance of frequency differences, was calculated using the SnpSift CaseControl tool

in GATK. Variants from particular chromosomal regions of interest were selected using SnpSift Intervals and converted into Excel format for further evaluation.

Variant Prioritization and Follow-up with RFLP or Sequencing: Based on the GWAS performed in 162 horses (i.e. as described under *Selection of Horses for Whole-Genome Sequencing by Haplotype Analysis*), variants were evaluated within nine broad chromosomal regions for 1) predicted functional effect; and 2) segregation with disease status. Variants with a coverage depth of at least 50 that passed all quality filters were considered for follow-up if they had a predicted functional effect within a protein-coding gene and if they were present in at least 5 affected or unaffected individuals, but fewer individuals of the opposite disease status (preferably with a genotypic model p-value of $<5 \times 10^{-2}$ as calculated by SnpSift CaseControl). Primers were designed to amplify regions surrounding 16 SNPs on four chromosomes using Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi). Restriction enzymes were successfully designed for nine of these SNPs using NEBcutter v2.0 (New England BioLabs, Inc., http://tools.neb.com/NEBcutter2/); the remainder of the SNPs required genotyping by sequencing.

The master mix for each PCR reaction included 1.5µl 10x PCR buffer (QIAGEN, Valencia, CA), 1.0µl each of [20µM] forward and reverse primer, 1.5µl [300µM] dNTPs, 0.15µl HotStarTaq® DNA polymerase (QIAGEN, Valencia, CA), 8.25µl water, and 3.0µl DNA (2.5ng/µl). PCR reaction conditions were as described in **Appendix 1**, with the following exceptions: the primers for variants in *TRAF3IP1* and *ESPNL* used an annealing temperature of 57°C; primers for variants in *GRIA2*, *FSTL5*, and ENSECAG00000003042 used an annealing temperature of 58°C and an extension time

of 45sec; the primers for the variant in *SCLY* required a touchdown procedure (20min at 95°C followed by 24 cycles of 30sec at 94°C, 30sec at 67°C, 30sec at 72°C with the annealing temperature dropping by 0.5°C each cycle, then 11 cycles at an annealing temperature of 55°C, and finally 15min at 72°C). PCR products were visualized on 2% agarose gels with ethidium bromide prior to submission for sequencing. To prepare the PCR products for Sanger sequencing, 1µl USB® ExoSAP-IT® (Affymetrix, Santa Clara, CA) was added to 4µl PCR product and incubated in the thermocycler for 15 min at 37°C for PCR product cleanup followed by 15 min at 80°C for enzyme inactivation. Subsequently, 1µl [20µM] primer and 6µl water were added and the sample (12µl total volume) submitted to the University of Minnesota BioMedical Genomics Center (UMGC) for sequencing. Sequence was analyzed in Sequencher and assembled against the equine reference sequence. Enzyme digests were performed according to manufacturer recommendation for each enzyme.

Sequenom Assay: A custom Sequenom assay was designed for high-throughput genotyping within the study cohort. Variants were selected from within the top regions of interest in the GWAS performed in 162 horses. Additional variants were selected from chromosomal regions previously reported to be associated with hock OC (see **Chapter 4**, **Table 1** for complete summary of previously published GWAS). Regions of interest included: ECA1 117-119Mb, ECA2 98-100Mb, ECA3 88-114Mb, ECA4 56-60Mb, ECA5 76-92Mb, ECA10 55-60Mb and 80-81Mb, ECA14 15-19Mb and 34-37Mb, ECA16 6-24Mb and 33-43Mb, ECA18 35-47Mb and 74-82Mb, ECA21 5-17Mb and 43-54Mb. Variants discovered through whole-genome sequencing were filtered to include

only SNPs (no indels) that passed all quality control filters, and were subsequently prioritized according to the following parameters:

1) present in 3+ more cases than controls, or vice versa;
2) not intergenic;
3) non-synonymous, then synonymous changes;
4) if intronic, close to intron-exon boundary (preferably <100bp);
5) coding genes preferred over non-coding; and
6) if upstream/downstream, as close as possible to start/stop codon.

An attempt was made to include at least one variant per coding gene within each region of interest; if multiple variants of equally low predicted function were the only ones available within a gene, then the one with the higher genomic p-value was selected. In addition to the experimental SNPs, 98 ancestry informative markers (AIMs) were included in the Sequenom assay to help control for population structure (**Table 5**).

Genotyping results were initially analyzed using an uncorrected association analysis in PLINK[212] (--assoc), pruning for MAF <1% (--maf 0.01) and missingness <95% (--geno 0.95). Subsequently, they were analyzed using GEMMA (Genome-wide Efficient Mixed Model Analysis) software[214] to account for population structure and relatedness. The association test in GEMMA was performed using the options to create a centered relatedness matrix (-gk2) and perform all three possible frequentist tests: Wald, likelihood ratio, and score (-fa 4). The relatedness matrix was constructed using the AIMs. SNPs were pruned prior to analysis using the default GEMMA parameters of MAF <1% and missingness <95%.

Results

Selection of Horses for Whole-Genome Sequencing by Haplotype Analysis: The region on ECA2 spanned from ~70-79Mb and included 169 SNPs after pruning. The region on ECA6 spanned from ~22-28Mb and included 123 SNPs. The region on ECA14 spanned from ~15-19Mb and included 84 SNPs. There were no haplotypes that were very common in one affectation group and absent in the other. On average, the selected haplotypes of interest were present in twice as many affected as unaffected individuals, or vice versa, but were present in only 15-30% of individuals overall. Eighteen horses, 9 affected with OC and 9 unaffected, were selected for whole-genome sequencing; a summary of these individuals is in **Table 1**.

Whole-Genome Sequencing: Actual coverage for the twelve individuals sequenced for a target of 6x ranged from 4.7x to 7.9x (mean 6.4x). Actual coverage for the six individuals sequenced for a target of 12x ranged from 10x to 13.1x (mean 12.2x). Summary metrics for sequence alignment are reported in **Table 2**.

After filtering, 14,588,812 variants were called, at an average of 1 variant every 162 base pairs. Of these, 13,157,608 were SNPs, 671,144 were insertions, and 760,060 were deletions. The vast majority of variants, over 14 million (99.1%), were not predicted to have any functional effect. Of the 152,700 variants predicted to have some functional effect, 85,916 were of low effect (mostly synonymous SNPs), 57,122 were of moderate effect, and 9,662 were of high effect. A summary of predicted effects by type and region is reported in **Table 3**. Graphical summaries of distribution of variant types, variant depth of coverage, indel lengths, and allele frequency spectrum are shown in **Figure 1**.

Variant Prioritization and Follow-up with RFLP or Sequencing: Variants considered for follow-up are summarized in **Table 4**. Although primers were designed and optimized for 16 variants (highlighted in the table), only four variants were ultimately followed up using RFLP (*SCLY*) or Sanger sequencing (*PER2, ESPNL, TRAF3IP1*) because of the decision to pursue high-throughput genotyping as an alternative approach (see *Sequenom Assay*). After genotyping 10 cases and 10 controls, only the variant in *ESPNL* continued to show any segregation with disease status. After genotyping in an additional 10 cases and 10 controls, and including data from the horses that underwent whole-genome sequencing, individuals with a T allele were 2.97 times more likely to be cases than those with the C allele (95% CI 1.38-6.39). The variant was additionally sequenced in 10 randomly selected Quarter Horses (7 C/C; 3 C/T) and 10 randomly selected Thoroughbreds (5 C/C; 5 C/T). None of these individuals had a T/T genotype at this locus, although T is the reference allele (from a Thoroughbred).

Sequenom Assay: 240 SNPs on 10 chromosomes were included in the final Sequenom assay design (**Table 6**). These SNPs were selected from regions of interest identified in the GWAS performed in 162 horses as well as from chromosomal regions previously reported to be associated with hock OC in other populations (see Materials and Methods). These SNPs were multiplexed in groups of 48 for genotyping in 180 individuals; two horses from the final study cohort reported in **Chapter 4** did not have sufficient remaining DNA for genotyping. 218 SNPs were available for analysis in PLINK; 168 SNPs passed filtering in GEMMA and were available for analysis by this method. The top results from PLINK and GEMMA analyses are shown in **Table 7**. The most significantly associated SNP in both analyses was located in the first intron of

ARHGAP26 (Rho GTPase activating protein 26) on ECA14 (chr14.34391965). The alternate allele for this SNP was found in 20% of cases and 8% of controls. Other top hits in both analyses were located just downstream of the *PREP* (prolyl endopeptidase) stop codon on ECA10 (chr10.55605051; 17% of cases, 7% of controls) and in intron 13 of *SEMA5A* (sema domain, seven thrombospondin repeats [type 1 and type 1-like], transmembrane domain [TM] and short cytoplasmic domain, [semaphorin] 5A) on ECA21 (chr21.50348105; 35% of cases, 21% of controls).

Discussion

OC is a complex disease, and as such it is unlikely that a variant in a single gene underlies all risk for disease development. Instead, it is likely that a combination of genetic risk factors, in addition to environmental context, determines the extent of disease expression in an individual. The variety of GWAS results that have been previously reported may reflect differences in disease definition, environmental risk factors, and/or computational approaches between studies, but it is also possible that they reflect real differences in risk alleles between populations. The challenge lies in sorting out the risk alleles of moderate to major effect that, given the shared pathophysiology of disease, are likely shared across populations, from the modifying alleles that may be breed- or predilection site-specific. In this study, we have capitalized on a population of Standardbreds with a shared environment to reduce this potential source of confounding, and have selected analytical techniques that account for population structure and relatedness among individuals. Investigation of risk alleles based on a GWAS in this population thus may be more fruitful than in a more heterogeneous group.

Although many potential candidate genes within reported QTLs for OC have been suggested, only a single candidate gene has been investigated. This gene, *XIRP2* (Xin actin-binding repeat containing 2) is located within a putative QTL on equine (ECA) chromosome 18 reported in South German Coldbloods.[208] Two SNPs in intron 2 of *XIRP2* were found to be significantly associated with OC, with relative risks of disease reported to be 1.3-2.4 higher in individuals homozygous or heterozygous for the reference allele at these markers.[156] However, there was no physiologic justification for the role of this gene, which is primarily expressed in cardiomyocytes and at the myotendinous junction of skeletal muscle cells, in OC. Further, this gene has not been reported to be expressed in cartilage.

The lack of intense investigation into specific candidate genes for OC may, in part, be a reflection of the challenges inherent in the traditional candidate gene approach to variant discovery. The process can be time-consuming and difficult, and is especially hampered by incomplete annotation of the reference genome. We initially made an effort to sequence a gene considered to be a strong biologic candidate for disease risk, histone deacetylase 4 (*HDAC4*), but discovered that multiple approaches, including the use of genomic DNA, cDNA isolated from cartilage and subchondral bone, and RNAseq data, were insufficient to bridge a 20kb gap in the middle of the gene and sequence missing exons. Attempts to identify the first exon and 5'UTR were also minimally successful (see **Appendix 1** for a summary of these efforts). Although alignment of next-generation sequencing is also primarily reliant on the reference genome, this approach allows for *de novo* assembly of unaligned reads into longer contiguous reads that can bridge gaps. Further, in this study, the use of next-generation sequencing allowed variant discovery to

be carried out in 18 horses, rather than just two or three. The inevitable outcome of this is a more complete picture of what variants are present in the population, as well as a better estimate of how those variants segregate with disease status, which helps with prioritization for follow-up in the larger group.

Next-generation sequencing also allows for widespread variant discovery in non-exonic sequence, something that is not particularly feasible (and is generally cost-prohibitive) using Sanger sequencing, especially in large genes. As our understanding of the importance of non-coding/regulatory regions of the genome improves, it is likely that variants found outside of exons will be increasingly recognized for their roles in disease. Indeed, many SNPs found to be associated with complex diseases in large human GWAS have been recognized to overlap with regulatory regions annotated as part of the ENCODE (Encyclopedia of DNA Elements) project, suggesting their plausibility as functional alleles underlying disease risk.[250] As yet, an equivalent of the ENCODE project is not available for agricultural species, including the horse, but it is logical to assume that regulatory elements could be of similar importance for disease development, and this will likely become an important focus of future studies.

For this project, investigation of variants was confined to specific regions of the genome corresponding to GWAS findings in our study cohort and selected additional regions published by others as putative QTLs for hock OC. After an initial attempt at variant follow-up using Sanger sequencing and RFLP, it was decided that a high-throughput genotyping assay (Sequenom) was a more efficient and cost-effective approach to variant investigation in this group. Using this approach, 240 variants were investigated in a population of 180 horses. Since this was the population in which our

GWAS was conducted, it is perhaps not surprising that the most highly associated SNPs from the Sequenom assay were found within the chromosomal regions of interest from that analysis. However, it is of note that the genes associated with the top three SNPs are each viable candidates for playing a role in OC risk. GTPase activating proteins, such as *ARHCAP26* are crucial mediators of the activity of Rho GTPases, which play an important role in chondrocyte differentiation and normal long bone development.[225] Increased activity of the murine equivalent of *PREP* (prolyl endopeptidase) has been reported in naturally-occurring temporomandibular joint osteoarthrosis, suggesting that it plays an important role in cartilage metabolism.[251] Additionally, another member of the prolyl oligopeptidase family has been shown to play important roles in tissue remodeling within the cartilage primordia during embryonic development.[252] Finally, although *SEMA5A* is primarily known for its role in axonal guidance during neural development, it has also been shown to play an important role in angiogenesis, supporting migration of endothelial cells from pre-existing vessels and facilitating extracellular matrix breakdown via matrix metalloproteinase 9 (MMP9).[253] Vascular abnormalities, particularly failure to establish appropriate anastomoses during endochondral ossification, are thought to be central to the pathophysiology of OC[254], and MMP9 is known to be an important mediator of cartilage breakdown and plays a role in growth plate cartilage response to injury.[255]

Limitations: Although the use of next-generation sequencing allowed variant discovery in a larger pool than Sanger sequencing would have, it was still a relatively small group of individuals. Thus, it is possible that many variants present within the larger population were not detected and therefore not available to follow up. This

limitation would best be addressed by sequencing additional horses, although it is not certain that the benefits of this approach would outweigh the additional cost. Alternatively, the availability of a shared resource for sequence within the equine genetics community, similar to the 1000 Genomes project, would be of great value to individual researchers to augment project-specific sequencing. A second limitation is that variants were selected for follow-up largely based on SnpEff annotations. Since this is based on the current equine reference genome, which is known to be incompletely annotated, it is possible that variants with functional effect were missed. An updated "EquCab 3.0" with improved annotation is currently under development, however, in the interim, this limitation could be addressed by manually annotating regions of interest by comparison with human and mouse genomes. Finally, only variants within or near protein-coding genes were selected for follow-up. As mentioned above, this ignores the vast number of regulatory regions throughout the genome, and it is possible that important risk alleles were missed. The development of "AgENCODE" (discovery and annotation of regulatory elements in agricultural species) will help to address this limitation, but this resource is not likely to be available for some time.

Future Directions: The results from the Sequenom analysis for individual SNPs were only marginally significant if a conservative Bonferroni correction is applied. However, since there are likely many alleles interacting with each other to confer risk for OC, evaluation of individual SNPs may not be the most informative approach. An additional approach that might be considered is Gene Set Enrichment Analysis (GSEA).[256] GSEA assesses the significance of known pathways in expression or genotyping data and can reveal biologically relevant pathway enrichment even among

genes that are not individually significant. A "seed" gene list is provided for this analysis that would include genes that have been identified as important in skeletogenesis, particularly endochondral ossification. GSEA has been used to identify pathways involved in risk of development of other complex diseases, including lung cancer.[257] Alternatively, a computational approach could be taken to determine interactions between SNPs and their relative contributions to the OC phenotype. In a random forest approach to a case-control study, the predicted probability of an individual being affected or unaffected with disease is based on the aggregation of a number of decision trees.[258;259] Within these decision trees, each node is an attribute – in this case, the genotype at a given SNP. The importance of each SNP is determined by quantifying the increase of misclassified individuals when the genotype at that SNP is randomly permuted.[258] This approach requires no prior knowledge of gene function and can accommodate multiple variants within the same gene. Random forest analysis has been used successfully to identify pathway-phenotype associations in complex diseases such as bladder cancer in humans[259] and economically important traits such as feed efficiency in cattle.[260] Either approach might reveal novel interactions between genes and/or specific variants that play a role in the development of OC.

It will be important to validate the findings reported here in one or more independent populations before declaring the top variants from this study to be true risk alleles for OC. As mentioned in **Chapter 4**, an appropriate second population in which to follow up these results might be similar to the one reported by Lykkjen et al. (2010)[65] which consists of Standardbreds phenotyped for tarsal OC. However, to determine if putative risk alleles are specific to the Standardbred breed, or to tarsal OC, or are

universal risk alleles for OC (i.e. across all predilection sites and breeds), validation in Standardbreds with OC at locations other than the tarsus, and in additional breeds, will be required. The pre-existing Sequenom assay could be applied to any number of additional individuals. However, as only a small fraction of the discovered variants within chromosomal regions of interest could be evaluated using this method, it is possible that additional fine mapping may be necessary to identify the actual functional variants underlying disease risk. The long-term goal would be to construct a genetic risk model for OC that allows for genetic testing and quantification of risk in individual horses. This risk model will likely contain 6-15 putative risk alleles, similar to those that have been used successfully to predict recurrence and survival in patients with cancer.[261] Improved risk assessment will facilitate management changes and early intervention in high-risk horses and allow for informed breeding decisions in high-risk pedigrees.

Table 1: Summary of 18 individuals selected for whole-genome sequencing (from a total cohort of 162 horses). An "x" in the column of a given chromosome indicates that the individual had the haplotype of interest for that region of interest. Depth of coverage for whole-genome sequencing is indicated in the "coverage" column. M = mare; G = gelding; S = stallion; P = pacer; T = trotter; OC+ = affected; OC- = unaffected.

	Gender	Gait	Sire	ECA2	ECA6	ECA14	Coverage
OC+	M	P	Western Ideal	x	x	x	12x
	M	T	Glidemaster	x			6x
	M	P	Yankee Cruiser	x	x		6x
	G	T	Andover Hall	x	x	x	12x
	M	T	Cantab Hall		x	x	6x
	M	P	Somebeachsomewhere		x	x	6x
	S	P	Western Ideal	x	x	x	6x
	M	T	SJs Caviar		x	x	12x
	G	P	Allamerican Native			x	6x
OC-	S	P	Badlands Hanover	x	x	x	6x
	S	T	Cantab Hall	x	x	x	6x
	M	P	Dragon Again	x			12x
	M	T	Credit Winner		x	x	6x
	M	P	Somebeachsomewhere	x	x		12x
	S	T	Muscles Yankee		x	x	6x
	S	T	Revenue S			x	12x
	S	P	Cam's Card Shark	x		x	6x
	S	T	Windsong's Legacy		x	x	6x

Table 2: Summary metrics for whole-genome sequencing of 18 horses. High quality reads were aligned with a mapping quality of phred-scale Q20 or higher (1/100 or smaller chance of error). Error rate is the percentage of bases that are mismatched with the reference in the high quality aligned reads. OC status: + = affected; - = unaffected.

	OC Status	Total Reads	High Quality Aligned Reads	Error Rate	Mean Read Length (bp)
M968	+	93,947,899	92,086,487	0.0038	94
M977	+	165,878,647	162,796,957	0.0038	94
M989	+	91,232,388	89,376,289	0.0038	94
M992	-	66,950,985	65,622,111	0.0041	93
M1005	-	81,144,917	79,500,675	0.004	94
M1009	+	99,788,604	97,950,326	0.0062	98
M1012	-	98,326,500	96,238,544	0.0062	98
M1027	+	166,302,245	162,966,092	0.004	93
M1048	-	99,232,541	97,442,218	0.0062	98
M5256	-	150,507,994	147,636,767	0.0038	94
M5259	+	76,150,808	74,930,615	0.0062	98
M5260	+	63,823,525	62,743,449	0.0063	98
M5269	+	140,001,498	137,479,949	0.0038	94
M5271	+	70,289,886	69,034,449	0.0063	98
M5287	-	58,972,694	57,901,005	0.0062	98
M5300	-	146,203,843	143,486,911	0.0039	94
M5304	-	67,130,607	65,827,168	0.0064	98
M5306	-	123,271,316	121,016,175	0.0038	94

Table 3: Summary of variants by type and region. Some variants were predicted to have more than one possible effect, so were assigned to more than one type or region.

TYPE			REGION		
Type	**Number**	**Percent**	**Type**	**Number**	**Percent**
codon change + codon deletion	119	0.001	downstream	802,140	4.65
codon change + codon insertion	66	<0.001	exon	170,216	0.99
codon deletion	155	0.001	intergenic	9,741,652	56.43
codon insertion	114	0.001	intron	4,438,806	25.71
downstream	802,140	4.65	none	1,221,763	7.08
exon	20,282	0.12	splice site acceptor	921	0.005
exon deleted	2	<0.001	splice site donor	1,099	0.006
frame shift	6,946	0.04	upstream	867,574	5.03
intergenic	9,741,652	56.43	3’UTR	13,024	0.08
intragenic	101	0.001	5’UTR	7,472	0.04
intron	4,438,806	25.71			
none	1,221,763	7.08			
nonsynonymous coding	56,668	0.33			
nonsynonymous start	16	<0.001			
splice site acceptor	921	0.005			
splice site donor	1,099	0.006			
start gained	746	0.004			
start lost	58	<0.001			
stop gained	596	0.003			
stop lost	40	<0.001			
synonymous coding	85,097	0.49			
synonymous start	3	<0.001			
synonymous stop	54	<0.001			
upstream	867,574	5.03			
3’UTR	13,024	0.08			
5’UTR	6,726	0.04			

Table 4: Variants of predicted functional effect within broad regions of interest based on a GWAS performed for OC in 162 horses. Variants that were initially selected for follow-up using RFLP or Sanger sequencing are highlighted. Eventually, only the four variants on ECA6 were actually followed up with these techniques. The p-value is for a genotypic model, as calculated by SnpSift CaseControl. NS = nonsynonymous mutation; Homo = homozygous for alternate allele; Het = heterozygous for alternate allele.

Region	bp	predicted effect	gene	case homo/het	control homo/het	p-value
ECA1 118Mb	118292615	NS	*ODF3L1*	5/2	8/1	4.11×10^{-2}
	118292698	NS	*ODF3L1*	5/3	8/1	5.12×10^{-2}
	11829330	NS	*ODF3L1*	5/3	8/1	5.12×10^{-2}
ECA2 70-78Mb	72866120	NS	*FSTL5*	1/6	0/2	9.05×10^{-3}
	76167028	NS	*FAM198B*	1/3	5/3	1.78×10^{-2}
	76167357	NS	*FAM198B*	2/1	4/2	9.46×10^{-2}
	77012861	NS	*GRIA2*	8/1	4/2	2.85×10^{-2}
ECA2 98-100Mb	98927677	frame shift	*ENS3042*	2/3	0/2	2.28×10^{-2}
ECA6 24-26Mb	24075423	NS	*SCLY*	2/4	6/3	2.20×10^{-2}
	24123697	NS	*ESPNL*	1/4	6/2	9.87×10^{-3}
	24232581	NS	*PER2*	2/6	7/2	9.05×10^{-3}
	24339281	NS	*TRAF3IP1*	2/4	6/2	6.34×10^{-2}
ECA10 55-57Mb	55677884	NS	*PREP*	1/2	0/0	3.20×10^{-2}
	56695299	start gained	*AIM1*	1/2	0/1	6.38×10^{-2}
ECA14 15-19Mb	18198820	NS	*GABRA6*	3/3	7/2	1.45×10^{-2}
	18198966	NS	*GABRA6*	2/1	3/3	1.35×10^{-1}
	18209193	NS	*GABRA6*	0/1	0/6	1.18×10^{-2}
	18322233	NS	*GABRB2*	2/2	6/3	5.46×10^{-3}
ECA14 32-39Mb	35110220	NS	*KIAA0141*	1/6	1/3	7.79×10^{-2}
	35832132	NS	*PCDHB2*	0/7	1/3	2.07×10^{-1}
	35832191	NS	*PCDHB2*	0/8	1/3	1.03×10^{-1}
	36471455	NS	*SLC4A9*	2/5	1/3	9.26×10^{-2}
	37348824	NS	*MATR3*	0/4	0/1	9.91×10^{-2}
ECA15 27-29Mb	28635757	NS	*CCDC142*	1/5	4/3	1.03×10^{-1}
	28684616	NS	*C2orf81*	4/4	1/4	4.89×10^{-2}
	28685440	NS	*C2orf81*	4/3	1/5	4.51×10^{-2}
ECA21	52586463	NS	*MED10*	6/3	2/5	1.70×10^{-2}

47-55Mb	52586465	frame shift	*MED10*	8/1	2/6	3.02×10^{-3}
	52586476	frame shift	*MED10*	5/4	2/4	2.34×10^{-2}

ODF3L1: outer dense fiber of sperm tails 3-like 1
FSTL5: follistatin-like 5
FAM198B: family with sequence similarity 198, member B
GRIA2: glutamate receptor, ionotropic, AMPA2
ENS3042: novel predicted protein-coding ENSECAG00000003042
SCLY: selenocysteine lyase
ESPNL: espin-like
PER2: period circadian clock 2
TRAF3IP1: TNF receptor-associated factor 3 interacting protein 1
PREP: prolyl endopeptidase
AIM1: absent in melanoma 1
GABRA6: gamma-aminobutyric acid (GABA) A receptor, alpha 6
GABRB2: GABA A receptor, beta 2
KIAA0141: KIAA0141
PCDHB2: protocadherin beta 2
SLC4A9: solute carrier family 4, sodium bicarbonate cotransporter, member 9
MATR3: matrin 3
CCDC142: coiled-coil domain containing 142
C2orf18: chromosome 2 open reading frame 81

Table 5: Summary of 98 ancestry informative markers (AIMs) included on the Sequenom assay. ECA = equine chromosome; BP = base pair.

ECA	BP		ECA	BP		ECA	BP
1	26033759		6	51217265		16	30093970
1	31093333		7	30488035		16	82803597
1	36023288		7	89613245		17	21816238
1	77690698		8	6785301		17	51747591
1	98055129		8	24411688		18	14108728
1	101165799		8	76124128		18	61433985
1	114904651		8	81128171		18	78466909
1	121506138		8	91677016		19	2769487
1	122002161		9	32762141		19	40676833
1	132799749		9	50598697		19	45976675
1	171153599		9	59641922		20	23355754
1	178233312		10	15121923		20	61202862
2	570369		10	60511069		21	15019427
2	5875892		11	50758904		22	7497373
2	8981099		11	54196626		22	22921048
2	19755735		11	60287593		23	12912268
2	59537511		12	24911632		23	16879993
2	73953110		13	2038848		23	22113070
2	92833858		13	8585372		23	25501899
3	21799559		13	10076303		23	28319673
3	23661007		13	20421961		23	34664623
3	62784961		14	35687375		23	37253564
3	89085846		14	44985427		24	36459249
3	89642978		14	51513604		26	5658510
3	107488067		14	81448528		26	23350773
4	73237092		15	9687152		26	33986358
4	86531899		15	27839435		27	24849168
4	93380231		15	40905019		28	30534194
5	2238907		15	46042792		28	35326133
5	5581906		15	64311775		29	5002164
5	14281995		15	80094415		31	2520050
5	24558888		16	7245207		31	5221387
5	61309309						
6	46740223						

Table 6: Summary of 240 SNPs putatively associated with OC that were selected for inclusion in the Sequenom assay. SNPs were multiplexed in groups of 48 in five separate wells.

WELL	ECA	BP		WELL	ECA	BP
W1	2	98927433		W4	21	50321052
W1	18	75969689		W4	21	52586460
W1	18	77703956		W4	10	57468165
W1	18	75780078		W4	14	35652608
W1	21	4598516		W4	14	36115804
W1	2	29591774		W4	14	37213468
W1	18	77894560		W4	14	35832191
W1	18	75504584		W4	16	36711709
W1	2	32190656		W4	14	38496150
W1	16	34620840		W4	10	59685445
W1	18	77725062		W4	14	67874878
W1	14	15545589		W4	21	4707138
W1	21	4800562		W4	10	57209370
W1	14	35338969		W4	1	117892759
W1	16	41804869		W4	14	38120832
W1	2	32641671		W4	14	35359271
W1	18	76006633		W4	10	57134088
W1	2	18190159		W4	14	38261144
W1	16	17404735		W4	1	118771557
W1	14	18322233		W4	14	38237645
W1	14	38234471		W4	10	57167028
W1	21	7509248		W4	14	35042619
W1	16	20901110		W4	14	16830511
W1	16	41546606		W4	14	36226740
W1	21	6300122		W4	14	37321714
W1	18	40478429		W4	10	56817838
W1	14	18034557		W4	1	118401125
W1	21	51305453		W4	14	34391965
W1	18	75992716		W4	14	35638840
W1	14	16782922		W4	14	16776824
W1	16	12795866		W4	14	18528304
W1	14	72832737		W4	1	117907704
W1	1	117899604		W4	21	53794214

W1	14	35710575	[redacted]	W4	1	140238061
W1	18	75187666	[redacted]	W4	14	38421896
W1	5	77353905	[redacted]	W4	14	36386541
W1	2	30971496	[redacted]	W4	10	55518157
W1	5	78709303	[redacted]	W4	14	18209193
W1	14	35733635	[redacted]	W4	14	72832742
W1	14	38640879	[redacted]	W4	14	70569943
W1	2	98925499	[redacted]	W4	14	16857186
W1	14	73999237	[redacted]	W4	14	36762857
W1	14	16538670	[redacted]	W4	16	38404778
W1	21	51402003	[redacted]	W4	14	38740729
W1	14	35681098	[redacted]	W4	14	18029925
W1	21	53443537	[redacted]	W4	14	36302342
W1	21	49950751	[redacted]	W4	14	16854653
W1	14	37281732	[redacted]	W4	21	51325270
W2	14	37348824	[redacted]	W5	14	36243090
W2	16	43467550	[redacted]	W5	21	4512138
W2	4	6190928	[redacted]	W5	2	28136111
W2	16	23943994	[redacted]	W5	16	41459922
W2	16	34073553	[redacted]	W5	3	89027561
W2	10	59079917	[redacted]	W5	2	33902705
W2	1	118105257	[redacted]	W5	2	30971463
W2	14	35749215	[redacted]	W5	14	36975745
W2	18	39910627	[redacted]	W5	16	39306626
W2	16	14358731	[redacted]	W5	2	23390833
W2	16	34954141	[redacted]	W5	10	80739334
W2	10	80792903	[redacted]	W5	14	35160061
W2	16	20892756	[redacted]	W5	21	51353146
W2	21	4800528	[redacted]	W5	1	140205123
W2	16	41787035	[redacted]	W5	14	36174501
W2	18	44859933	[redacted]	W5	14	36078935
W2	5	77353904	[redacted]	W5	16	41794953
W2	10	55605051	[redacted]	W5	21	51448245
W2	5	77536297	[redacted]	W5	14	35581792
W2	14	38011286	[redacted]	W5	16	38384099
W2	3	88076689	[redacted]	W5	1	139695746
W2	18	40807803	[redacted]	W5	21	49216451
W2	14	36239254	[redacted]	W5	2	19959258
W2	10	80690472	[redacted]	W5	21	4898282
W2	1	139375281	[redacted]	W5	14	72031059
W2	21	4515908	[redacted]	W5	14	17882983

W2	10	59873648		W5	14	37568111
W2	1	139685697		W5	21	50383063
W2	14	36012505		W5	21	6611863
W2	14	34929440		W5	14	38224509
W2	1	117511240		W5	21	50238955
W2	21	52365784		W5	14	38578258
W2	1	118839637		W5	14	38157667
W2	14	33108459		W5	14	38297106
W2	18	46490552		W5	21	53288223
W2	18	42386473		W5	14	35110220
W2	3	107352236		W5	14	36321021
W2	10	57350466		W5	21	50250540
W2	14	16840478		W5	10	56789024
W2	21	53928489		W5	14	35796385
W2	21	50348105		W5	14	36231214
W2	1	117508428		W5	14	35713816
W2	1	117896863		W5	1	118293860
W2	10	55512346		W5	14	34945056
W2	16	20876274		W5	10	55657837
W2	14	33234861		W5	1	117503692
W2	10	57303131		W5	14	34256372
W2	14	34940505		W5	14	32504217
W3	1	118796012				
W3	14	34803961				
W3	1	118846185				
W3	14	16802524				
W3	14	18323534				
W3	14	36238870				
W3	16	43285189				
W3	14	18757945				
W3	14	35480068				
W3	14	36270564				
W3	2	31863561				
W3	4	5924012				
W3	14	36627081				
W3	16	41794782				
W3	10	56727782				
W3	1	117500403				
W3	14	18198820				
W3	1	139944477				
W3	1	118324956				

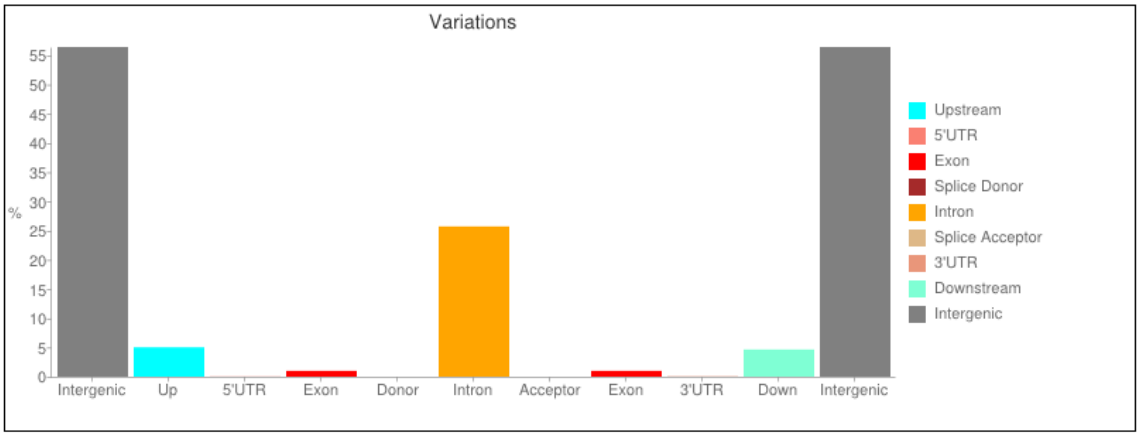
W3	21	52594185		
W3	14	16857276		
W3	14	17365436		
W3	14	37127327		
W3	14	34156670		
W3	14	33691422		
W3	14	36098913		
W3	14	33820804		
W3	21	53591449		
W3	14	35750986		
W3	14	35353077		
W3	21	51408645		
W3	2	99999249		
W3	2	30472121		
W3	21	48664783		
W3	21	49368721		
W3	14	71698112		
W3	14	35363931		
W3	14	35727280		
W3	14	17825358		
W3	21	6590487		
W3	21	49882816		
W3	14	18059791		
W3	14	17829592		
W3	14	34520718		
W3	2	99336592		
W3	18	39195340		
W3	14	16782779		
W3	1	117545952		

Table 7: Comparison of the top 25 association results from PLINK (uncorrected) and GEMMA (corrected for population structure and relatedness) based on Sequenom genotyping data in 180 horses. P-values for GEMMA analysis are based on the likelihood ratio test (Wald test and score test not shown). Highlighted SNPs are described in greater detail in the main text. F_A = frequency of alternate allele in affected population; F_U = frequency of alternate allele in unaffected population; OR = odds ratio.

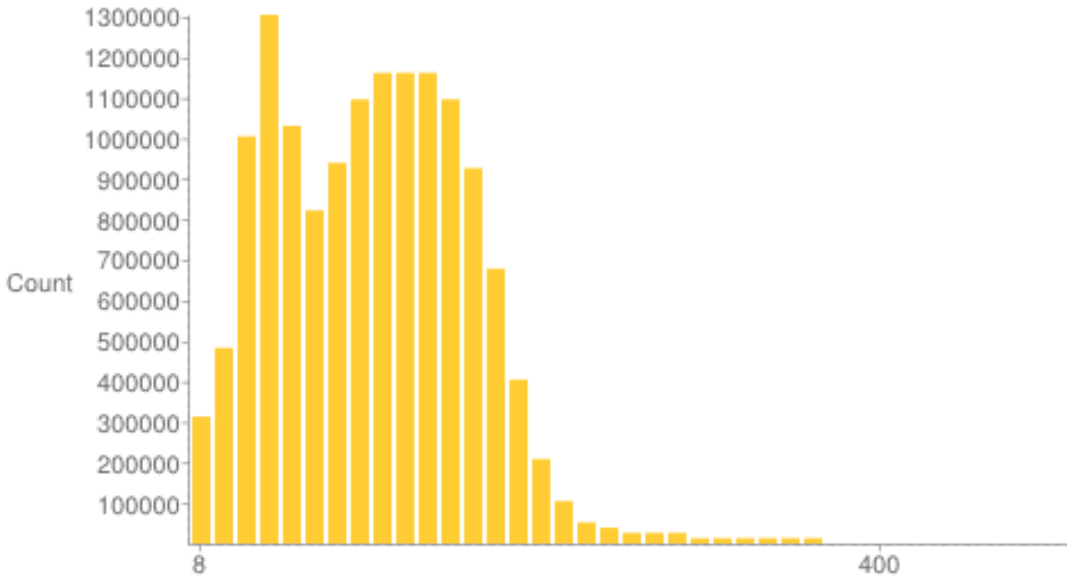
PLINK					GEMMA	
SNP	**F_A**	**F_U**	**P**	**OR**	**SNP**	**P**
chr14.34391965	0.20	0.08	0.001288	2.819	chr14.34391965	0.000867
chr14.37213468	0.36	0.21	0.00235	2.116	chr10.55605051	0.00758
chr21.50348105	0.35	0.21	0.004298	1.99	chr21.50348105	0.00802
chr21.53443537	0.38	0.24	0.004616	1.947	chr2.99999249	0.00929
chr10.55605051	0.17	0.07	0.006654	2.5	chr21.53443537	0.011
chr14.35363931	0.12	0.05	0.009192	2.838	chr14.35363931	0.0113
chr2.99999249	0.06	0.15	0.01034	0.3593	chr10.57350466	0.0153
chr10.57350466	0.54	0.41	0.01224	1.732	chr14.34803961	0.0176
chr21.50383063	0.35	0.23	0.01536	1.792	chr14.37127327	0.0176
chr14.16782922	0.16	0.26	0.0179	0.5201	chr14.16782922	0.0194
chr14.34803961	0.11	0.05	0.01846	2.606	chr21.50383063	0.022
chr14.37127327	0.11	0.05	0.01846	2.606	chr14.18528304	0.0282
chr21.49882816	0.34	0.23	0.01892	1.765	chr21.49882816	0.0314
chr14.18528304	0.31	0.21	0.02742	1.721	chr16.14358731	0.0348
chr16.14358731	0.33	0.44	0.03281	0.6163	chr21.48664783	0.0435
chr10.55657837	0.10	0.04	0.03761	2.552	chr14.16854653	0.0485
chr21.48664783	0.27	0.38	0.03935	0.6147	chr18.46490552	0.0597
chr18.46490552	0.29	0.39	0.04595	0.6286	chr10.57303131	0.0690
chr14.3624309	0.05	0.02	0.04618	3.699	chr14.17365436	0.0691
chr14.16854653	0.13	0.22	0.04997	0.562	chr21.4800562	0.0712
chr10.56817838	0.10	0.05	0.05001	2.275	chr10.56817838	0.0719
chr10.56789024	0.22	0.31	0.05719	0.6075	chr1.140238061	0.0743
chr21.51353146	0.48	0.38	0.05746	1.523	chr14.3849615	0.0749
chr14.17882983	0.10	0.17	0.05809	0.5222	chr14.3572728	0.0753
chr14.17365436	0.36	0.46	0.0628	0.6623	chr14.18034557	0.0782

Figure 1: Graphical summaries of variant calling. (A) Distribution of called variants by type/location. (B) Distribution of depth of coverage over called variants. (C) Distribution of insertions and deletions by length. (D) Alternate allele frequency distribution among 36 chromosomes (18 individuals).

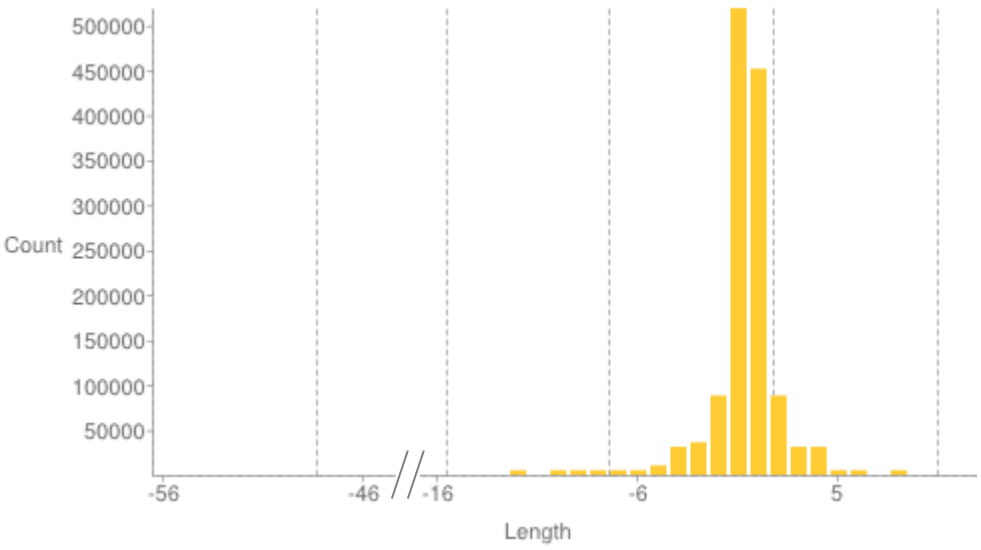
(A)



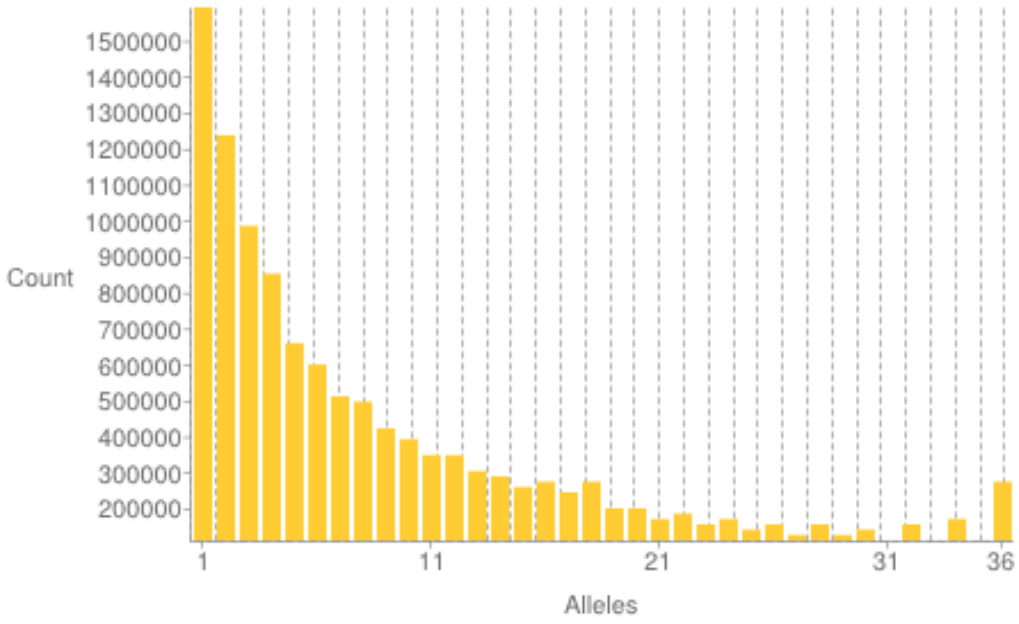
(B)



(C)


500000
450000
400000
350000
300000
Count 250000
200000
150000
100000
50000
-56
-46
-16
-6
5
Length

(D)


1500000
1400000
1300000
1200000
1100000
1000000
Count 900000
800000
700000
600000
500000
400000
300000
200000
1
11
21
31
36
Alleles

PART II:

Genetic Determinants of Gait and Performance in the Standardbred

Chapter 7

Investigation of Putative Modifying Variants Underlying Pacing in the Standardbred

Investigation of Putative Modifying Variants Underlying Pacing in the Standardbred

Annette M. McCoy

From the Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA

Acknowledgements: Thanks to R.J. Piercy (Royal Veterinary College), S. Lykkjen (Norwegian School of Veterinary Science), A. Moore (Moore Equine Services, Cambridge, ON), and P. Caputo (Paul Caputo, DVM, Pompano Beach, FL) for assistance with sample collection. Thanks to L.S. Andersson and C.-J. Rubin (Swedish Agricultural University) for providing data from pooled whole-genome sequencing. S. Beeson performed genotyping of ECA17 variants and manual curation of the surrounding region under the guidance of the author as part of a DVM Summer Scholars project.

Sources of Funding: Funding provided by USDA-NIFA-AFRI, USDA-NRICGP, Morris Animal Foundation, and Minnesota Agricultural Experiment Station. Dr. McCoy was funded by an institutional NIH T32 Comparative Medicine and Pathology Training Grant (University of Minnesota) and a Doctoral Dissertation Fellowship (University of Minnesota); partial funding for Dr. McCue was provided by NIH NIAMS 1K08AR055713-01A2. Ms. Beeson was funded by the University of Minnesota Summer Scholars Program.

Summary

During the course of domestication, horses have been selected for a number of traits related to their uses in labor and transportation. Among these was the ability to exhibit alternative patterns of locomotion, or gaits, a trait which is unique to horses among quadrupeds. Horses exhibiting these alternative gaits were most frequently prized for their smoothness under saddle (i.e. Tennessee Walking Horse, Missouri Foxtrotter) or their speed while in harness (i.e. pacing Standardbreds). A recently described premature stop codon in the gene *DMRT3* appears to be permissive for "gaitedness" across breeds.[11] However, this mutation is nearly fixed in Standardbreds, despite the fact that not all Standardbreds naturally pace. This suggests that the *DMRT3* mutation is necessary, but not sufficient for pacing ability, and that modifying alleles must exist in a subset of the population. The purpose of this study was to identify putative modifying alleles underlying the ability to pace in the Standardbred horse.

A genome-wide association study (GWAS) was performed in a large group of Standardbred horses phenotyped for gait (n = 374 in the initial cohort, n = 542 in the final cohort). GWA analysis was performed using the program GEMMA. After accounting for population structure and relatedness, thirteen SNPs on five chromosomes (ECA1, 6, 17, 23, 25) reached genome-wide significance ($p < 5 \times 10^{-7}$) in the final study population. Variant discovery within these regions was carried out via whole-genome sequencing in eighteen individuals, and variants were prioritized for follow-up based on segregation with gait in the sequenced horses and predicted functional effect. Of six variants selected for initial follow-up genotyping via Sanger sequencing or restriction enzyme digest, five retained their association with gait in the larger population. For greater efficiency,

genotyping of 303 additional variants was performed in 500 individuals using a high-throughput Sequenom assay. After correction for relatedness, 156 SNPs on twenty chromosomes were statistically significantly associated with gait. Several SNPs were located within feasible candidate genes based on known physiologic roles in neural development, including *PCHD9*, *NAA15*, and *KIF20B*. Additional evaluation of significant variants using Gene Set Enrichment Analysis (GSEA) and/or random forest analysis will help to elucidate relationships among genes/alleles as well as further prioritize variants of interest. Putative modifying alleles will need to be evaluated in independent populations to determine if they are unique to Standardbreds, or to all breeds that pace, or if they are universal "gaitedness" alleles similar to *DMRT3*.

Introduction

The Standardbred is relatively unique among racing breeds in that it was founded upon a performance standard – speed over a mile – rather than upon particular breeding lines.[2] Although its ancestry can be traced back to an imported English Thoroughbred stallion and further influence came from the Hackney and the Morgan, mares of several now-extinct breeds also played a crucial role in the development of the Standardbred. These included the Norfolk Trotter, Narrangansett Pacer, and Canadian Pacer.[3] The Norfolk Trotter was a sturdy carriage horse of great stamina that excelled at trotting and was developed in England by crossing Thoroughbred stallions to "native road mares." Presumably, these mares were the source of this breed's trotting prowess.[262] The introduction of an alternative gait, pacing, to the Standardbred was through the influence of the Narrangansett and Canadian Pacers, both of which were "ambling" breeds developed in North America in the 1700s.[6;7] Narrangansett and Canadian Pacers also contributed to the development of other gaited breeds, including the American Saddlebred, Tennessee Walking Horse, and Rocky Mountain Spotted Horse.[6]

Pacing Standardbreds and horses of other gaited breeds have been specifically selected over generations of breeding for their ability to perform alternative patterns of locomotion. Heritabilities of the pace and tölt in the Icelandic horse have been estimated to range between 0.53 and 0.73,[263] and there are strong signatures of selection evident when comparing gaited and non-gaited breeds.[62] However, until recently, the specific genetic determinants underlying these alternative gaits were completely unknown. In 2012, a mutation was reported in *DMRT3* (an isoform of the doublesex and mab-3 related transcription factor) on equine (ECA) chromosome 23 that appears to be permissive for

"gaitedness" across breeds.[11] A genome-wide association study (GWAS) in four-gaited (walk, trot, tölt, and gallop) and five-gaited (also pace) Icelandic horses revealed a strongly associated SNP on ECA23. Deep (30x coverage) whole-genome sequencing of one four-gaited and one five-gaited individual revealed a premature stop codon in the last exon of *DMRT3*, resulting in protein truncated by 174 amino acids. Subsequent genotyping of additional Icelandic horses revealed that nearly all five-gaited individuals were homozygous for the mutation, compared to only a third of the four-gaited horses. Of even greater interest, when horses of other breeds were genotyped for the mutation, it was found to be nearly fixed in gaited breeds (i.e. Paso Fino, Peruvian Paso, Tennessee Walking Horse, etc.), but absent in non-gaited breeds (i.e. Arabian, Thoroughbred, etc.).[11] The functional importance of *DMRT3* was confirmed in a mouse model. Mice null for *DMRT3* exhibited an abnormal gait characterized by an increased stride, prolonged, stance and swing phases of both the fore and hind limbs, and near absence of coordinated hind limb movements. *DMRT3* was localized to the spinal cord both pre- and postnatally, and null mice had fewer commissural interneurons, suggesting that this gene is important for the development of normal locomotor coordination.[11]

Although the *DMRT3* mutation appears to be necessary for "gaitedness," it is not sufficient for this trait, as demonstrated by the fact that it is nearly fixed in the Standardbred, regardless of whether they pace or not.[11] It is noteworthy that approximately 20% of the offspring of trotter stallions go on to race as pacers.[10] It is unknown whether this is due to genetic predisposition, training, or a combination of the two, but it is likely that modifying genetic factors exist in a subset of the Standardbred population that interact with *DMRT3* and determine an individual's ability to pace. We

hypothesize that across all gaited breeds, including the Standardbred, a combination of modifying alleles determine the specific gait exhibited, and that breeds selected for similar gaits likely share modifying alleles (**Figure 1**). The purpose of this study was to identify putative modifying alleles associated with gait in a large cohort of Standardbred pacers and trotters using a combination of genome-wide association and variant discovery via whole-genome sequencing.

Materials and Methods

Horses: The initial study cohort consisted of 374 Standardbred pacers (n = 84) and trotters (n = 290). Horses were classified on the basis of race records; if a horse never raced, it was classified according to whether the sire and dam trotted or paced. All of the pacers and 77 of the trotters were from North America. Sixty-six trotters were from Sweden, and the remaining 147 trotters were from Norway. The North American and European trotters were related to each other. The pacers belonged to a distinct family separate from the trotters with minimal admixture between groups.

To address concerns over the unequal numbers of pacers and trotters, and the fact that most of the pacers were from a single breeding farm, samples were collected from an additional 168 North American horses (92 pacers and 76 trotters). This resulted in a final cohort of 542 individuals – 176 pacers and 366 trotters. Consistent with the original cohort, the pacers and trotters formed genetically distinct clusters (**Figure 2**).

DNA Isolation and Whole-Genome Genotyping: DNA was isolated from whole blood samples using the Gentra® Puregene® Blood Kit (Qiagen, Valencia, CA) per manufacturer recommendations. Briefly, RBC lysis solution was added to samples at a

3:1 ratio, incubated, and centrifuged. After discarding the supernatant, Cell lysis solution was added to the white blood cell pellet and the cells were re-suspended, after which protein was precipitated and discarded. DNA was precipitated in isopropanol and subsequently washed in ethanol prior to final hydration. Quantity and purity of extracted DNA were assessed using spectrophotometric readings at 260 and 280nm (NanoDrop 1000, Thermo Scientific, Wilmington, DE).

Genome-wide genotyping of single nucleotide polymorphism (SNP) markers was performed by Neogen GeneSeek (Lincoln, NE) using the Illumina Custom Infinum SNP genotyping platform. The majority of the horses in the original study cohort were sampled during the course of unrelated projects in the Equine Genetics and Genomics Laboratory and had already been genotyped either at 54,602 SNPs using the first generation Illumina Equine SNP50 chip (n = 306) or at 65,157 SNP markers using the second generation Illumina Equine SNP70 chip (n = 68). All additional samples in the final study cohort were genotyped using the SNP70 chip (n = 168).

Genotype Imputation: The two equine genotyping platforms share only 45,703 SNPs. While the data can be pruned down to this shared marker list for the purposes of merging files, the information from tens of thousands of markers is lost. Genotype imputation is a technique that statistically estimates genotypes from non-assayed SNPs by comparing haplotype blocks in the study population with haplotype blocks in a more densely genotyped reference population. A pipeline for imputation of equine genotyping data was established and validated utilizing BEAGLE[211] software for imputation (see **Chapter 5**). This pipeline was used to impute the ~18,000 markers unique to the SNP70 chip in those horses genotyped on the SNP50 chip, and likewise to impute the ~9,000

markers unique to the SNP50 chip in those horses genotyped on the SNP70 chip. Imputed files were merged for subsequent analysis using the --merge command in PLINK.[212]

Genotyping for DMRT3 Mutation: Primers (5'-AGAGTCTGCGGAAAA CCTCA-3'/5'-CAACCGAAAGTTCGACTTCC-3') were developed using Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) to encompass ~300bp surrounding the published *DMRT3* mutation (Ser301STOP) based on the EquCab 2.0 gene sequence as reported in Ensembl (https://useast.ensembl.org/Equus_caballus/Info/ Index; accessed 21 Feb 2013). Sanger sequencing of three Standardbred horses confirmed that the PCR product was targeted correctly. Genotyping in the entire study cohort (n = 542) was subsequently performed by restriction fragment length polymorphism (RFLP) using the enzyme *DdeI*. A sequenced Standardbred served as a positive control (A/A), while a Thoroughbred served as a negative control (C/C).

Genome-Wide Association (GWA) Analysis: A GWA analysis with gait as the phenotype of interest was performed after imputation in the original study cohort (n = 374) using GEMMA (Genome-wide Efficient Mixed Model Analysis) software.[214] This program accounts for population structure and relatedness through the use of a marker-based relationship matrix, can incorporate covariates into the model, is highly efficient for large data sets, and estimates variance for each individual SNP (rather than average variance across all SNPs). The GWA was carried out using the options to create a centered relatedness matrix [-gk 2] and perform all three possible frequentist tests: Wald, likelihood ratio, and score [-fa 4]. A covariate file including gender and origin (North America or Europe) was incorporated into the mixed model [-c]. SNPs were pruned prior to GWA using the default GEMMA parameters of MAF <1% and missingness <95%.

GWA analysis in GEMMA was repeated in the final study cohort (n = 542) as described above except that the relatedness matrix was constructed using a linkage-disequilibrium (LD)-pruned set of markers (100 SNP windows, sliding by 25 SNPs along the genome, pruned at $r^2 > 0.2$; PLINK command --indep-pairwise 100 25 0.2).[59] The analysis was run with and without the use of gender and origin covariates.

Association plots were generated using the base graphics package in the R statistical computing environment.[187] On the basis of previously published guidelines, p-values of less than 5×10^{-7} were considered to indicate genome-wide significant association and p-values between 5×10^{-5} and 5×10^{-7} were considered to indicate moderate association.[67]

Whole-Genome Sequencing: Nine pacers and nine trotters were selected from the study cohort for whole-genome sequencing. A detailed description of the selection process for these individuals can be found in **Chapter 6**. Genomic DNA (2-6μg) was submitted to the University of Minnesota Genomics Center (UMGC) and subjected to standard library preparation including fragmentation, polishing and adaptor ligation. All samples were labeled with an indexed barcode for a paired-end run on the Illumina HiSeq sequencer. Depth of coverage was determined by the calculation C = LN/G, where C is the coverage, L is the read length, N is the number of reads, and G is the genome length. For 100bp paired-end reads, given an approximate genome length of 3 billion base pairs, 160 million reads results in ~12x coverage of the genome, and 80 million reads results in ~6x coverage of the genome. Of the nine pacers, 3 were sequenced at 12x coverage and 6 at 6x coverage; the same distribution was used for the nine trotters (**Table 3**). Samples were split between all eight lanes on each of two flowcells (10 samples on one flowcell, 8

on the other). Raw sequence data was deposited within designated storage at the Minnesota Supercomputing Institute (MSI).

All data analysis, including quality control, alignment and variant detection, was carried out following published best practices[249] within the Galaxy framework hosted by MSI. A detailed description of this data analysis can be found in **Chapter 6**. Briefly, reads were mapped to the reference sequence (EquCab 2.0) using BWA for Illumina. Variant discovery was carried out using the UnifiedGenotyper utility of the Broad Institute's Genome Analysis ToolKit (GATK). Predicted functional effect for each called variant was determined based on the current equine reference genome annotation using the SnpEff tool in GATK. Frequency of variants within pacers and trotters, and the significance of frequency differences, was calculated using the SnpSift CaseControl tool in GATK. Variants from particular chromosomal regions of interest were selected using SnpSift Intervals and converted into Excel format for further evaluation.

Variant Prioritization and Follow-up with RFLP or Sequencing: Based on the GWAS performed in 374 horses (see *Genome-Wide Association (GWA) Analysis*), variants were evaluated within the regions of interest on ECA17 and 13 for 1) predicted functional effect; and 2) segregation with gait. Variants with predicted functional effect that segregated nearly perfectly with gait (i.e. present in 7 or more pacers and no more than 2 trotters, or vice versa) were considered for follow-up. Primers were designed to amplify regions surrounding 12 variants on ECA17 using Primer3 (http://biotools.umassmed.edu/bioapps/primer3_ www.cgi). Restriction enzymes were successfully designed for six of these SNPs using NEBcutter v2.0 (New England BioLabs, Inc., http://tools.neb.com/NEBcutter2/); the remainder of the SNPs required

genotyping by sequencing. Six of these variants were selected for follow-up on the basis of known gene function and/or potential impact of the mutation (four genotyped by RFLP and two by sequencing). PCR reactions were carried out under the following thermocycler conditions: 20 min at 95°C; 35 cycles of 30 sec at 94°C, 30 sec at 58° or 60°C, 30 sec at 72°C; 15 min at 72°C. PCR products were visualized on 2% agarose gels with ethidium bromide prior to submission for sequencing. To prepare the PCR products for Sanger sequencing, 1µl USB® ExoSAP-IT® (Affymetrix, Santa Clara, CA) was added to 4µl PCR product and incubated in the thermocycler for 15 min at 37°C for PCR product cleanup followed by 15 min at 80°C for enzyme inactivation. Subsequently, 1µl [20µM] primer and 6µl water were added and the sample (12µl total volume) submitted to the UMGC for sequencing. Enzyme digests were performed according to manufacturer recommendation for each enzyme.

Sequenom Assay: While follow-up of individual variants on ECA17 provided proof of principal for the experimental approach, it was impractical to use this technique to evaluate variants across all of the regions of interest identified on the GWAS in the final study cohort. Therefore, a custom Sequenom assay was designed for high-throughput genotyping within the study cohort. The majority of the variants were selected from the top regions of interest in the GWAS. These included: ECA1 5.3-5.6Mb, 17.5-18.1Mb, and 35.3-55.1Mb, ECA2 17.3-19.7Mb, ECA3 2.4-3.9Mb and 44.1-77.7Mb, ECA6 6.5-7.8Mb, ECA9 7.5Mb, ECA11 46.8-58.4Mb, ECA16 25.7-31.3Mb, ECA17 27.6-29.3Mb, 39.4Mb, and 60.5Mb, ECA18 77.6Mb, ECA19 21.4-37.9Mb, ECA23 14.1-14.9Mb, ECA25 11.5-16.9Mb, and ECA26 3.3-3.6Mb.

Additional variants were selected from chromosomal regions found to have high differentiation between pacers and trotters (F_{ST} > 0.35) or a combination of low heterozygosity (het < 0.1) in one of the groups and high differentiation (F_{ST} > 0.30). These data were generated by collaborators at Uppsala based on pooled whole-genome sequencing data in 20 pacers and 20 trotters selected from our study cohort. The included individuals were specifically chosen to be as unrelated as possible to each other on the basis of coancestry coefficients. Selected pacers had coancestry coefficients <0.06 (no more closely related than first cousins); selected trotters had coancestry coefficients <0.14 (one pair of half-siblings, the rest less closely related). Some of these regions overlapped regions of interest identified from the GWAS, and additional variants were generally not selected from these regions. Regions of interest from this data set from which variants were selected included: ECA1 38.5-38.8Mb and 106.8-106.9Mb, ECA3 24.8-25.2Mb and 52.3Mb, ECA4 89.9-91.1Mb, ECA5 55.3-81.7Mb, ECA6 81.3-81.7Mb, ECA9 29.1-29.2Mb, ECA11 29.5-36.8Mb, ECA12 16.2-16.4Mb, ECA14 1.3-1.7Mb and 5.4Mb, ECA15 10.1Mb, ECA16 59.3Mb, ECA17 61.7-65.7Mb, ECA20 25.1-27.7Mb and 46.7-47.1Mb, ECA23 20.6-20.7Mb, ECA24 6.7-10.3Mb, ECA25 11.7-15.1Mb, ECA29 10.0-10.1Mb, and ECA30 14.0-15.1Mb.

Variants discovered through individual whole-genome sequencing were filtered to include only SNPs (no indels) that passed all quality control filters, and were subsequently prioritized according to the following parameters:

1) segregation with gait by one of 4 measures (in order of priority):
    a. for the alternate allele, no trotter homozygotes and ≥ 5 pacer homozygotes;

b. present in 8 or 9 pacers and ≤4 trotters;

c. present in 7 pacers and ≤2 trotters;

d. variant of important predicted effect present in more pacers than trotters within a region of interest with no other potential markers to use;

2) not intergenic;

3) non-synonymous, then synonymous changes;

4) if intronic, close to intron-exon boundary (preferably <100bp);

5) coding genes preferred over non-coding; and

6) if upstream/downstream, as close as possible to start/stop codon.

Variants from pooled sequencing data were prioritized on criteria 2-6. In addition to the experimental SNPs, 98 ancestry informative markers (AIMs) were included in the Sequenom assay to help control for population structure (see **Chapter 6, Table 5** for a list of these AIMs).

Genotyping results were analyzed using GEMMA to account for population structure and relatedness. The association test in GEMMA was performed using the options to create a centered relatedness matrix (-gk2) and perform all three possible frequentist tests: Wald, likelihood ratio, and score (-fa 4). The relatedness matrix was constructed using the AIMs. SNPs were pruned prior to analysis using the default GEMMA parameters of MAF <1% and missingness <95%.

Results

Genotyping for DMRT3 Mutation: All 542 horses were successfully genotyped for the *DMRT3* Ser301STOP mutation by RFLP. All were homozygous for the mutation (A/A), supporting published findings in which the mutation was present in 100% of US Standardbred pacers and trotters and 97% of Swedish Standardbred trotters.[11] Since every individual was homozygous for this mutation, the genotype at this locus was not included in the GWA mixed model analysis.

GWA Results for Original Study Cohort: After pruning, 40,616 SNPs were available for the final analysis in 374 horses. After correction for population structure and relatedness, regions on two equine (ECA) chromosomes, ECA13 and ECA17, were found to be highly significantly associated with gait, with five SNPs reaching genome-wide significance ($p < 5 \times 10^{-7}$) (**Figure 3**). **Table 1** reports the p-values using three different test statistics for the top 50 SNPs in this mixed model analysis. Four of the genome-wide significant SNPs were located within two consecutive, though not contiguous regions on ECA17. Three SNPs were loosely clustered from ~40.9-45.7Mb (with slightly less significant hits at 38.5 and 20.7Mb) and a single SNP was located at 60.5Mb (with a second slightly less significant SNP within 50kb). There are 4 genes within the region from ~40.9-45.7Mb. The SNP at 60.5Mb is approximately 100kb downstream from an unnamed predicted protein-coding gene and 1.3Mb upstream from a cluster of micro-RNAs and the gene *GPC5* (glypican 5). The hit on ECA13 was located between the closely-spaced genes *GET4* (golgi to ER traffic protein 4 homolog [S. cerevisiae]) and *SUN1* (Sad1 and UNC84 domain containing 1). Thirty-seven additional SNPs on 16 chromosomes were moderately associated with gait ($p < 5 \times 10^{-5}$).

GWA Results for Final Study Cohort: Among the final study cohort (n = 542), 306 were genotyped on the Equine SNP50 chip and 236 were genotyped on the Equine SNP70 chip. In order to combine these data without the loss of marker information, genotype imputation was performed (see Materials and Methods). Imputation was successfully carried out in the final study cohort using the pipeline described above and in **Chapter 5**. After imputation, there were 73,691 markers in the complete data set, an increase of nearly 28,000 markers over the shared set. SNP pruning for MAF and genotyping success was performed during mixed model analysis in GEMMA as described above. After pruning, 62,901 SNPs were available for the final analysis in 542 horses. Inclusion of gender and origin as covariates did not alter the analysis, so the results of the simpler model are presented here. After correction for population structure and relatedness, thirteen SNPs on five chromosomes (ECA1, 6, 17, 23, 25) reached genome-wide significance ($p < 5 \times 10^{-7}$) (**Figure 4**). Similarly to the original GWA analysis, the seven SNPs on ECA17 were located within three consecutive, but not contiguous, regions: 28.5Mb (1 SNP), ~36.3-41.9Mb (3 SNPs), and 60.5Mb (3 SNPs). The SNP at 28.5Mb was within intron 26 of *VWA8* (von Willebrand factor A domain containing 8), and the larger region included one of the four named genes, *PCDH9* (protocadherin 9), from the region defined in the original GWA. The two SNPs on ECA1 were widely spaced, one at 18.1Mb, the other at 155.2Mb. The SNP at 18.1Mb fell between *CASP7* (caspase 7) and *NRAP* (nebulin-related anchoring protein) and was within 100kb of 4 additional named genes. The SNP at the opposite end of ECA1 was in a region with only a large number of predicted single exon "genes." Two SNPs on ECA6 demarcated a region spanning from ~7-7.3Mb, which did not contain any named genes,

but was 300kb downstream of *TNP1* (transition protein 1 [during histone to protamine replacement]). The SNP on ECA23 was located at 14.6Mb and was not related to the previously described *DMRT3* mutation (~23.0Mb), but was instead found to be within intron 1 of a predicted protein-coding gene. The single SNP on ECA25 was located at 2Mb and fell within intron 8 of *PAX5* (paired box 5). One hundred one additional SNPs on twenty chromosomes were moderately associated with gait ($p < 5 \times 10^{-5}$). **Table 2** reports the p-values using three different test statistics for the top 50 SNPs in this mixed model analysis.

Whole-Genome Sequencing: Actual coverage for the twelve individuals sequenced for a target of 6x ranged from 4.7x to 7.9x (mean 6.4x). Actual coverage for the six individuals sequenced for a target of 12x ranged from 10x to 13.1x (mean 12.2x). Summary metrics for sequence alignment and distribution of variants and the predicted effects of 14,588,812 called variants are detailed in **Chapter 6**, **Tables 4-5** and **Figure 1**. ECA17 appeared to have especially large haplotype blocks segregating with gait, including a ~27Mb region from ~19.7-46.7Mb. Within one 5Mb segment of this region (~24-29Mb) there were 712 variants that segregated perfectly with pace; of these, 12 were predicted to have functional effect.

Variant Prioritization and Follow-up with RFLP or Sequencing: A summary of the six variants selected for follow-up is shown in **Table 4**. Of these, the SNP in *FAM124A* did not continue to segregate with gait after initial investigation and was abandoned. The remaining five variants were genotyped in 400-500 additional horses and were highly significantly associated with gait ($p < 2.2 \times 10^{-16}$). For two of the SNPs, located in *PCDH9* and ENSECAG00000016793, the alternate allele was more common

in trotters than in pacers. In contrast, for the variants located in *VWA8, EPSTI1* (epithelial stromal interaction 1 [breast]), and *NAA16* (N(alpha)-acetyltransferase 19, NatA auxiliary subunit), the alternate allele was more common in pacers. In fact, for these three variants, there were no trotters homozygous for the alternate allele. A summary of the genotyping results for these five variants is in **Table 5**.

Sequenom Assay: 303 SNPs were included in the final Sequenom assay, including 190 SNPs from the whole-genome sequencing regions of interest and 113 SNPs from the pooled sequencing regions of interest (**Table 6**). The six SNPs on ECA17 described above (*Variant Prioritization and Follow-up with RFLP or Sequencing*) were not included on the Sequenom assay since they had already been genotyped in the larger population. The 303 SNPs were multiplexed in groups of 48 (except for a single well with 17 SNPs) for genotyping in 500 horses. Most of these horses were selected from the GWAS cohort; however, additional pacers had been sampled in the interim and were included in the Sequenom assay (in total, 262 pacers and 238 trotters were genotyped). After pruning in GEMMA, 244 SNPs were available for analysis. The top SNP (chr30.14067984) had a p-value of 1.71×10^{-31}. After Bonferroni correction, p-values < 2.06×10^{-4} would be considered statistically significant; 156 SNPs on 20 chromosomes met this criteria. The top 25 results from GEMMA analysis are shown in **Table 7**. Within these top hits were multiple SNPs clustered together in 6 loci on 4 chromosomes (ECA1, 17, 23 and 30), as well as two single SNPs on ECA1 and ECA3. Fourteen of these SNPs were derived from the whole-genome sequencing and 11 were from the pooled sequencing. The top SNPs were on ECA30, but were not located within genes; however, the most significantly associated SNP (chr30.14067984) was 750bp upstream of *RRP15*

(ribosomal RNA processing 15 homolog [S. cerevisiae]). The alternate allele for this SNP was found in 78% of trotters and 7% of pacers. Other top SNPs were found within *NHLRC2* (NHL repeat containing 2; ECA1), *FER1L3* (fer-1-like-3, myoferlin; 6 SNPs, ECA1), *KIF20B* (kinesin family member 20B; 3 SNPs, ECA1), *HEATR3* (heat repeat containing 3; ECA3), *VWA8* (von Willebrand factor A domain containing 8; 3 SNPs, ECA17), and *MAMDC2* (MAM domain containing 2; 2 SNPs, ECA23). Three SNPs on ECA23 that were highly associated with gait were located within a novel predicted protein-coding gene of unknown function.

Discussion

The horse is unique among quadrupeds in that it can exhibit alternative patterns of locomotion as a physiologic, rather than pathologic, adaptation. In fact, certain breeds of horses, such as the Standardbred, have been strongly selected for this ability over generations of breeding. Beyond giving insight into an economically important trait, improved understanding of the pathways that underlie alternative gaits in the horse may also provide insight into pathways that are dysregulated with disease in other species. However, a challenge arises in identifying the most appropriate candidate genes to investigate because there is little known about the development of normal limb coordination. It is likely that many genes important to the expression of alternative gaits have not previously been described to have any such function. This is aptly illustrated by *DMRT3*, which had initially been described as primarily playing a role in gonadal development and sexual differentiation.[264] The *DMRT3* nonsense mutation originally reported by Andersson et al. (2012)[11] has now been reported to occur at some frequency

in 68 out of 141 breeds tested from around the world, and at high frequency (>50%) in all "gaited" breeds.[265] This example demonstrates that a strongly associated mutation cannot be ruled out as a putatively functional modifying allele for gait simply because it falls within a gene that does not have a described role in neural development or locomotion.

In this study, GWAS in a large cohort of horses revealed several chromosomal regions highly associated with gait. However, as some of these regions were quite large and contained many genes, a traditional candidate gene approach for variant discovery was not feasible. The use of next-generation sequencing allowed for large-scale variant discovery in 18 individual horses (9 pacers, 9 trotters), as well as in pooled samples from 40 horses (20 pacers, 20 trotters). Of the tens of thousands of variants discovered within regions of interest, only a small fraction could be genotyped in a large population, so it was possible that actual functional alleles would not selected via the prioritization process. However, the genotyping results for six variants genotyped by Sanger sequencing and RFLP provided proof of principal for this approach. Five of these six variants maintained strongly significant association with gait when genotyped in 400-500 horses, and three fell within genes that are plausible candidates for having a role in gait based on the published literature. *VWA8* (KIAA0564) has been associated with risk for autism[266] and bipolar disorder with complex migraine[267] in large human genome-wide association analyses, although a functional mechanism has not been proposed in either case. Little is known about *NAA16*, but its close paralog *NAA15*, also called *NARG1* (NMDA-receptor regulated gene 1) is highly expressed in the developing brains in areas of neuronal proliferation and migration, and may play an important role in neuronal differentiation. While these genes are certainly worth investigating further, the strongest

evidence exists for *PCDH9*. Members of the protocadherin family, including *PCDH9*, are thought to play an important role in various processes during neural development, including cell-cell adhesion, neural projection, and synapse formation. Among other regions, *PCDH9* was localized to the vestibulocochlear nerve, vestibular nuclei, and vestibulocerebellum during early development in the mouse.[268] Given the importance of these in the development of coordinated movement, this makes *PCDH9* an excellent candidate to play a role in the development of alternative gait.

Of the top 25 variants from GEMMA analysis of Sequenom genotyping in 500 horses, sixteen were located within named genes (n = 6). Of these genes, only two, *VWA8* (KIAA0564) and *KIF20B*, have any reported relationship to the central nervous system. *VWA8*'s association with complex neurological disease is discussed above, but *KIF20B* is perhaps even more interesting as a candidate because it has been shown to be crucial in development of the cerebral cortex. Mice with a splice mutant of *KIF20B* resulting in premature stop codons exhibit microencephaly secondary to impaired division and increased apoptosis of neural stem/progenitor cells in the midbrain.[269] *KIF20B* has further been shown to play an important role in migration of polarized neurons in the developing mouse brain, as well as the transition from multipolar to bipolar cell states.[270] Two other genes have known functions unrelated to the central nervous system: *FER1L3* (myoferlin) is important for muscle development[271] and is thought to play a role in tumor invasion[272;273], while *HEATR3* plays a role in the NF-κB pro-inflammatory signaling pathway and has been associated with increased risk of Crohn's Disease in an Ashkenazi Jewish population.[274] Little is known about the function of *NHLRC2*, although it falls within a region that has been associated with Alzheimer's Disease.[275] Similarly,

MAMDC2 falls within a region associated with Kabuki syndrome, a complex disorder with mental retardation and craniofacial deformities, although no specific mutations within the gene were associated with the condition.[276] None of these genes has been investigated for a role in coordinated movement patterns or altered gait.

Limitations: Although variant discovery was carried out in a relatively large number of horses (n = 58) when both the individual and pooled sequencing is considered, there is still a possibility that variants were missed that are present in the larger population. This could be addressed by performing additional whole genome-sequencing, although the benefits of this approach are unlikely to outweigh the costs. Instead, targeted sequencing (i.e. sequence capture) of specific regions of interest might be used to search for additional variants with putative functional effect. A second limitation to this study is that variants were prioritized and selected for follow-up genotyping on the basis of SnpEff annotations. These are based on the current reference genome, which is known to be incompletely annotated. In particular, the first exon of many genes, as well as the 5' and 3' untranslated regions (UTRs) are often missing in the existing annotation. This limitation could be addressed by manual annotation within regions of interest by comparison with the human and mouse genomes, and this was, in fact, done for a region spanning from ~17.2-46.7Mb on ECA17. Although 51 additional variants of putative functional effect were identified within this 29.5Mb region, these results may not justify the substantial time and effort required for this approach across multiple regions (or the entire genome). A more efficient way to do this would be to try to identify putative functional effects of specific markers found to be highly associated with gait that are near annotated genes, but not currently assigned an effect. Release of the updated "EquCab

3.0" reference genome, currently under development, will also aid in addressing this limitation. An offshoot of this limitation is that only variants within or near protein-coding genes were selected for follow-up. There is ample evidence that a much larger portion of the genome is involved in regulatory actions than in protein coding[250], and so it is possible that important modifying alleles were missed by our approach. The development of "AgENCODE," a project seeking to discover and annotate regulatory elements in agricultural species, will help address this limitation in the long term, but this resource is not likely to be available for some time.

Future Directions: Of 303 selected variants from the Sequenom assay, 244 were included in the final GEMMA analysis, and of these, 156 were statistically significantly associated with gait ($p < 2.05 \times 10^{-4}$). Clearly, much work remains to be done to assess these variants for a potential role in gait. Although potential individual candidate genes based on the top 25 variants were discussed above, it is likely that a higher-level analysis will be necessary to tease out which of the 156 significant variants show the greatest promise as potential modifying alleles involved in gait. There are two approaches that may be used here. First, pathway analysis can be used to look for connections between genes (or protein products) based on published evidence of interactions from the literature. Tools available for this type of analysis include IPA (Ingenuity Pathway Analysis; Ingenuity Systems, Redwood, CA), STRING[277;278], ClueGO[279], and GRAIL (Gene Relationships Across Implicated Loci).[280] All of these approaches are based on text mining, so the lack of published information about gait may be a disadvantage to this approach, although for GRAIL, "seed" genes/regions known to be important for neural development could be selected to guide the analysis. Second, a random forest

computational approach can be taken to discover relationships between predictors (SNPs) and to assign numerical contributions of each SNP to the trait of interest (gait) (see **Chapter 6**).[258;281] This approach has the advantage of not requiring any prior knowledge about gene function to establish relationships between pathways and phenotype. A combination of the two approaches is likely to prove of greatest use in future investigations.

It will be important to validate the findings reported here in independent populations. These should include 1) an additional population of Standardbreds phenotyped for gait; 2) a population of Icelandic horses that pace; 3) a population of gaited and non-gaited horses from a variety of breeds. This will help determine if the modifying alleles discovered in this population are unique to Standardbreds, are unique to breeds that pace, or are, like the *DMRT3* mutation, universal alleles related to "gaitedness." The pre-existing Sequenom assay described here could be applied to any number of these additional individuals, but it is likely that as our focus narrows and additional fine mapping is performed, an updated group of selected variants will be considered for follow-up validation.

Table 1: Top 50 SNPs from GEMMA mixed model analysis in 374 individuals (gender and origin covariates). After pruning, analysis included 40,616 SNPs. Uncorrected p-values are presented for the Wald test, the Likelihood ratio test (lrt) and the Score test. CHR = chromosome; BP = base pair.

RANK	CHR	BP	p_wald	p_lrt	p_score
1	17	60554458	9.08E-09	7.39E-09	3.12E-08
2	17	40999944	9.23E-09	7.51E-09	3.16E-08
3	13	6175513	1.54E-08	1.26E-08	4.90E-08
4	17	45678720	3.76E-07	3.20E-07	7.91E-07
5	17	45679062	3.76E-07	3.20E-07	7.91E-07
6	17	38461954	6.19E-07	5.29E-07	1.23E-06
7	17	20690428	1.57E-06	1.36E-06	2.82E-06
8	1	18091709	2.04E-06	1.77E-06	3.57E-06
9	28	6098105	4.23E-06	3.70E-06	6.90E-06
10	19	6093264	5.10E-06	4.47E-06	8.17E-06
11	17	60503138	5.37E-06	4.71E-06	8.56E-06
12	23	14650375	5.90E-06	5.18E-06	9.33E-06
13	1	104906681	5.91E-06	5.19E-06	9.34E-06
14	3	67042453	9.51E-06	8.39E-06	1.44E-05
15	3	67045788	9.51E-06	8.39E-06	1.44E-05
16	25	32055492	9.53E-06	8.41E-06	1.44E-05
17	23	26109618	1.01E-05	8.89E-06	1.52E-05
18	17	41566253	1.31E-05	1.16E-05	1.92E-05
19	17	42347955	1.35E-05	1.19E-05	1.98E-05
20	17	74662387	1.40E-05	1.24E-05	2.05E-05
21	17	32444026	1.46E-05	1.30E-05	2.13E-05
22	3	64348556	1.52E-05	1.35E-05	2.21E-05
23	17	72658830	1.69E-05	1.51E-05	2.44E-05
24	23	20712252	1.96E-05	1.75E-05	2.80E-05
25	20	27347017	2.07E-05	1.84E-05	2.94E-05
26	17	33932846	2.23E-05	1.99E-05	3.14E-05
27	5	98040238	2.58E-05	2.30E-05	3.60E-05
28	32	12381464	2.87E-05	2.57E-05	3.97E-05
29	15	934287	3.28E-05	2.94E-05	4.49E-05
30	1	121531223	3.35E-05	3.00E-05	4.58E-05
31	21	51617903	3.43E-05	3.07E-05	4.68E-05
32	8	87181631	3.45E-05	3.09E-05	4.70E-05

33	20	27217212	3.56E-05	3.19E-05	4.83E-05
34	7	38199052	3.84E-05	3.44E-05	5.19E-05
35	7	38243678	3.84E-05	3.44E-05	5.19E-05
36	7	39700987	3.84E-05	3.44E-05	5.19E-05
37	7	40454665	3.84E-05	3.44E-05	5.19E-05
38	8	89831586	4.09E-05	3.67E-05	5.50E-05
39	13	16327809	4.14E-05	3.72E-05	5.57E-05
40	5	97885183	4.49E-05	4.04E-05	6.00E-05
41	8	90545119	5.14E-05	4.63E-05	6.81E-05
42	2	1624948	5.41E-05	4.88E-05	7.14E-05
43	2	1674613	6.52E-05	5.89E-05	8.49E-05
44	17	34146408	6.59E-05	5.95E-05	8.57E-05
45	8	66103181	6.75E-05	6.09E-05	8.76E-05
46	8	90527471	6.78E-05	6.13E-05	8.80E-05
47	5	98020885	6.79E-05	6.13E-05	8.81E-05
48	19	47460400	7.29E-05	6.59E-05	9.41E-05
49	3	33622587	7.37E-05	6.66E-05	9.51E-05
50	17	26071186	7.45E-05	6.73E-05	9.60E-05

Table 2: Top 50 SNPs from GEMMA mixed model analysis in 542 individuals (no covariates). After pruning, analysis included 62,901 SNPs. Uncorrected p-values are presented for the Wald test, the Likelihood ratio test (lrt) and the Score test. CHR = chromosome; BP = base pair.

RANK	CHR	BP	p_wald	p_lrt	p_score
1	17	60554458	1.15E-11	1.03E-11	6.79E-11
2	1	18091577	2.93E-10	2.67E-10	1.11E-09
3	23	14650375	1.57E-09	1.44E-09	4.81E-09
4	6	7372690	1.62E-09	1.48E-09	4.94E-09
5	17	60503138	3.25E-08	3.02E-08	7.16E-08
6	17	60523882	3.25E-08	3.02E-08	7.16E-08
7	6	7000504	4.09E-08	3.81E-08	8.83E-08
8	17	40999944	4.88E-08	4.54E-08	1.03E-07
9	17	28460851	6.00E-08	5.60E-08	1.25E-07
10	1	155226154	8.67E-08	8.09E-08	1.74E-07
11	17	36291973	1.23E-07	1.15E-07	2.39E-07
12	17	41905502	2.65E-07	2.49E-07	4.81E-07
13	25	2021044	4.30E-07	4.04E-07	7.48E-07
14	9	76324169	5.87E-07	5.53E-07	9.95E-07
15	31	18200337	6.28E-07	5.91E-07	1.06E-06
16	19	24644810	6.80E-07	6.41E-07	1.14E-06
17	1	55259288	9.26E-07	8.74E-07	1.51E-06
18	31	18194086	9.31E-07	8.79E-07	1.52E-06
19	31	18163586	1.06E-06	9.98E-07	1.71E-06
20	31	18207378	1.06E-06	9.98E-07	1.71E-06
21	31	18263790	1.06E-06	9.98E-07	1.71E-06
22	2	19755735	1.21E-06	1.14E-06	1.94E-06
23	31	18083836	1.59E-06	1.51E-06	2.50E-06
24	1	70111131	1.68E-06	1.59E-06	2.62E-06
25	17	60468732	2.45E-06	2.32E-06	3.71E-06
26	17	60468135	2.56E-06	2.42E-06	3.87E-06
27	31	5160203	2.56E-06	2.43E-06	3.88E-06
28	1	18091709	2.87E-06	2.72E-06	4.31E-06
29	10	60356610	2.91E-06	2.76E-06	4.36E-06
30	31	5160132	3.00E-06	2.84E-06	4.48E-06
31	17	25963835	3.83E-06	3.64E-06	5.63E-06
32	16	28803066	3.97E-06	3.77E-06	5.82E-06

33	1	155323247	3.98E-06	3.78E-06	5.83E-06
34	2	17532819	4.14E-06	3.93E-06	6.05E-06
35	3	46116569	4.51E-06	4.29E-06	6.56E-06
36	19	30643317	4.80E-06	4.56E-06	6.95E-06
37	20	43319164	5.00E-06	4.75E-06	7.21E-06
38	31	18562346	5.72E-06	5.44E-06	8.18E-06
39	2	1673079	6.01E-06	5.72E-06	8.57E-06
40	2	1674613	6.01E-06	5.72E-06	8.57E-06
41	1	43852372	6.17E-06	5.88E-06	8.78E-06
42	31	17759802	6.28E-06	5.98E-06	8.92E-06
43	9	78229420	6.47E-06	6.16E-06	9.17E-06
44	1	39331337	6.80E-06	6.48E-06	9.61E-06
45	10	5679757	6.92E-06	6.59E-06	9.77E-06
46	25	16689693	6.95E-06	6.62E-06	9.81E-06
47	16	28745680	6.95E-06	6.62E-06	9.82E-06
48	31	1904633	6.98E-06	6.65E-06	9.85E-06
49	1	5649300	7.11E-06	6.77E-06	1.00E-05
50	2	1919026	7.12E-06	6.78E-06	1.00E-05

Table 3: Summary of 18 individuals selected for whole-genome sequencing (from a total cohort of 542 horses). Depth of coverage for whole-genome sequencing is indicated in the "coverage" column. M = mare; G = gelding; S = stallion.

	Gender	Sire	Coverage
Pacers	M	Western Ideal	12x
	M	Dragon Again	12x
	M	Somebeachsomewhere	12x
	S	Cam's Card Shark	6x
	S	Badlands Hanover	6x
	M	Yankee Cruiser	6x
	M	Somebeachsomewhere	6x
	S	Western Ideal	6x
	G	Allamerican Native	6x
Trotters	G	Andover Hall	12x
	M	SJs Caviar	12x
	S	Revenue S	12x
	M	Glidemaster	6x
	M	Cantab Hall	6x
	S	Cantab Hall	6x
	M	Credit Winner	6x
	S	Muscles Yankee	6x
	S	Windsong's Legacy	6x

Table 4: Summary of the six variants on ECA17 selected for follow-up with Sanger sequencing and/or RFLP. For each primer pair, the forward primer is listed first, then the reverse primer. BP = base pair; RFLP = restriction fragment length polymorphism enzyme.

Gene	BP	Variant type	Primers	RFLP
FAM124A	20049014	missense	5'-GGAGAAAATGGGGAAGATGC-3' 5'-CAGATCTGCAGCTGTTCAGG-3'	*NotI*
NAA16	28717149-28717150	missense (2)	5'-GAGCCACTGTTCTGCCTACC-3' 5'-TTTGCTGTTTGCCTTTTGTG-3'	n/a
EPSTI1	27382018	insertion	5'-GCATAACTTTTAGGGGAGGATAAG-3' 5'-CCGGATACACACCTTTCAGG-3'	n/a
VWA8	28468055	missense	5'-TCATGGTGCCACTTACATACG-3' 5'-TCAGTCCTGACAAGTCTTCACG-3'	*PstI*
ENS16793	32765728	missense	5'-ATCCAGCTGGTCGTCTTCC-3' 5'-TGTCAAATCAGAATCAAAGAATGG-3'	*BsaI*
PCDH9	41520282	missense	5'-GAGGGGCACCAACTTAAAGG-3' 5'-GATCTGGACCGAAAGACAGG-3'	*PsiI*

FAM124A: family with sequence similarity 124A
NAA16: N(alpha)-acetyltransferase 16, NatA auxiliary subunit
EPSTI1: epithelial stromal interaction 1 (breast)
VWA8: von Willebrand factor A domain containing 8 (also known as *KIAA0564*)
ENS16793: ENSECAG00000016793, uncharacterized protein-coding gene
PCDH9: protocadherin 9

Table 5: Summary of genotyping results for five segregating variants on ECA17 in a large population of horses.

	PCDH9 SNP1			
	G/G	**G/T**	**T/T**	**Total**
Trotters	14.8%	53.0%	32.2%	236
Pacers	71.1%	26.2%	2.7%	263
	ENS16793 SNP1			
	A/A	**A/G**	**G/G**	**Total**
Trotters	34.2%	49.8%	16.0%	219
Pacers	81.1%	17.7%	1.2%	249
	VWA8 SNP2			
Trotters	**T/T**	**T/G**	**G/G**	**Total**
Pacers	97.2%	2.8%	0.0%	211
	25.5%	54.9%	19.6%	255
	EPSTI1 INSERTION			
	G/G	**G/GA**	**GA/GA**	**Total**
Trotters	88.9%	11.1%	0.0%	234
Pacers	19.3%	50.2%	30.5%	259
	NAA16 SNPs1 & 2			
	AA/AA	**AA/CC**	**CC/CC**	**Total**
Trotters	93.5%	6.5%	0.0%	184
Pacers	22.5%	47.6%	30.0%	227

Table 6: Summary of 303 SNPs putatively associated with gait that were selected for inclusion in the Sequenom assay. The 190 SNPs selected from whole-genome sequencing were multiplexed with the AIMs into groups of 48 in six wells; the 113 SNPs selected from pooled sequencing were multiplexed into two wells with 48 samples each and a third well with 17 samples (well assignments not shown). ECA = equine chromosome; BP = base pair.

SNPs selected from whole-genome sequencing			SNPs selected from pooled sequencing	
ECA	**BP**		**ECA**	**BP**
1	5322242		1	38563255
1	5532596		1	38573734
1	5532686		1	38837069
1	5632465		1	106883816
1	17548101		1	106928205
1	17552161		3	2489598
1	17617018		3	2506253
1	17945265		3	2521561
1	17955548		3	52318025
1	18065598		4	8996717
1	18109069		4	9091283
1	35670985		4	9116614
1	35720250		5	55291788
1	35721326		5	55301141
1	35726345		5	55317127
1	35729338		5	55333664
1	35731283		5	61126642
1	35731849		5	61144680
1	38306816		5	61163899
1	38591441		5	66187039
1	38592096		5	66199885
1	38592542		5	66221515
1	38599532		6	81299480
1	38646291		6	81651604
1	38987953		6	81668230
1	38988003		9	29141611
1	39070730		9	29155106

1	39102187	[redacted]	9	29211591
1	39376366	[redacted]	11	29532466
1	39376415	[redacted]	11	29564206
1	39560619	[redacted]	11	29599837
1	39589062	[redacted]	11	31303355
1	39691699	[redacted]	11	31319004
1	39772282	[redacted]	11	31470618
1	42307163	[redacted]	11	36608682
1	43245721	[redacted]	11	36669528
1	43245806	[redacted]	11	36714823
1	48293517	[redacted]	11	36775489
1	48896092	[redacted]	11	36796850
1	49602985	[redacted]	12	16262259
1	50226814	[redacted]	12	16270318
1	50226817	[redacted]	12	16380392
1	55106029	[redacted]	12	16412374
2	17358298	[redacted]	14	1368081
2	17616769	[redacted]	14	1388861
2	17690098	[redacted]	14	1403046
2	18282822	[redacted]	14	1427118
2	18364832	[redacted]	14	1570169
2	18538576	[redacted]	14	5442438
2	18622200	[redacted]	15	10100242
2	18987527	[redacted]	15	10103035
2	19016749	[redacted]	15	10108201
2	19326982	[redacted]	16	59352686
2	19327652	[redacted]	16	59382124
2	19327821	[redacted]	16	59391893
2	19327827	[redacted]	17	50983052
2	19698739	[redacted]	17	51017875
2	19714056	[redacted]	17	51044268
2	19724672	[redacted]	17	61717590
2	19731591	[redacted]	17	61721785
2	19775173	[redacted]	17	61728019
3	2384676	[redacted]	17	61744016
3	2494992	[redacted]	17	61749334
3	3051017	[redacted]	17	65640738
3	3581003	[redacted]	17	65645147
3	3581548	[redacted]	17	65659694
3	3581676	[redacted]	20	25103556
3	3977000	[redacted]	20	25113918

3	46785415	[redacted]	20	25127274
3	46785561	[redacted]	20	27650699
3	46785621	[redacted]	20	27691110
3	47817508	[redacted]	20	27711111
3	48135062	[redacted]	20	27727105
3	48816809	[redacted]	20	27768899
3	49488838	[redacted]	20	46929785
3	49530033	[redacted]	20	47062579
3	49601762	[redacted]	20	47092658
3	49601886	[redacted]	23	14640812
3	49601896	[redacted]	23	14645077
3	49785110	[redacted]	23	14649864
3	49857337	[redacted]	23	20639151
3	49857369	[redacted]	23	20652865
3	49857478	[redacted]	23	20658789
3	52563310	[redacted]	23	20662320
3	52680823	[redacted]	24	6712987
3	53257281	[redacted]	24	6736928
3	53721793	[redacted]	24	10276151
3	53834511	[redacted]	24	10285906
3	54166034	[redacted]	24	10296168
3	56561263	[redacted]	24	10299566
3	56755586	[redacted]	25	3657454
3	57629621	[redacted]	25	3666056
3	57929520	[redacted]	25	3694284
3	58044431	[redacted]	25	3724550
3	58070704	[redacted]	25	3860478
3	58077312	[redacted]	25	11783623
3	58174699	[redacted]	25	11800074
3	58434545	[redacted]	25	15026761
3	58903953	[redacted]	25	15044553
3	76844764	[redacted]	29	3291497
3	76844777	[redacted]	29	3447596
3	77739534	[redacted]	29	3471572
6	6609510	[redacted]	29	10074576
6	7832752	[redacted]	29	10087117
6	7841413	[redacted]	29	10109015
6	7881374	[redacted]	30	14059751
9	75715699	[redacted]	30	14067984
9	75803120	[redacted]	30	14107178
9	75813719	[redacted]	30	14936139

9	75816548	[redacted]	30	14947553
9	75896366		30	15055793
11	46845778		30	15068782
11	47003926		30	15124747
11	47016874			
11	47016889			
11	48341945			
11	50634935			
11	50846315			
11	50918059			
11	52424214			
11	57422057			
11	58376457			
16	25722825			
16	28197056			
16	28198996			
16	28199539			
16	28786474			
16	29331872			
16	30331912			
17	27685585			
17	28054635			
17	28293289			
17	28347510			
17	28361747			
17	28458432			
17	28485796			
17	28540291			
17	28658850			
17	28658966			
17	28711958			
17	29271555			
17	29274637			
17	39403989			
17	39404292			
17	60460198			
18	77603381			
19	21446218			
19	31393832			
19	37986794			
23	14182456			

23	14211839	[redacted]	
23	14648590		
23	14714942		
23	14813929		
23	14814635		
23	14980071		
25	11689674		
25	11691308		
25	11793191		
25	11811829		
25	12758770		
25	12791659		
25	13016645		
25	13021802		
25	13021844		
25	13031250		
25	13031714		
25	13036645		
25	13047191		
25	13050176		
25	13052616		
25	14418173		
25	14531198		
25	14557888		
25	14735220		
25	14737344		
25	14760167		
25	15576483		
25	15621763		
25	15829342		
25	15839070		
25	15845420		
25	16817685		
25	16820893		
26	3315794		
26	3315939		
26	3315959		
26	3315992		
26	3641990		
26	3642510		

Table 7: Comparison of the top 25 association results from GEMMA (corrected for population structure and relatedness) based on Sequenom genotyping data in 500 horses. Reported p-values are based on the Wald, likelihood ratio (lrt), and score tests. ECA = equine chromosome; BP = base pair.

ECA	BP	p_wald	p_lrt	p_score
30	14067984	3.35E-32	1.71E-31	1.22E-24
30	14947553	1.25E-30	4.98E-30	9.05E-24
23	14649864	3.20E-30	1.27E-29	1.64E-23
23	14648590	9.36E-30	4.30E-29	3.80E-23
30	15055793	1.27E-29	5.68E-29	4.47E-23
23	14640812	1.39E-29	5.87E-29	4.45E-23
30	15068782	9.73E-29	3.63E-28	1.37E-22
17	28540291	4.80E-28	1.13E-27	2.30E-22
30	14936139	6.40E-28	1.85E-27	3.51E-22
1	35729338	2.29E-27	5.32E-27	6.39E-22
1	35731849	2.66E-27	6.22E-27	7.12E-22
1	35731283	2.88E-27	6.63E-27	7.36E-22
1	17945265	1.48E-27	1.07E-26	1.72E-21
1	35720250	7.74E-27	1.81E-26	1.45E-21
17	28458432	1.57E-25	1.78E-25	4.50E-21
1	35726345	9.29E-26	2.07E-25	7.19E-21
1	35721326	2.54E-24	6.44E-24	8.07E-20
30	15124747	6.38E-24	8.30E-24	6.61E-20
1	38573734	1.65E-23	3.75E-23	2.59E-19
17	28361747	4.90E-23	6.30E-23	2.70E-19
1	38591441	3.46E-23	7.70E-23	4.26E-19
1	38592542	2.30E-22	5.15E-22	1.64E-18
23	20662320	2.40E-22	6.18E-22	1.99E-18
23	20652865	3.59E-22	8.66E-22	2.45E-18
3	3051017	3.65E-21	4.27E-21	5.05E-18

Figure 1: Schematic of a hypothetical polygenic model for expression of alternative gait across breeds. In this model, there are 7 modifying alleles that interact with *DMRT3* to produce different gaits in various breeds. Breeds that share the same alternative gait also share modifying alleles. These modifying alleles may be present in non-gaited breeds, but in the absence of *DMRT3*, no alternative gait is exhibited. *DMRT3* is thus necessary, but not sufficient for alternative gaits.

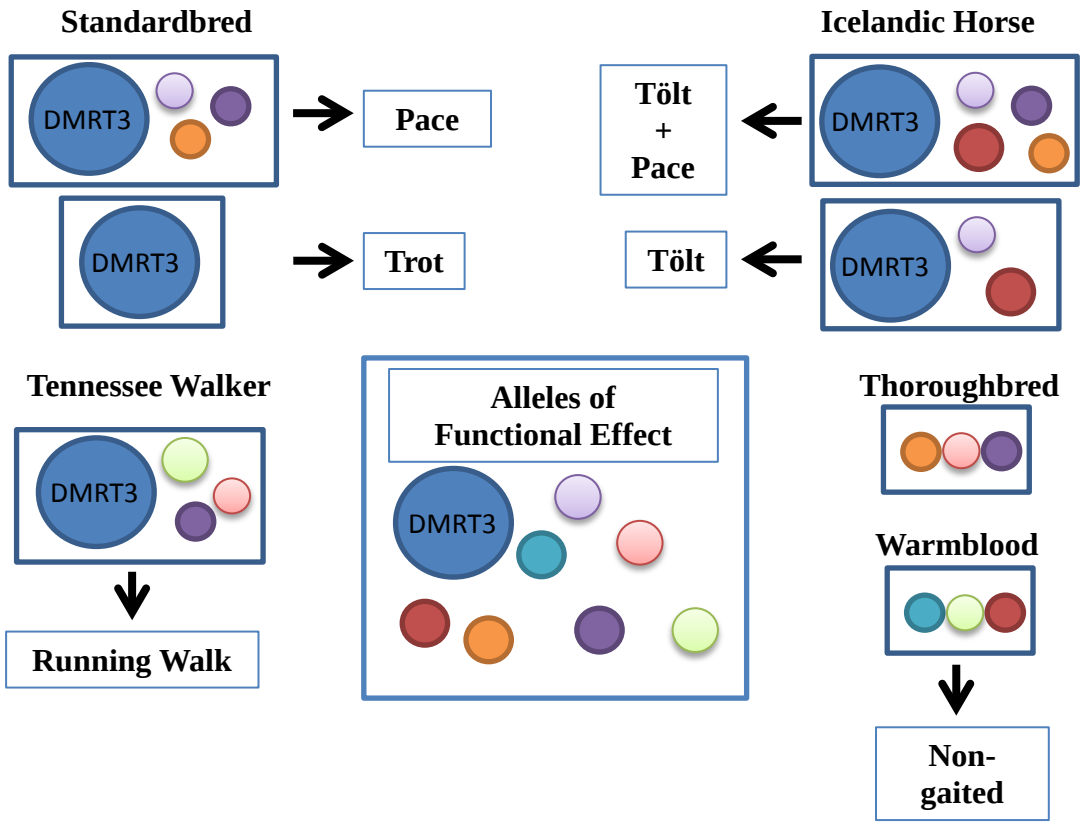


Figure 2: Multidimensional scaling plot of 542 Standardbreds based on genome-wide genotyping data. The horses cluster distinctly by gait, with the pacers more similar to pacers than trotters to trotters.

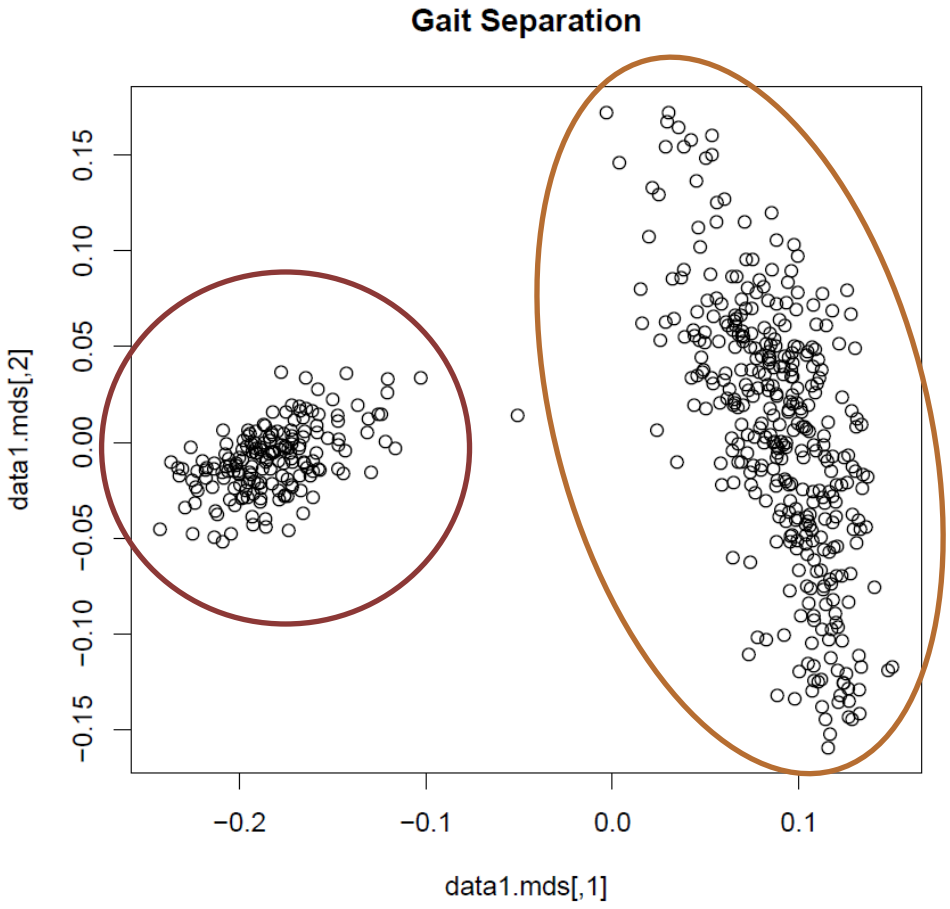


Figure 3: Manhattan plot of results from mixed model analysis using GEMMA in a population of 374 pacers and trotters (gender and origin covariates). The 31 autosomal and X chromosome (32) are represented in different colors along the x-axis and the –log(p-value) of the likelihood ratio test is on the y-axis. Each colored dot represents a SNP. The blue line marks a p-value of 5 x 10^{-5} while the red line marks a p-value of 5 x 10^{-7}. SNPs falling above the red line are considered genome-wide significant Top hits are on ECA13 and 17.

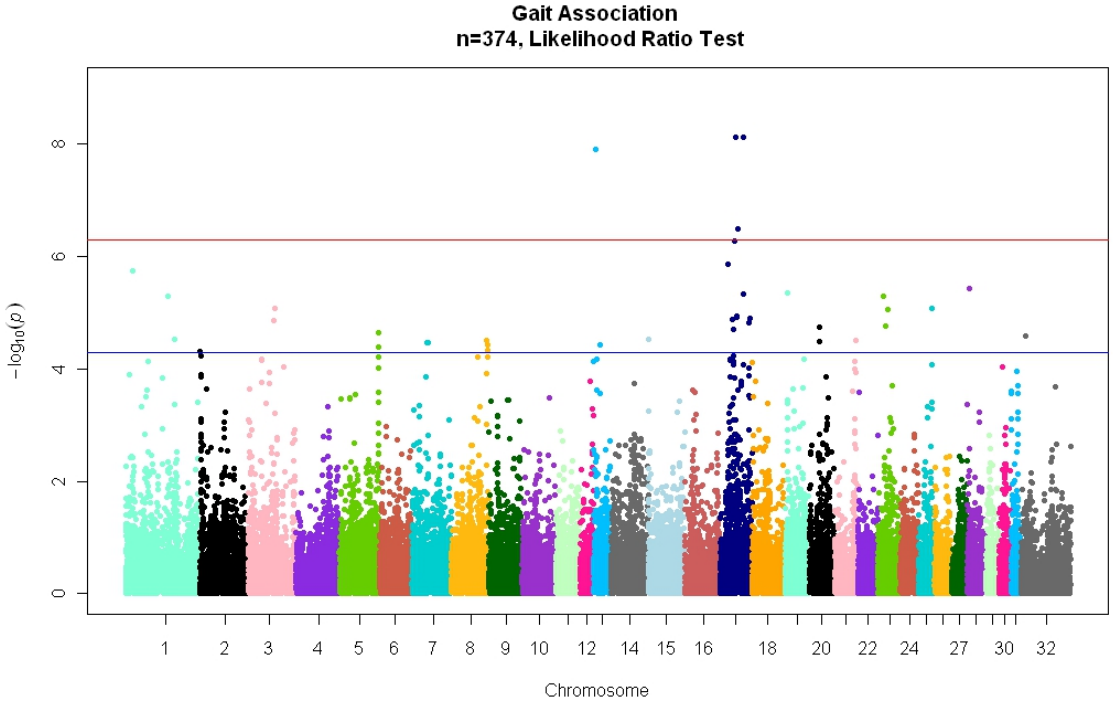
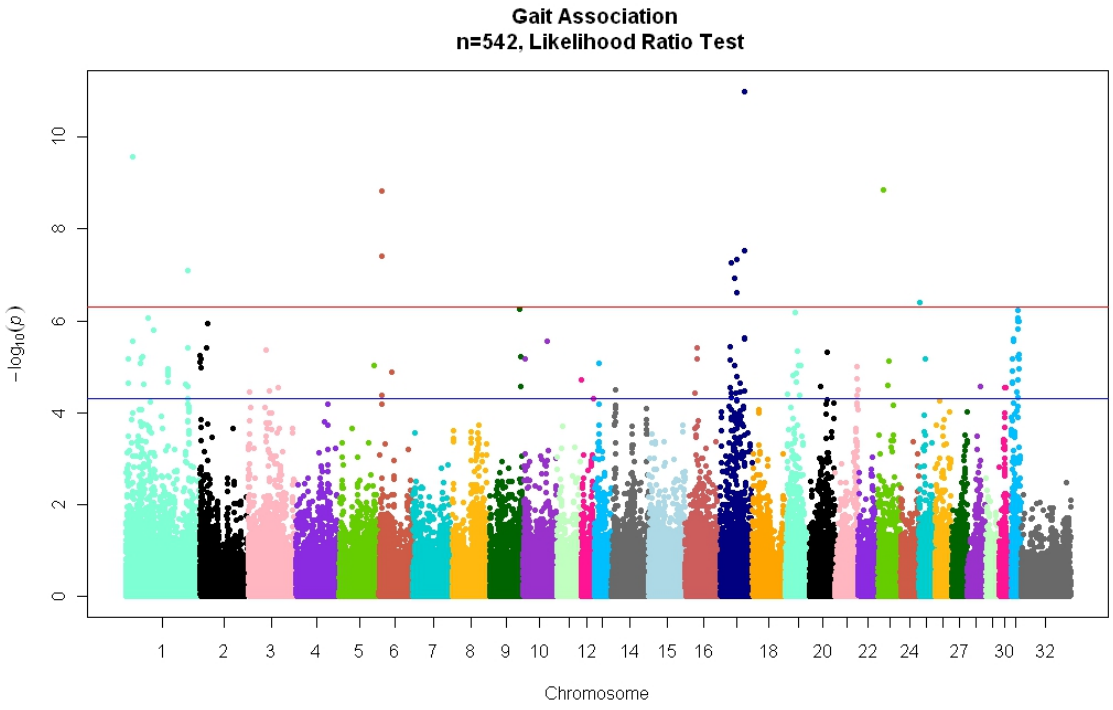


Figure 4: Manhattan plot of results from mixed model analysis using GEMMA in a population of 542 pacers and trotters (no covariates). See Figure 3 for compete legend. Genome-wide significant hits ($p < 5 \times 10^{-7}$) are on ECA1, 6, 17, 25 and 26.



Chapter 8

Investigation of Genetic Determinants of Performance in the Standardbred Horse

Investigation of Genetic Determinants of Performance in the Standardbred Horse

Annette M. McCoy

From the Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, St Paul, MN 55108, USA

Sources of Funding: Funding provided by USDA-NIFA-AFRI and the United States Equestrian Federation, Inc. Dr. McCoy was funded by an institutional NIH T32 Comparative Medicine and Pathology Training Grant (University of Minnesota) and a Doctoral Dissertation Fellowship (University of Minnesota); partial funding for Dr. McCue was provided by NIH NIAMS 1K08AR055713-01A2.

Summary

Horses have been selectively bred to perform specific tasks for thousands of years, and while the definition of what constitutes “good” performance varies by breed and discipline, there is no question that performance is one of the most important factors in judging the potential breeding value of an individual horse. However, while it is widely accepted that performance characteristics are inherited, the specific genetic factors underlying athletic performance in the horse are largely unknown.

Among racing breeds, the performance characteristics that are most valued are number and quality of wins, and total earnings. However, these measures can be influenced by a large number of outside factors and so are less than ideal phenotypes for investigating genetic factors underlying race performance. Although physiologic measurements related to cardiovascular capacity and musculoskeletal characteristics would likely be the best phenotypes upon to which to base these investigations, they are difficult to obtain from a large population of horses. Speed is an essential determinant of wins and earnings, and an individual’s maximum speed is determined by their intrinsic physiologic characteristics. Fastest recorded race speed, although still influenced to a degree by outside factors, is thus a better measure of intrinsic athletic ability than other easily obtained measures, and was selected as the primary phenotype of interest for this study.

A genome-wide association study (GWAS) was performed in a large group of Standardbred horses with fastest race times available in their public records (n = 94 in the original cohort, n = 414 in the final cohort). GWA analysis was performed using GEMMA software. After accounting for population structure and relatedness, 4 SNPs on

4 chromosomes (ECA2, 7, 11, 22) were moderately associated with speed ($p < 5 \times 10^{-5}$) when gender, gait, origin, and age at fastest time were included as covariates in the mixed model. An additional GWAS was performed using linear regression analysis in PLINK in a subset of the population (n = 208) using optimal distance (i.e. the distance at which the horse's fastest time was recorded) as the outcome variable. Five SNPs on three chromosome (ECA3, 14, 23) were moderately associated with optimal distance. Although several plausible candidate genes for performance and optimal distance were identified within the GWAS regions of interest, these findings should be validated in one or more independent populations. Alternative phenotypes may also need to be considered given the particularly complex nature of "performance" as a trait.

Introduction

Performance in horses is evaluated in a myriad of ways, depending on breed, discipline, and the purpose for which assessment is being conducted. While performance as defined by wins and earnings unquestionably drives breeding and buying decisions in the race industry, the most common use of racing performance measures in the scientific literature is to evaluate the detrimental effects of a disease or condition[e.g.282;283], or, more commonly, to demonstrate the effectiveness of a medical or surgical intervention.[e.g.22;284-290] However, the lack of a single accepted measure of performance in racehorses makes it difficult to compare results between published studies. Single parameters considered often include number of races, earnings (cumulative or per start), finish position, and number of top three finishes.[22;285;288;290] A variety of performance indices (PI) have also been proposed, which assign point values to various parameters, such as order of finish and earnings, and combine them together or transform them in some way.[22;282;285-287;289] The advantage to PIs is that they purportedly allow for comparison across different classes of performance.[289] A complex statistical model has been developed[291] and validated[292] in Thoroughbreds that accounts for a number of environmental factors including track, track conditions, race distance, purse size, age, time of year, placement of the race in the schedule (i.e. time of day), number of horses in the race, post position, and weight carried by the horse, and this has been used to report outcome in at least one study.[288] However, this equation has limited utility outside of the region in which it was developed because all of the data upon which it was based came from only five tracks.[291] Models have also been developed by racing boards to facilitate comparison of

performance across race seasons, ages, and genders, but these are not widely utilized in the literature.[283;285;287;293]

Although it is commonly accepted that performance characteristics are inherited – indeed, this is the basis of selective breeding in the horse and other agricultural species – the genetic factors that contribute to "performance" are largely unknown. Given the complexity of merely defining performance, as illustrated above, this is perhaps not surprising. Breeders of production species, such as cattle and pigs, have long used complex breeding values based on progeny testing to make decisions about which animals to keep and which to cull, although recently, genomic-based selection has become the standard.[294] In contrast, in horses, the use of genetic evaluation in breeding decisions tends to be highly breed-specific (and discipline-specific) and is generally based more on the performance and conformation characteristics of an individual than of their progeny.[263;295-297] Widespread genomic selection has not become the norm in this species, where there is still as much "art" as science in the quest to breed elite performers.

There is not generally standardized testing of young stock in the racing breeds as is seen in sport horses. For these horses, pedigree and the number and quality of wins are more likely to make an individual desirable for breeding. A standardized genetic merit evaluation based on a combination of race variables, including number of starts, number of top 3 finishes (normalized for number of starts), earnings, earnings per start, and fastest race time was evaluated in Swedish trotters[12], but has not been widely accepted. The reported heritability of racing traits varies widely, which makes the utility of a universal breeding value questionable. For example, heritability of speed in North American Standardbreds (trotters and pacers pooled) was reported to be 0.29.[14] Similarly,

heritability of speed in Finnish Standardbred trotters was reported to be 0.28[13], but in German trotters it ranged from 0.01-0.18 when age was taken into account.[15] In Quarter Horses, heritability of speed ranged from 0.26 to 0.41 and increased with the distance over which individuals were raced (301m-402m).[298] Conversely, in Thoroughbreds, heritability of speed decreased from 0.29 in horses racing 1000m to only 0.05 in horses racing 1600m.[299] The influence of age and race distance on the heritability estimates reported here emphasizes the point that even on an optimal genetic background, external influences will always play a role in the success of an individual. That being said, better understanding of the genes and pathways that underlie performance traits can only improve genetic selection schemes. The interest of the racing community in such information is aptly illustrated by the popularity of the commercial test for the so-called "speed gene," which genotypes individuals for an intronic variant in the myostatin gene (*MSTN*) that has been reported to be highly associated with optimal race distance in Thoroughbred horses.[79]

The primary purpose of this study was to identify chromosomal regions associated with performance in Standardbred racehorses as a first step in investigating specific genetic determinants of performance in this breed. For the purposes of this work, "performance" was primarily defined as fastest speed over a mile. While, as discussed above, this may be considered an oversimplification of performance, it is one of the most basic and essential determinants of success in a racing individual and was therefore considered to be an appropriate starting point for investigation of this particularly complex trait.

Materials and Methods

Horses and Performance Records: The initial study cohort was comprised of 94 Standardbred yearlings born in 2007 and raised on a single breeding farm in the eastern United States. There were 48 pacers and 46 trotters in this group, belonging to distinctly bred families with little crossover between lines, although all were related within thirteen generations to a single breed foundation sire. Management practices, including diet and exercise regimen, were the same for all foals during their first year of life. As yearlings (14-20 months of age), horses were sold, at which time they were lost to direct follow-up. Although the primary trait of interest in these individuals at the time of sample collection was osteochondrosis (OC; see **Chapter 4**), they were leveraged for investigation into genetic factors underlying performance as well. OC status was shown to not affect race performance in this cohort using a variety of outcome measures (see **Chapter 3**), so for the purposes of the present study, OC was not included as a covariate in any analysis. All available performance data was obtained from the United States Trotting Association.

While preliminary results from the initial study cohort were interesting, they were based on small numbers of horses and could have been biased by the close relationships between individuals. Thus, an expanded study cohort was assembled, consisting of 414 Standardbred horses from North America (n = 201) and Europe (n = 213). There were 113 pacers and 301 trotters in this group. Consistent with the original cohort, the pacers and trotters were members of distinct families. For the North American horses, all available performance records were obtained from the United States Trotting Association. For the European horses (and the small number of horses bred in the U.S.

that had been exported to Europe), records were collected from the appropriate country's trotting association.

Performance records obtained for every individual included fastest career time and age at fastest time. As needed, fastest time was converted from time over 1 km to time over 1 mile so direct comparisons could be made. Only times recorded during auto start races were considered for European racers (not volte start). For horses that raced in Europe, the distance at which their fastest time was achieved was also recorded.

Whole-Genome Genotyping: All horses utilized in this study cohort had been sampled during the course of unrelated projects in the Equine Genetics and Genomics Laboratory and had already been genotyped either at 54,602 SNPs using the first generation Illumina Equine SNP50 chip (n = 306) or at 65,157 SNP markers using the second generation Illumina Equine SNP70 chip (n = 68).

Genotype Imputation: The two equine genotyping platforms share only 45,703 SNPs. To avoid loss of information from the non-overlapping markers, genotype imputation was utilized. Imputation statistically estimates genotypes from non-assayed SNPs by comparing haplotype blocks in the study population with haplotype blocks in a more densely genotyped reference population. As part of the validated pipeline detailed in **Chapter 5**, BEAGLE[211] was utilized to impute the ~18,000 markers unique to the SNP70 chip in those horses genotyped on the SNP50 chip, and to impute the ~9,000 markers unique to the SNP50 chip in those horses genotyped on the SNP70 chip. Imputed files were merged using the --merge command in PLINK[212] prior to analysis.

Genome-Wide Association (GWA) Analysis: Initial GWA analysis was performed in the original study cohort (n = 94) using a linear regression model in PLINK (--linear)

with fastest speed in seconds as the continuous outcome variable. Gait and gender were included as covariates in the model. To control for multiple testing, 10,000 t-max label-swapping permutations were applied (--mperm 10000). This analysis was repeated in pacers only and in trotters only; no covariates were applied in these models.

The GWA in the final cohort (n = 414) was carried out after imputation using GEMMA (Genome-wide Efficient Mixed Model Analysis) software.[214] GEMMA is computationally efficient for large data sets, accounts for population structure and relatedness using a marker-based relationship matrix, and calculates variation for individual SNPs rather than an average variation across all SNPs. These features make this software particularly appropriate for the present application. The GWA was performed using the options to create a centered relatedness matrix (-gk 2) and perform all three possible frequentist tests: Wald, likelihood ratio, and score (-fa 4). The relatedness matrix was constructed using a linkage-disequilibrium (LD)-pruned set of markers (100 SNP windows, sliding by 25 SNPs along the genome, pruned at $r^2 > 0.2$; PLINK command --indep-pairwise 100 25 0.2).[59] A covariate file including gender, gait, and origin (North America or Europe), +/- age at fastest time and distance at which the fastest time was achieved was incorporated into the mixed model [-c] for the entire group. When pacers were considered separately, gender, origin, and age were included as covariates in the model. When trotters were considered separately, distance at which the fastest time was achieved was also added as a covariate. SNPs were pruned prior to GWA using the default GEMMA parameters of minor allele frequency (MAF) <1% and missingness <95%.

Since European trotters race over different distances (in contrast to North American horses, which are all raced over 1 mile), a GWA was performed in this subset of individuals (n = 208) with the distance at which an individual achieved their fastest time as the outcome of interest. A simple linear regression model was used in PLINK with age at fastest time and gender as covariates included in the model. SNPs were pruned for MAF <1% and missingness <90%.

Association plots were generated using the base graphics package in the R statistical computing environment.[187] Based on previously published guidelines, uncorrected p-values of less than 5×10^{-7} were considered to indicate genome-wide significant association, while uncorrected p-values between 5×10^{-5} and 5×10^{-7} were considered to indicate moderate association.[67] When permutations were applied, a p-value of <0.05 was considered to be genome-wide significant.

Results

GWA Results for Original Study Cohort: 54,602 SNPs were considered during the analysis in PLINK. Sixty-six of 94 horses had a fastest time recorded in USTA records and were included in the analysis. After 10,000 label-swapping permutations, there were no SNPs reaching statistical significance. The most significant SNP was located on ECA1 (chr1.120115552; p = 0.021); three other less significant SNPs were located nearby (chr1.118806817, chr1.125287062, and chr1.127072698; p = 0.47). The ~9Mb region defined by these 4 SNPs was quite gene-rich, containing 88 named genes, 12 novel protein-coding genes, and numerous predicted pseudogenes and noncoding RNAs. The second-most significant SNP was located on ECA11 (chr11.5367142; p = 0.27)),

~22kb upstream of *MGAT5B* (mannosyl(alpha-1,6)-glycoprotein beta-1,6-N-acetylglucosaminyltransferase, isozyme B).

Forty of 48 pacers had a recorded fastest time and were included in the gait-specific analysis. After 10,000 label-swapping permutations, there were no SNPs reaching statistical significance. The top two SNPs were the same as in the combined analysis (chr1.120115552, chr11.5367142), but with less significant p-values (0.16 and 0.31, respectively). Twenty-six of 46 trotters had a recorded fastest time and were included in the gait-specific analysis. As above, there were no SNPs reaching statistical significance after permutations. The top SNP in this analysis was located on ECA13 (chr13.27836637; p = 0.49), followed by a cluster of 3 SNPs on ECA20 (chr20.14408119, chr20.14481448, chr20.14509997; p = 0.70). The ECA13 SNP was located within intron 3 of *COQ7* (coenzyme Q7 homolog, ubiquinone [yeast]). The region demarcated by the ECA20 SNPs is ~300kb from the nearest gene, *JARID2* (jumonji, AT rich interactive domain 2).

GWA Results for Final Study Cohort: Among the final study cohort (n = 414), 306 were genotyped on the Equine SNP50 chip and 68 were genotyped on the Equine SNP70 chip. Genotype imputation was performed in order to combine these data without the loss of marker information (see Materials and Methods). Imputation was successfully carried out in the final study cohort using the pipeline described above and in **Chapter 5**. After imputation, there were 74,595 markers in the complete data set, an increase of nearly 29,000 markers over the shared set. SNP pruning for MAF and genotyping success was performed during analysis in GEMMA and PLINK as described above. After pruning, 62,795 SNPs were available for analysis in the entire final study cohort of 414

horses (pacers and trotters combined) in GEMMA. The addition of age at which the fastest time was achieved as a covariate did not substantively change the mixed model results over gender, gait, and origin alone, however, the top SNPs were slightly different when age and distance at which the fastest time was achieved were both included. In each model, only 3-4 SNPs showed moderate evidence of association with speed ($p \leq 5 \times 10^{-5}$ as determined by the likelihood ratio test) (**Figure 1**). A comparison of the top 25 results from each analysis is shown in **Table 1**. Moderately associated SNPs were located on ECA2, 7, 11, and 22 when gender, gait, and origin, +/- age were considered as covariates. None of these SNPs were located particularly close to any protein coding genes (chr2.109162364: ~62kb upstream of *NDST3* [N-deacetylase/N-sulfotransferase (heparin glucosaminyl) 3], chr7.94667256: ~30kb downstream of *METTL15* [methyltransferase like 15], chr11.4759103: ~160kb downstream of *SEPT9* [septin 9], chr22.43741335: ~30kb downstream of *BMP7* [bone morphogenetic protein 7]). In addition to the same ECA11 SNP as above, markers on ECA15 and 17 were moderately significant when distance was added as a covariate. The SNP on ECA15 (chr15.52466022) was ~6kb upstream of *TMEM247* (transmembrane protein 247); the SNP on ECA17 (chr17.5851284) was ~3kb upstream of *SHISA2* (shisa family member 2).

After pruning, 59,003 SNPs were available for analysis in 113 pacers. Fifteen SNPs on seven chromosomes were moderately associated with speed ($p \leq 5 \times 10^{-5}$ as determined by the likelihood ratio test). None of these SNPs were the same as the moderately associated markers from the combined analysis. Five SNPs demarcated a region of interest on ECA15 from 10.4-19.6Mb (two clustered on one end, three on the other), while four SNPs on ECA21 were clustered between 46.3-46.5Mb. The remaining

six SNPs were singletons (ECA1, 2, 11, 16 and two distantly located SNPs on ECA26). The top 25 SNPs from this analysis are shown in **Table 2**. The region on ECA15 is extremely gene rich; two of the associated SNPs are located within genes (chr15.1189429 in intron 2 of a novel gene; chr15.1961695 in intron 58 of *DNAH6* [dynein, axonemal, heavy chain 6]). One of the ECA21 SNPs (chr21.46337201) is within intron 75 of *DNAH5* (dynein, axonemal heavy chain 5), while the rest are clustered ~200kb downstream of this gene. Of the singleton SNPs, only two are within 10kb of named genes. The SNP on ECA2 (chr2.25314465) is ~3kb downstream of *SNRNP40* (small nuclear ribonucleoprotein 40kDa [U5]), while the SNP on ECA11 (chr11.5654983) is ~5kb downstream of *CYGB* (cytoglobin). This latter SNP is the fourth highest hit in the combined analysis when age and distance are included as covariates.

After pruning, 61,114 SNPs were available for analysis in 301 trotters. When age, gender, origin, and age at fastest time were considered as covariates, only one SNP on ECA22 was moderately associated with speed ($p \leq 5 \times 10^{-5}$ as determined by the likelihood ratio test). This SNP was identical to the one that was moderately significant in the combined analysis (above). When distance at which the fastest speed was achieved was added as a covariate, four SNPs were moderately associated with speed, including the SNP on ECA22 from the simpler model. Significantly associated SNPs on ECA2 and 15 were identical to those seen in the combined analysis (above). The SNP on ECA 10 (chr10.2162356) was only 631bp downstream of *ZNF536* (zinc finger protein 536) and likely lies within the 3'UTR (untranslated region) for this gene.

The only phenotype other than speed that was considered as an outcome measure in this study was "optimum distance" – that is, the distance at which the fastest recorded

time for an individual was achieved. Since North American horses are all raced over a mile, they were excluded from this analysis. Two hundred thirteen trotters racing in Europe were included in this GWA. Distance was evaluated as a continuous phenotype measured in meters, ranging from 1600m to 2640m. The two most common distances were 1609m (1 mile) and 1640m, with nearly twice as many horses recording their fastest time at these distances than the next most common distance. After pruning, 62,177 SNPs were available for analysis in 208 individuals in PLINK. Five SNPs were moderately associated with optimal distance under the additive logistic regression model (uncorrected p-value $p \leq 5 \times 10^{-5}$), although there were several others nearly as significant (**Figure 2**). The top 25 SNPs from this analysis are reported in **Table 4**. Of the five moderately associated SNPs, only one, chr14.25019300, is located within or near a named gene (*GRIA2* [glutamate receptor, ionotropic, AMPA2]). This SNP is one of 7 markers defining a region of interest from ~24.5-25.5Mb on ECA14. In addition to *GRIA2*, there are two other named genes within this region, *MFAP3* (microfibrillar-associated protein 3) and *FAM114A2* (family with sequence similarity, member A2), as well as a novel protein-coding transcript and three pseudogenes.

Discussion

Determination of an accurate phenotype is one of the most important aspects of designing a genome-wide association analysis so that misclassification bias can be avoided.[55] This becomes especially challenging with a complex trait such as performance, in which there is no single best phenotype to evaluate. Although outcomes such as number of wins, order of finish, and winnings are widely used as measures of

success for racehorses after an intervention/treatment[22;284-290], their appropriateness for use as phenotypes when investigating genetic factors involved in performance is questionable because they are so heavily influenced by external factors. Speed (i.e. as measured by fastest time over a set distance) can be considered a more "intrinsic" measure of performance ability, but even this can be affected by race distance, track conditions, and other factors.[291] From a physiological perspective, the inherited aspects of performance are likely related to a combination of conformation, muscle structure (i.e. fiber type distribution), and cardiovascular capacity. Objective measures of these parameters would unquestionably provide a more appropriate alternative to the more commonly used "performance" phenotypes. However, these measures are difficult to obtain from a population large enough to conduct an appropriately powered GWAS. Thus, for this study, fastest recorded speed over a mile was selected as the primary phenotype of interest as a starting point for investigating genetic factors underlying performance in Standardbred racehorses.

Initial GWA analysis was performed in a study population made up of highly related individuals from a single breeding farm and was heavily biased towards pacers. The top SNP on ECA1 was clearly driven by the pacers, as it was not among the top results when trotters were evaluated separately. Interestingly, haplotype analysis in the pacers revealed a region of interest ECA1 spanning from ~120-122Mb that was shared by racing Quarter Horses and Thoroughbreds (J. Petersen, personal communication). Within this region were 13 named genes, of which two (*THSD4* [thrombospondin, type I, domain containing 4] and *PKM* [pyruvate kinase, muscle]) were considered plausible candidates for playing a role in performance. *PKM* was selected for further follow-up in the

Standardbred population and a small deletion located 12bp after the end of exon 4 (potentially affecting splicing) was discovered via Sanger sequencing. However, genotyping results in 515 Standardbreds failed to demonstrate a significant association between this variant and fastest time. It is possible that this variant would have been significantly associated with a different phenotype, but since 96% of pacers and 88% of trotters carried at least one copy of the deletion, it is more likely that this was not a functional mutation differentiating good performers from poor performers.

Expansion of the GWAS population to 414 horses markedly changed the SNPs found to be associated with speed when compared to the original GWAS. This population was biased towards trotters, and the top SNPs in the combined analysis were the same as those found when trotters were evaluated alone, while the top SNPs in pacers alone were completely different. This raises the possibility that there may be different genetic factors underlying speed in pacers and trotters. Given the distinct biomechanical differences between gaits[197-200], this is not an unreasonable supposition.

A few of the genes in which associated SNPs from the final GWAS were located could be considered plausible candidate genes for playing a role in performance based on their known physiologic functions. *NDST3*, for example, is important for heparin sulfate sulfation in the brain and is a target of *RUNX2* (runt-related transcription factor 2) in osteoblasts[300;301], and could thus play a role in performance via effects in the nervous system or skeletogenesis (i.e. conformation). *DNAH5* and *DNAH6*, which contained associated SNPs in pacers only, are paralogous genes that play an important role in ciliary structure and function in the respiratory tract.[302] Mutations in *DNAH5* account for 15% of human primary ciliary dyskinesia, which results in abnormal clearance of mucous

and other substances from the lungs[302]; the potential negative impact of suboptimal respiratory tract clearance on performance is not hard to imagine. Finally, *CYGB* (also associated only in pacers) has multiple important physiologic roles, including muscle repair and regeneration[303], modulation of cardiovascular and respiratory reflexes[304], and protection from hypoxic injury[305], any of which could impact performance.

Our GWAS results when "optimal distance" was used as the phenotype of interest do not support the variant in *MSTN* previously reported in Thoroughbreds.[79;306] It is possible that this is because the majority of the horses in our cohort were raced at the same distance (approximately 1 mile), with relatively few horses raced at longer distances. It is unlikely that the link between *MSTN* and muscle fiber type (predicted to be the underlying physiologic explanation for the role of *MSTN* on this performance trait)[62] would differ between Thoroughbreds and Standardbreds. It is interesting to note that one of the moderately associated SNPs in the GWAS reported here was in *GRIA2*, which plays a crucial role in excitatory neurotransmission.[307] A link between neurotransmission and the propensity to be a "sprinter" or "stayer" is not inconceivable.

Limitations: As discussed above, the primary limitation of this study is likely the choice of phenotype used to measure performance. Fastest recorded speed was chosen as the phenotype of interest here because it is easily measured, readily available in public records, and should be a reflection of intrinsic athletic ability. However, a study of soccer players demonstrated that individuals did not exhibit their maximal sprint speed during match play, and in fact, that faster players attained a lower percentage of their maximal speed than did slower players.[308] This was speculated to be due, in part, to the tactical demands of the game, and it is not unreasonable to suppose that a similar situation could

arise among a field of horses under race conditions. It has been suggested that actual racing time for an individual should be adjusted based on the time of the winning horse for that race as a more objective measure of race performance that accounts for track conditions, class of race, and other factors.[14] This adjustment would be possible based on publically available data, and could be considered in the future.

In studies of human athletic performance, fine motor skills are evaluated on a sport-specific basis (i.e. passing accuracy and dribbling speed for soccer players[309]), but measures of speed, power, and agility tend to be consistent across disciplines.[309-311] Objective physiologic parameters, such as VO_{2max} (maximal oxygen uptake), velocity at lactate threshold, maximal heart rate, and heart rate variability are also widely utilized as measures of performance.[311-313] Similar measurements can be made in horses, but they generally require access to specialized equipment including a dry-land treadmill, and are therefore not done on a routine basis although published reports suggest that they may be predictive of performance.[314-316] For example, the velocity resulting in a blood lactate concentration of 4mmol/L was reportedly a predictor of "good" performance in Standardbred trotters, but it also varied by age and track.[316] It has been suggested that combined analysis over multiple measures of performance (speed, power, agility, and fine motor skills) might be a more accurate way to assess an individual's underlying athletic ability than any single measure alone.[309] This idea is already embraced in the equine industry, at least when making decisions about the breeding value of an individual[263;295-297], and it is possible that utilizing a "combined" phenotype may be a better approach to investigating the genetic factors underlying performance than any single phenotype alone.

Future Directions: Although there are limitations to the phenotype investigated here, the next step based on the current GWAS findings, or those from future work using alternative phenotypes, is investigation of specific putatively functional variants that may play a role in performance. As discussed extensively in **Chapter 6**, there are many challenges inherent in a traditional candidate gene approach for variant discovery, and this is illustrated in the present study by the results from *PKM* genotyping. A more efficient alternative approach would be to use whole-genome sequencing for variant discovery within chromosomal regions of interest identified from the GWAS, and then to perform additional fine-mapping to hone in on specific variants of interest. This approach is described in **Chapters 6** and **7**, and the whole-genome sequencing data set used in those studies could be applied here as well since the sequenced individuals were part of the final performance study cohort.

As for any GWAS, it will be important to replicate the findings reported here in an independent population. An appropriate initial validation population would consist of Standardbred trotters and pacers, but it would also be interesting to see if the findings are consistent across racing breeds (i.e. populations of Thoroughbred and Quarter Horse racehorses). Genes involved in musculoskeletal development and cardiovascular capacity are likely to play a role in performance across breeds, but it is possible that specific variants of functional effect may be different. To further investigate the "optimal distance" phenotype, a larger population of trotters racing at longer distances could be added to the existing cohort; however, since all North American Standardbreds are raced over a distance of 1 mile, this phenotype is of less interest to the breed as a whole than to other racing breeds such as the Thoroughbred.

Table 1: Top 25 SNPs from GEMMA mixed model analysis in 414 individuals. All three models included gender, gait, and origin as covariates. After pruning, analysis included 62,795 SNPs. Uncorrected p-values are shown for the Likelihood ratio test only. CHR = chromosome; BP = base pair.

Gender, Gait, Origin covariates			+ Age at fastest time			+ Age and Distance at fastest time		
CHR	**BP**	**P**	**CHR**	**BP**	**P**	**CHR**	**BP**	**P**
2	109162364	2.85E-06	**11**	4759103	4.78E-06	**15**	52466022	1.60E-05
11	4759103	1.58E-05	**2**	109162364	1.69E-05	**11**	4759103	2.89E-05
7	94667256	2.43E-05	**22**	43741335	2.26E-05	**17**	5851284	3.58E-05
22	47953960	3.91E-05	**7**	94667256	4.24E-05	**11**	5654983	5.71E-05
22	43741335	5.27E-05	**15**	52466022	8.51E-05	**21**	46337210	9.87E-05
11	7271702	7.65E-05	**21**	46337210	9.16E-05	**7**	23509046	1.24E-04
21	46337210	8.01E-05	**21**	46521182	1.03E-04	**26**	18163019	1.25E-04
22	48130443	9.78E-05	**21**	46521240	1.03E-04	**21**	46521182	1.25E-04
22	21499797	1.12E-04	**26**	18717165	1.35E-04	**21**	46521240	1.25E-04
20	26411124	1.25E-04	**23**	27442828	1.68E-04	**22**	43741335	1.88E-04
15	52466022	1.29E-04	**20**	26411124	1.73E-04	**26**	18717165	1.96E-04
23	27442828	1.31E-04	**17**	5851284	1.73E-04	**26**	25224953	2.05E-04
21	46521182	1.39E-04	**26**	19098773	1.87E-04	**7**	94667256	2.11E-04
21	46521240	1.39E-04	**5**	42703456	1.90E-04	**20**	26411124	2.21E-04
32	24398075	1.51E-04	**6**	7081402	2.29E-04	**5**	42703456	2.25E-04
22	45095964	1.60E-04	**11**	5654983	2.37E-04	**23**	27442828	2.27E-04
11	4069250	1.78E-04	**4**	9497538	2.45E-04	**5**	74779634	2.55E-04
4	9497538	1.93E-04	**22**	45095964	2.51E-04	**26**	19098773	2.71E-04
7	23509046	1.98E-04	**26**	18163019	2.64E-04	**4**	44608273	3.01E-04
11	7252839	2.25E-04	**11**	4069250	2.77E-04	**4**	44872987	3.28E-04
10	83124963	2.71E-04	**2**	115780809	2.82E-04	**19**	26993661	3.49E-04
5	42703456	2.76E-04	**7**	23509046	3.68E-04	**2**	116255680	3.80E-04
6	7081402	3.48E-04	**11**	4586622	3.69E-04	**17**	4360227	3.99E-04
11	4586622	3.66E-04	**26**	18161281	4.02E-04	**1**	59808939	4.01E-04
4	44608273	3.85E-04	**26**	18163088	4.02E-04	**26**	18161281	4.05E-04

Table 2: Top 25 SNPs from GEMMA mixed model analysis in 113 pacers (gender, origin, and age covariates). After pruning, analysis included 59,003 SNPs. Uncorrected p-values are shown for the Likelihood ratio test only. CHR = chromosome; BP = base pair.

CHR	BP	P
15	1859121	6.04E-06
15	1961695	7.41E-06
21	46337210	8.11E-06
21	46521182	8.11E-06
21	46521240	8.11E-06
15	1045079	1.36E-05
15	1189429	1.36E-05
2	25314465	1.39E-05
16	50243195	1.82E-05
21	46537316	1.90E-05
15	1960766	2.14E-05
26	12566139	2.14E-05
11	5654983	4.34E-05
26	9096160	4.39E-05
1	13703668	5.00E-05
15	814072	5.04E-05
2	20240145	7.02E-05
2	20264159	7.02E-05
2	20287080	7.02E-05
2	20332069	7.02E-05
15	1364962	7.40E-05
15	1711463	7.40E-05
15	1874783	7.40E-05
15	1918178	7.40E-05
6	28433490	7.90E-05

Table 3: Top 25 SNPs from GEMMA mixed model analysis in 301 trotters. Comparison is between models including gender, origin, and age +/- distance as covariates. After pruning, analysis included 61,114 SNPs. Uncorrected p-values are shown for the Likelihood ratio test only. CHR = chromosome; BP = base pair.

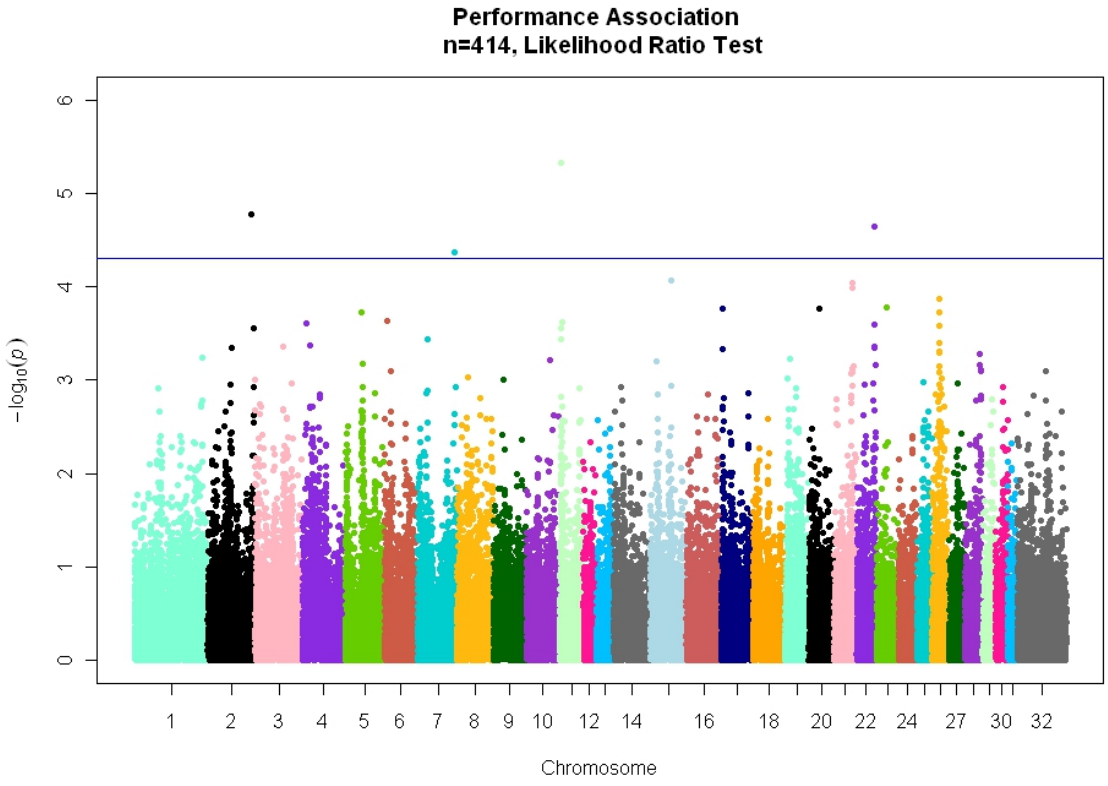
Gender, Origin, Age at fastest time covariates			+ Distance at fastest time		
CHR	**BP**	**P**	**CHR**	**BP**	**P**
22	43741335	3.02E-06	**15**	52466022	7.91E-06
7	94321082	8.86E-05	**19**	28704353	1.92E-05
15	52466022	9.98E-05	**10**	2162356	2.98E-05
2	1.09E+08	1.07E-04	**22**	43741335	3.45E-05
7	94667256	1.13E-04	**10**	2171393	5.04E-05
22	22494583	1.43E-04	**19**	28510403	6.43E-05
22	22494595	1.43E-04	**19**	29690494	6.98E-05
6	7081402	1.76E-04	**19**	29715200	6.98E-05
20	26411124	1.80E-04	**17**	4360227	9.80E-05
27	23343485	1.98E-04	**1**	1.68E+08	1.48E-04
17	4360227	2.28E-04	**20**	26411124	1.63E-04
1	1.69E+08	2.41E-04	**4**	94787923	1.73E-04
27	23348567	2.51E-04	**11**	2867571	1.78E-04
23	29846926	3.17E-04	**1**	1.69E+08	1.98E-04
1	1.69E+08	3.24E-04	**1**	1.69E+08	2.47E-04
18	11016556	3.26E-04	**7**	94667256	2.48E-04
18	11151692	3.93E-04	**7**	94321082	2.52E-04
7	1738119	3.95E-04	**32**	28223572	2.70E-04
23	29650377	4.08E-04	**19**	29958331	2.82E-04
20	27865962	4.37E-04	**19**	29959446	2.82E-04
1	1.72E+08	4.37E-04	**19**	29967544	2.82E-04
7	1969080	5.09E-04	**3**	45882542	3.07E-04
3	70453544	5.95E-04	**1**	1.69E+08	3.14E-04
3	70453579	5.95E-04	**1**	1.17E+08	3.49E-04
3	70453647	5.95E-04	**30**	9818990	4.21E-04

Table 4: Top 25 SNPs from PLINK linear regression model (additive) in 208 European trotters (age and gender covariates). The outcome of interest was the distance at which an individual's fastest time was recorded. After pruning, analysis included 62,177 SNPs. CHR = chromosome; BP = base pair; A1 = tested (minor) allele; BETA = regression coefficient.

CHR	BP	A1	BETA	P
3	90145541	3	-92.04	1.53E-05
23	3700429	1	99.1	2.03E-05
14	51731935	1	117	2.46E-05
23	12948388	1	100.1	4.38E-05
14	25019300	3	-89.91	4.85E-05
16	11729336	1	243.4	6.02E-05
16	11853791	3	243.4	6.02E-05
14	24507021	3	-90.57	7.26E-05
14	24550312	1	-90.57	7.26E-05
14	25000674	3	-88.52	7.34E-05
16	11949745	2	194.7	8.48E-05
23	12947247	3	100.5	8.59E-05
4	1.04E+08	1	329.3	0.000101
7	40454393	3	310.4	0.000108
21	2393047	1	264.1	0.000111
14	50550458	1	123.7	0.000116
14	40329699	1	106.1	0.000153
14	21184338	1	153.8	0.000178
14	25294678	1	-82.52	0.000206
14	25403200	3	-82.52	0.000206
7	40216150	3	413.6	0.00023
23	5094115	1	80.35	0.000232
14	50468965	2	117.5	0.000245
14	51638043	1	85.06	0.000255
14	25481429	1	-81.44	0.000283

Figure 1: Manhattan plot of results from mixed model analysis using GEMMA in a population of 414 pacers and trotters. Fastest recorded time over a mile was the outcome variable of interest (continuous). A) gender, gait, origin, and age at fastest time included as covariates in the model; B) distance at fastest time included as an additional covariate. The 31 autosomal and X chromosome (32) are represented in different colors along the x-axis and the –log(p-value) of the likelihood ratio test is on the y-axis. Each colored dot represents a SNP. The blue line marks a p-value of 5×10^{-5}. SNPs falling above the blue line are considered moderately significant. Top hits in the first model (A) are on ECA 2, 7, 11, and 22. Top hits in the second model (B) are on ECA 11, 15, and 17.

A)



B)

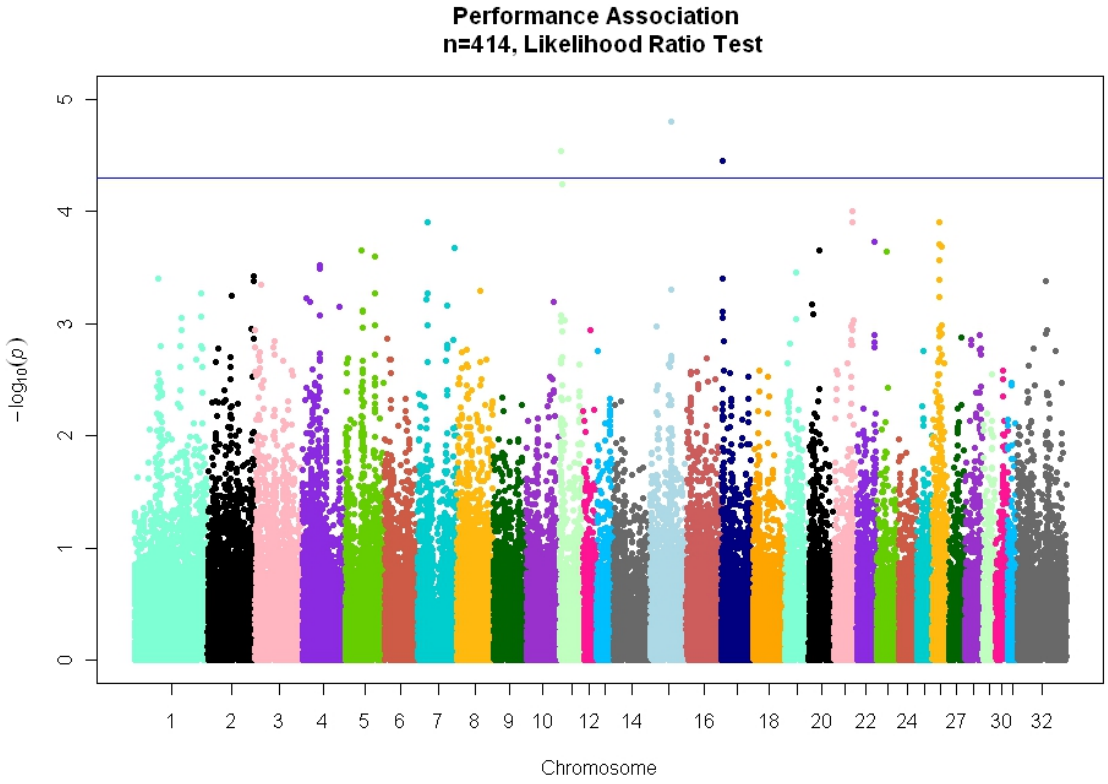
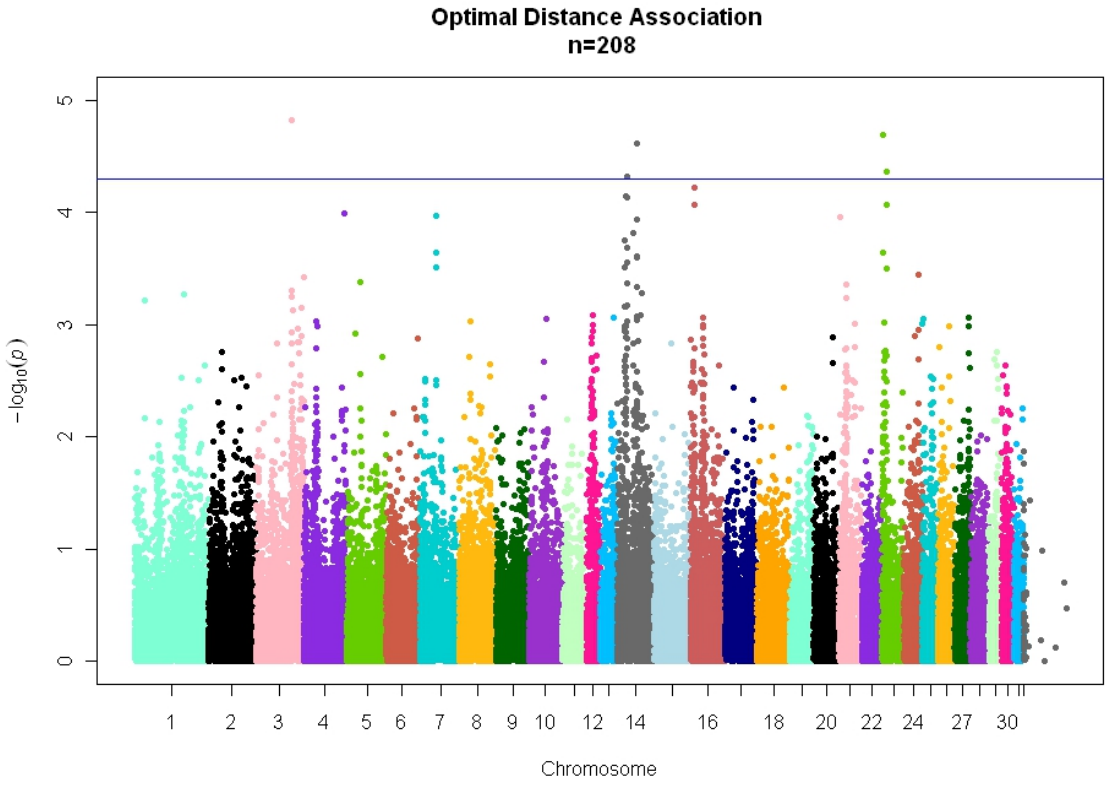

Performance Association
n=414, Likelihood Ratio Test
−log10(p)
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
12
14
16
18
20
22
24
27
30
32
Chromosome

Figure 2: Manhattan plot from linear regression analysis (additive model) using PLINK in a population of 208 trotters raced in Europe over varying distances. The distance at which a horse's fastest time was recorded was the outcome variable of interest (continuous). Age and gender were included in the model as covariates. See Figure 1 for complete legend. Moderately significant hits ($p < 5 \times 10^{-5}$) are on ECA 3, 14, and 23.



Chapter 9

Conclusions and Future Directions

Conclusions

Studying complex diseases and traits in the horse: There has been much success over the past 20 years in identifying the causative alleles underlying simple traits – that is, those governed by Mendelian inheritance of a single gene – in the horse.[e.g.206;238;317;318] However, the same strategies that have led to the discovery of these simple traits, such as linkage analysis and candidate gene approaches, are unlikely to be successful (or at least, efficient) at identifying the multiple interacting alleles underlying complex/polygenic traits. An alternative investigational approach is needed that can account for environmental risk factors, issues related to population structure in large study cohorts, and epistatic interactions. The overarching goal of this thesis work was to develop and apply such an approach to three complex diseases/traits of importance in the Standardbred breed, namely, osteochondrosis (OC), gait, and performance.

This general approach, as described in Chapters 4, 6, 7, and 8, utilizes successive steps that allow for iterative narrowing and expansion of focus: genome-wide association (GWA) analysis to identify specific chromosomal regions of interest, followed by whole-genome sequencing (WGS) for variant discovery within the regions of interest, and then prioritization of the discovered variants for genotyping in a larger population. This genotyping may be followed by additional fine mapping as needed, but more importantly by computational analyses looking for potential interactions between genes containing putative functional variants (see **Future Directions**). Each of these steps has attendant challenges to be overcome. GWA studies require assembly of an adequately-sized, well-phenotyped cohort, and appropriate steps must be taken to account for population structure and known environmental influences. WGS, though more affordable than ever

before, still has a trade-off between coverage and number of individuals sequenced. Finally, while it is clear that large-scale genotyping cannot possibly be performed for every variant discovered within a region of interest, prioritization is heavily dependent on the annotation of the reference genome, and there is potential for important variants to be passed over. Despite these potential pitfalls, the results reported for each of the investigated diseases/traits are promising, and suggest that this approach can be successful for studying complex traits in the horse.

Studying OC illustrates the importance of environmental factors and population structure: The first central hypothesis of this thesis was that one or more genes of moderate to major effect underlie OC risk in the horse and that these risk alleles are shared across breeds. Standardbreds were selected as a model population for this study because the high prevalence and heritability of OC in this breed suggests that major risk alleles should be present at a high frequency. This study is certainly not the first to try to identify chromosomal loci associated with OC in the horse. Several GWAS have been published in recent years, including studies in populations of (Norwegian) Standardbreds and French Trotters.[53;63-66;208] However, the findings have not been consistent across studies; in fact only a few loci have overlapped even partially between two or more reports. Since these GWAS were performed in a variety of breeds and for OC lesions in multiple anatomical predilection sites, it is possible that some of these discrepancies might reflect true differences in risk alleles. However, it is well-known that environmental influences such as diet and exercise play an important role in the development of OC[137;138], and failure to account for these could also explain the lack of agreement. Additionally, the computational approaches used in these studies did not

always account for relatedness among individuals. Since all modern Standardbreds trace back to a single founder stallion[1;2], this could be a crucial oversight. Affected individuals that are related to each other are likely to exhibit identity by descent that is completely unrelated to their disease status, and failure to account for this can increase type I error.[55]

The OC study population reported in Chapter 4 was specifically selected to help overcome potential environmental confounders. All of the horses were raised on a single breeding farm with standardized management practices and therefore had a shared environment during the susceptible period for development of OC (during the first year of life). As would be expected, these horses were closely related to each other. In fact, the initial study cohort was designed to include case-control half sibling pairs, similar to a traditional family linkage analysis study. The use of a family-based cohort in a GWAS can actually be advantageous because the frequency of important rare alleles will be enhanced within the population[56;68;69], but relatedness must obviously be accounted for during analysis. Several approaches were tried to accomplish this, but the most effective seemed to be the use of a mixed-model analysis incorporating a marker-based relationship matrix in the program GEMMA (Genome-wide Efficient Mixed Model Analysis).[214] GEMMA has additional advantages, including potential inclusion of covariates in the mixed model, calculation of individual marker variance (which improves power when looking for a few markers of moderate to major effect), and high efficiency when handling large data sets. Based on all of these factors, the decision was made to use GEMMA to perform GWA analyses for gait and performance as well.

Using this approach, as reported in Chapter 4, two distinct loci of interest were identified on ECA14 that contained several plausible candidate genes based on their

known physiologic function. WGS was completed in 18 horses (see Chapter 6), sequenced at either shallow (6x, n = 12) or moderate (12x, n = 6) depth. This combination was selected in an effort to balance confidence in variant calls (related to depth) with efficient variant discovery representative of the larger population (related to number of individuals sequenced). Thousands of variants were found within the regions of interest, of which a subset were selected for follow-up genotyping using a Sequenom assay. The modest significance of the individual markers after analysis of the Sequenom genotyping data may indicate that the actual risk variants have yet to be discovered (i.e. through additional fine mapping), but more likely reflect the polygenic nature of OC. It is likely that a the combined effect of several markers will result in increased disease risk, rather than any single marker alone, and so a crucial next step in this work is to try to better understand the potential interactions between markers (see **Future Directions**).

Studying gait illustrates challenges related to annotation and known gene function: The second central hypothesis of this thesis was that modifying loci that interact with a known mutation in the gene *DMRT3* underlie the ability to perform specific alternative gaits and are shared across breeds that have been selected for similar gaits. Standardbreds were selected as a model population because while *DMRT3* is nearly fixed in the breed, not all Standardbreds exhibit pacing, the alternative gait for this breed. Thus, modifying variants must exist that determine the ability to pace, and should be present at high frequency in the population. As reported in Chapter 7, GWA analysis revealed regions on five chromosomes (ECA1, 6, 17, 23, and 25) that were strongly associated with gait, and hundreds of the variants within these regions discovered by WGS perfectly, or nearly perfectly, segregated between pacers and trotters. The situation here

was nearly the opposite of that seen with the OC data – instead of only a few moderately associated markers, there was an overabundance of highly associated markers – and yet, the challenge of how to appropriately prioritize these markers for follow-up genotyping remained the same.

In this study (as for OC), prioritization of markers was based on putative functional effects predicted by SnpEff, a utility in the Broad Institute's Genome Analysis ToolKit (GATK). These predictions are based on the current reference equine genome (EquCab 2.0, September 2007), which, while an excellent draft quality reference, is known to be incompletely annotated. In particular, the first exon and the 5' untranslated region (UTR) of many genes are known to be missing. Thus, it is possible that functional variants were passed over in our prioritization scheme. Additionally, since only a fraction of the discovered variants could be genotyped in a large population, ideally priority would have been given to variants falling within plausible candidate genes. However, as the horse is unique among quadrupeds in exhibiting alternative gaits as a physiologic rather than pathologic adaptation, very little is known about what genes might play important roles in this trait. We considered genes with a known role in neural development to be strong potential candidates, but it is likely that some important genes, like *DMRT3*, have never before been described to have such a function.

It is unlikely that each of the 156 variants that were found to be statistically significantly associated with gait after analysis of the Sequenom genotyping data are actually functional modifying variants. The presence of extensive long-range LD in the Standardbred[59] makes it likely that many of these variants have been inherited together and are merely "marking" one functional variant; this is supported by our discovery of a

5Mb region on ECA17 with over 700 variants segregating perfectly with gait, only 12 of which were predicted to have functional effect. It is also possible that some of these variants are related to differences between pacers and trotters other than gait since selective breeding over the past 100 years has resulted in genetic separation between the groups that is similar to that between other separate breeds.[10] Further prioritization can be performed by trying to establish interactions between the markers, however, the lack of knowledge about what genes play a role in the development of alternative gaits makes standard pathway analysis problematic. A computational approach, such as random forest analysis, offers an alternative that does not require *a priori* knowledge about gene function, and is likely the next best step in analysis of this data (see **Future Directions**).

Studying performance illustrates the importance of phenotype: The third central hypothesis of this thesis was that use of fastest recorded speed as a phenotype for GWA analysis could identify genetic factors underlying performance in the Standardbred. Although it is widely accepted that performance traits are inherited to some degree, and in fact, hundreds of years of selective breeding are based upon this premise, specific genetic factors underlying performance in the horse are largely unknown. This may be due, in large part, to the challenge in establishing the most appropriate phenotype upon which to base studies. In production animals, specific traits of interest are relatively easy to identify and quantify. Indeed, specific variants relating to traits such as fatty acid composition in milk[319] and back fat thickness in pigs[320] have been reported. However, in horses, not only do desirable performance traits vary widely by breed and discipline, but they are often subjective in nature (i.e. gait "quality" in a dressage horse) and/or can be affected by a number of external/environmental factors (i.e. lifetime earnings for a

racehorse). The importance of selecting an appropriate phenotype as part of designing a GWAS cannot be overemphasized. Ideally, the phenotype should be specific and measurable; the less certain the phenotype, the higher the risk of misclassification bias and false positive results.[55]

The obvious question becomes, how does one select a single specific, measurable phenotype reflecting performance in the horse? Clearly, this is not possible. Like so many complex traits, performance will need to be broken down into pieces to unravel its genetic influences. Fundamentally, physiologic capacity drives performance, and so objective measures relating to the cardiovascular, respiratory, and musculoskeletal systems would, in theory, serve as ideal phenotypes for investigation and should be applicable across breeds and disciplines. However, these measures are not readily available in large populations of horses, so in the meantime, proxy measures must be employed (such as fastest speed in the current study). Additional follow-up of the results presented in Chapter 8, including examination of specific putative functional variants within chromosomal regions of interest and validation in an independent population of Standardbreds, will need to be completed before a thorough evaluation can be made of the utility of speed as a phenotype in future investigations.

Future Directions

Each of the three traits investigated in this thesis are complex, involving the interaction of variants within multiple genes. A crucial next step for this work will be to investigate these interactions. It is also likely that the actual variants of functional effect have yet to be discovered for these traits, and additional fine mapping efforts will be

required. New tools and resources in the equine genetics community will aid these efforts in the future.

Investigate variant and gene interactions using pathway analysis and random forest analysis: GWAS, by its nature, evaluates the association of single markers with a given trait. As a result of this multiple testing, stringent p-values are set to determine statistical significance of each marker. Unfortunately, using this approach, genes that are truly associated with the trait may be missed, especially if the study has limited power.[321] As a complementary alternative, pathway-based approaches have been developed. Pathway analysis uses prior knowledge of gene function to determine if groups of functionally related genes are significantly associated with a trait of interest.[321] Although this approach was originally designed for use with microarray gene expression data[322;323], it has been modified for use with SNP data.[321] Pathway analysis of large-scale GWAS data has successfully identified candidate genes for several complex traits in humans, including multiple factors related to cholesterol metabolism.[324;325]

A number of tools are available that could be leveraged for follow-up of the Sequenom data presented for OC (Chapter 6) and gait (Chapter 7). These include IPA (Ingenuity Pathway Analysis; Ingenuity Systems, Redwood, CA), Gene Set Enrichment Analysis (GSEA)[256], ClueGO[279], and GRAIL (Gene Relationships Across Implicated Loci).[280] All are based on text mining, and many require the use of "seed" gene lists or regions upon which to build their analysis. These lists are user inputs based on genes that are known (or strongly suspected) to be important to the trait of interest. For example, for OC, the "seed" list would include those genes with reported roles in skeletogenesis, particularly endochondral ossification. For gait, genes known to be important in neural

development, particularly the cerebellum and other regions important to coordinated movement, would be included. No single pathway analysis tool is best for every situation, and it has been recommended that the results from multiple approaches be compared and subsequently validated in an independent population.[321]

It is likely that there are genes involved in OC and/or gait that have not had their physiologic role completely defined, and these could be missed using the methods described above. Random forest analysis provides an alternative computational approach that does not require prior knowledge about gene function to establish connections between genotype and phenotype. This approach has the additional advantage of being able to accommodate multiple variants within a single gene as well as non-genotype predictors (i.e. relatedness or other important covariates). In a random forest approach, a series of decision trees are constructed to classify individuals as "cases" or "controls" based on a randomly chosen subset of predictors (SNPs). The importance of each predictor is determined by the number of individuals that are misclassified when the value of that predictor (SNP genotype) is randomly permuted.[258] When applied to the Sequenom data reported for OC (Chapter 6) and gait (Chapter 7), random forest analysis will guide variant prioritization and will be able to help elucidate novel interactions between SNPs. Non-genotype predictors that will be included include principal components representing relatedness between individuals (calculated based on ancestry informative markers [AIMs]) for both traits, and gait for OC.

A combination of approaches may yield the most interesting results. The use of pathways as "synthetic features" in random forest analysis has been reported recently.[259] In this approach, prioritized SNPs were organized into physiologic pathways based on

gene ontology (GO) terms. Random forest analysis was first carried out at the individual SNP level, and the results of each SNP within a pathway were combined into a single continuous variable reflecting a predicted probability for that pathway (that is, the probability of an individual being a case given their aggregated genotypes across all SNPs in this pathway). Random forest analysis was then repeated for these pathway-level parameters to look for pathway-level interactions. Using this approach, the authors were able to identify novel putative biological mechanisms underlying bladder cancer in humans.[259] It is not hard to believe that a "higher order" analysis such as this would yield insights into a complex trait that are far beyond any that would be possible at the single marker level. Without a doubt, future justification of the functional effects of any variant thought to be truly associated with a complex disease or trait will necessitate a thorough understanding of many gene interactions at the cell and tissue level, rather than a superficial understanding of the function of a single gene.

Leverage new tools and resources for functional variant discovery: The human genome is considered "complete," yet updated releases are made public every three months as annotations are added and improved.[326] Additionally, DNA previously thought to be "junk" is increasingly being revealed to serve biologically relevant roles in diverse cell types.[250] As our knowledge of the genome improves, previous work can be re-evaluated using new tools and resources, and new insights can be gained. For example, many SNPs associated with complex human diseases that were previously thought to have no functional effect have been discovered to overlap with newly-annotated regulatory regions[250], providing new avenues of investigation. What is true for the human genome is also true for the horse, and as new tools and resources are developed by the

equine genetics community, they can be leveraged to improve data analysis and subsequently our understanding of complex diseases and traits.

Improved Genome Annotation: The current version of the equine genome (EquCab 2.0)[72] was released in September 2007. While this is an excellent draft-quality genome, it has known errors and shortcomings in its annotation. With respect to the latter, most notably, the first exon and 5' and 3' untranslated regions are missing from many genes, which makes recognition of variants with putative functional effect more challenging. Within the past two years, a concerted effort has been made to improve the assembly and annotation of the equine genome by using a combination of Sanger sequencing reads (from the original assembly), deep Illumina sequence data, long DNA reads (Illumina Moleculo technology), mate-paired reads, RNAseq data, and optical mapping.[327] Considerable progress has been made in this effort, and EquCab 3.0 is projected to be released within the next year. When this new reference is made available, our existing whole-genome sequencing data can be mapped to it. It is anticipated that new variants will be identified within our chromosomal regions of interest because gaps in the current reference will have been filled in. Many of these may have putative functional effect based on the improved gene model annotations and can be followed up in a larger population.

Imputation Resources: Genotype imputation is a technique that statistically estimates genotypes from non-assayed SNPs by comparing haplotype blocks in the study population with haplotype blocks in a more densely genotyped reference population.[328] Imputation is a validated and widely accepted technique in human genetics, where dense reference sets compiled from HapMap and the 1000 Genomes Project are publically

available. User-friendly data pipelines have been established for phasing, imputation, and subsequent association testing of human genotyping data[e.g.211;234], but until recently, a similar pipeline had not been validated in the horse ([329] and Chapter 5) and its use with experimental data has not previously been reported.

For each of the traits investigated in this thesis, genotype imputation between the two existing equine genotyping arrays (Illumina Equine SNP50 and SNP70 beadchips) was used to reduce the amount of data lost from non-overlapping SNPs on these platforms (Chapters 4, 7, 8). This approach resulted in the inclusion of 14,000-18,000 SNPs that would have otherwise been excluded from the various analyses. While this improvement may seem modest, the increased density of markers helps to narrow the chromosomal regions of interest in a GWAS. More importantly, however, the success of this approach in this thesis work provides proof of concept for the utility of imputation in the horse. As new genotyping platforms become available, (i.e. the Equine SNP670 chip, projected to go into production within the next few months) imputation will allow continued use of existing data without the cost of re-genotyping individuals.

Imputation can also be performed using sequencing data, raising the possibility of going from tens of thousands of markers to millions of markers. The newest version of the BEAGLE software used in this thesis (Beagle 4[330]) is designed to use variant calling files (VCF) for both input and output. While this is less convenient for existing genotype platform data, it facilitates the use of whole-genome sequencing data and eliminates many of the data transformation steps required in the current pipeline. The current limitation with this approach is that the number of sequenced horses available for use as a reference population within an individual laboratory is generally small, which may

reduce the accuracy of imputation.[329] However, there is an ongoing effort in the equine genetics community to consolidate whole-genome sequencing data from hundreds of horses from around the world into a single community resource, similar to the 1000 Genomes Project (albeit on a smaller scale). Additionally, genotyping data from dozens of horses genotyped on an experimental 2 million SNP chip will be made publically available (analogous to HapMap). The availability of these imputation resources will help to move this technique into mainstream use in equine genetics. It will also allow re-analysis of the data presented in this thesis at a fraction of the cost of repeating the experiments.

References

1. United States Trotting Association. Standardbred Breed Information. Available at: http://fanguide.ustrotting.com/standardbred-breed.cfm. Accessed 18 Nov 2013.

2. Standardbred Canada. History of the Standardbred. Available at: http://www.standardbredcanada.ca/content/new-racing-history-standardbred.html. Accessed 18 Nov 2013.

3. Lynghaug F. *The Official Horse Breeds Standards Guide: the complete guide to the standards of all North American equine breed associations*. Minneapolis, MN: Voyageur Press, 2009:318-322.

4. United States Trotting Association. The Trotting and Pacing Guide. Available at: http://www.ustrotting.com/trackside/tpg/tpg.cfm. Accessed 18 Nov 2013.

5. United States Trotting Association. USTA 2013/2014 Charter, Bylaws, Rules, and Regulations. Available at: http://www.ustrotting.com/pdf/USTARuleBook.pdf. Accessed 18 Nov 2013.

6. Edwards EH. *The New Encyclopedia of the Horse*. New York: Dorling Kindersley Publishing, Inc., 2000.

7. Crowell P. *Cavalcade of American Horses*. New York: McGraw-Hill Book Company, Inc., 1951:58-60.

8. Physick-Sheard PW (1986), Career profile of the Canadian Standardbred. II. Influence of age, gait and sex upon number of races, money won and race times, *Can.J.Vet.Res.* 50: 457-470.

9. Cheetham J, Riordan AS, Mohammed HO, McIlwraith CW, Fortier LA (2010), Relationships between race earnings and horse age, sex, gait, track surface and number of race starts for Thoroughbred and Standardbred racehorses in North America, *Equine Vet.J.* 42: 346-350.

10. Cothran EG, MacCluer JW, Weitkamp LR, Bailey E (1987), Genetic differentiation associated with gait within American standardbred horses, *Anim Genet.* 18: 285-296.

11. Andersson LS, Larhammar M, Memic F, Wootz H, Schwochow D, Rubin CJ, Patra K, Arnason T, Wellbring L, Hjalm G, Imsland F, Petersen JL, McCue ME, Mickelson JR, Cothran G, Ahituv N, Roepstorff L, Mikko S, Vallstedt A, Lindgren G, Andersson L, Kullander K (2012), Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice, *Nature* 488: 642-646.

12. Arnason T (1999), Genetic evaluation of Swedish standard-bred trotters for racing performance traits and racing status, *J.Anim Breed.Genet.* 116: 387-389.

13. Thuneberg-Selonen T, Poso J, Mantysaari E, Ojala M (1999), Use of individual race results in the estimation of genetic parameters of trotting performance for Finnhorse and Standardbred trotters, *Agr.Food Sci.Finland* 8: 353-363.

14. Tolley EA, Notter DR, Marlowe TJ (1983), Heritability and repeatability of speed for 2- and 3-year-old standardbred racehorses, *J.Anim Sci.* 56: 1294-1305.

15. Bugislaus AE, Roehe R, Willms F, Kalm E (2006), The use of a random regression model to account for change in racing speed of German trotters with increasing age, *J.Anim Breed.Genet.* 123: 239-246.

16. Kane AJ, McIlwraith CW, Park RD, Rantanen NW, Morehead JP, Bramlage LR (2003), Radiographic changes in Thoroughbred yearlings. Part 2: Associations with racing performance, *Equine Vet.J.* 35: 366-374.

17. Preston SA, Brown MP, Trumble TN, Chmielewski TL, Zimmel DN, Hernandez JA (2012), Effects of various presale radiographic findings for yearling Thoroughbreds on 2-year-old racing performance, *J.Am.Vet.Med.Assoc.* 241: 1505-1513.

18. Meagher DM, Bromberek JL, Meagher DT, Gardner IA, Puchalski SM, Stover SM (2013), Prevalence of abnormal radiographic findings in 2-year-old Thoroughbreds at in-training sales and associations with racing performance, *J.Am.Vet.Med.Assoc.* 242: 969-976.

19. Cohen ND, Carter GK, Watkins JP, O'Conor MS (2006), Association of racing performance with specific abnormal radiographic findings in Thoroughbred yearlings sold in Texas, *J.Equine Vet.Sci.* 26: 462-474.

20. Robert C, Valette JP, Jacquet S, Denoix JM (2013), Influence of juvenile osteochondral conditions on racing performance in Thoroughbreds born in Normandy, *Vet.J.* 197: 83-89.

21. Robert C, Valette JP, Denoix JM (2006), Correlation between routine radiographic findings and early racing career in French trotters, *Equine Vet.J.Suppl* 38(S36): 473-478.

22. Courouce-Malblanc A, Leleu C, Bouchilloux M, Geffroy O (2006), Abnormal radiographic findings in 865 French standardbred trotters and their relationship to racing performance, *Equine Vet.J.Suppl* 38(S36): 417-422.

23. Fortier LA, Nixon AJ (2005), New surgical treatments for osteochondritis dissecans and subchondral bone cysts, *Vet.Clin.North Am.Equine Pract.* 21: 673-90, vii.

24. Laws EG, Richardson DW, Ross MW, Moyer W (1993), Racing performance of standardbreds after conservative and surgical treatment for tarsocrural osteochondrosis, *Equine Vet.J.* 25: 199-202.

25. Beard WL, Bramlage LR, Schneider RK, Embertson RM (1994), Postoperative racing performance in standardbreds and thoroughbreds with osteochondrosis of the tarsocrural joint: 109 cases (1984-1990), *J.Am.Vet.Med.Assoc.* 204: 1655-1659.

26. Grondahl AM, Engeland A (1995), Influence of radiographically detectable orthopedic changes on racing performance in standardbred trotters, *J.Am.Vet.Med.Assoc.* 206: 1013-1017.

27. Jorgensen HS, Proschowsky H, Falk-Ronne J, Willeberg P, Hesselholt M (1997), The significance of routine radiographic findings with respect to subsequent racing performance and longevity in standardbred trotters, *Equine Vet.J.* 29: 55-59.

28. Brehm W, Staecker W (1999), Osteochondrosis (OCD) in the tarsocrural joint of Standardbred trotters - correlation between radiographic findings and racing, *Am.Assoc.Equine Practr.Proc.* 45: 164-166

29. Torre F, Motta M (2000), Osteochondrosis of the tarsocrural joint and osteochondral fragments in the fetlock joints: incidence and influence on racing performance in a selected group of Standardbred trotters, *Am.Assoc.Equine Practr.Proc.* 46: 287-294.

30. Denoix JM, Jeffcott LB, McIlwraith CW, van Weeren PR (2013), A review of terminology for equine juvenile osteochondral conditions (JOCC) based on anatomical and functional considerations, *Vet.J.* 197: 29-35.

31. Schougaard H, Falk RJ, Phillipson J (1990), A radiographic survey of tibiotarsal osteochondrosis in a selected population of trotting horses in Denmark and its possible genetic significance, *Equine Vet.J.* 22: 288-289.

32. Grondahl AM, Dolvik NI (1993), Heritability estimations of osteochondrosis in the tibiotarsal joint and of bony fragments in the palmar/plantar portion of the metacarpo- and metatarsophalangeal joints of horses, *J.Am.Vet.Med.Assoc.* 203: 101-104.

33. Alvaredo AF, Marcoux M, Breton L (1989), The incidence of osteochondrosis in a Standardbred breeding farm in Quebec, *Am.Assoc.Equine Practr.Proc.* 35: 293-307.

34. Lykkjen S, Roed KH, Dolvik NI (2012), Osteochondrosis and osteochondral fragments in Standardbred trotters: prevalence and relationships, *Equine Vet.J.* 44: 332-338.

35. Denoix JM, Jacquet S, Lepeule J, Crevier-Denoix N, Valette JP, Robert C (2013), Radiographic findings of juvenile osteochondral conditions detected in 392 foals using a field radiographic protocol, *Vet.J.* 197: 44-51.

36. Carlsten J, Sandgren B, Dalin G (1993), Development of osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints of Standardbred trotters. I. A radiological survey, *Equine Vet.J.Suppl* 16: 42-47.

37. Philipsson J, Andreasson E, Sandgren B, Dalin G, Carlsten J (1993), Osteochondrosis in the tarsocrural joint and osteochondral fragments in the fetlock joints in Standardbred trotters. II. Heritability, *Equine Vet.J.Suppl* 16: 38-41.

38. Hoppe F, Philipsson J (1985), A genetic study of osteochondrosis dissecans in Swedish horses, *Equine Practice* 7: 7-15.

39. Kane AJ, Park RD, McIlwraith CW, Rantanen NW, Morehead JP, Bramlage LR (2003), Radiographic changes in Thoroughbred yearlings. Part 1: Prevalence at the time of the yearling sales, *Equine Vet.J.* 35: 354-365.

40. Oliver LJ, Baird DK, Baird AN, Moore GE (2008), Prevalence and distribution of radiographically evident lesions on repository films in the hock and stifle joints of yearling Thoroughbred horses in New Zealand, *N.Z.Vet.J.* 56: 202-209.

41. Howard BA, Embertson RM, Rantanen NW, Bramlage LR (1992), Survey radiographic findings in Thoroughbred sales yearlings, *Am.Assoc.Equine Practr.Proc.* 38: 397-402.

42. Ricard A, Perrocheau M, Courouce-Malblanc A, Valette JP, Tourtoulou G, Dufosset JM, Robert C, Chaffaux S, Denoix JM, Guerin G (2013), Genetic parameters of juvenile osteochondral conditions (JOCC) in French Trotters, *Vet.J.* 197: 77-82.

43. Falconer DS, MackayT.F.C.*Introduction to Quantitative Genetics*, Essex, England: Pearson Education Limited, 1996:160-163.

44. Mike de Kock Racing. OCD: commercial minefield. Available at: http://mikedekockracing.com/wp/index.php/2013/01/ocd-commercial-minefield/. Accessed 23 Oct 2013.

45. Agarwala V, Flannick J, Sunyaev S, Altshuler D (2013), Evaluating empirical bounds on complex disease genetic architecture, *Nat.Genet.* 45: 1418-1427.

46. Gibson G (2011), Rare and common variants: twenty arguments, *Nat.Rev.Genet.* 13: 135-145.

47. Mackay TF, Stone EA, Ayroles JF (2009), The genetics of quantitative traits: challenges and prospects, *Nat.Rev.Genet.* 10: 565-577.

48. Goddard ME, Hayes BJ (2009), Mapping genes for complex traits in domestic animals and their use in breeding programmes, *Nat.Rev.Genet.* 10: 381-391.

49. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010), Missing heritability and strategies for finding the underlying causes of complex disease, *Nat.Rev.Genet.* 11: 446-450.

50. Queitsch C, Carlson KD, Girirajan S (2012), Lessons from model organisms: phenotypic robustness and missing heritability in complex disease, *PLoS.Genet.* 8: e1003041.

51. Lango AH, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segre AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon SA, Magi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin LK, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Zillikens MC, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Zhao JH, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani JM, Kaplan LM, Kettunen J, Konig IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Muller M, Suh NJ, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De GA, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpelainen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Pare G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietilainen KH, Pouta A, Ridderstrale M, Rotter JI, Sambrook JG, Sanders AR, Schmidt CO, Sinisalo J, Smit JH, Stringham HM, Bragi WG, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P, Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den HM, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kahonen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimaki T, Melander O, Mosley TH, Jr., Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tonjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Gronberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Lathrop GM, Salomaa V,

Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Cupples LA, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH (2010), Hundreds of variants clustered in genomic loci and biological pathways affect human height, *Nature* 467: 832-838.

52. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009), Common polygenic variation contributes to risk of schizophrenia and bipolar disorder, *Nature* 460: 748-752.

53. Dierks C, Lohring K, Lampe V, Wittwer C, Drogemuller C, Distl O (2007), Genome-wide search for markers associated with osteochondrosis in Hanoverian warmblood horses, *Mamm.Genome* 18: 739-747.

54. Spencer CC, Su Z, Donnelly P, Marchini J (2009), Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip, *PLoS.Genet.* 5: e1000477.

55. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN (2008), Genome-wide association studies for complex traits: consensus, uncertainty and challenges, *Nat.Rev.Genet.* 9: 356-369.

56. Strachan T, Read A. *Human Molecular Genetics*. New York: Garland Science, 2011:467-496.

57. Abecasis GR, Ghosh D, Nichols TE (2005), Linkage disequilibrium: ancient history drives the new genetics, *Hum.Hered.* 59: 118-124.

58. Wall JD, Pritchard JK (2003), Haplotype blocks and linkage disequilibrium in the human genome, *Nat.Rev.Genet.* 4: 587-597.

59. McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, Distl O, Guerin G, Hasegawa T, Hill EW, Leeb T, Lindgren G, Penedo MC, Roed KH, Ryder OA, Swinburne JE, Tozaki T, Valberg SJ, Vaudin M, Lindblad-Toh K, Wade CM, Mickelson JR (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: e1002451.

60. Sutter NB, Eberle MA, Parker HG, Pullar BJ, Kirkness EF, Kruglyak L, Ostrander EA (2004), Extensive and breed-specific linkage disequilibrium in Canis familiaris, *Genome Res.* 14: 2388-2396

61. Gray MM, Granka JM, Bustamante CD, Sutter NB, Boyko AR, Zhu L, Ostrander EA, Wayne RK (2009), Linkage disequilibrium and demographic history of wild and domestic canids, *Genetics* 181: 1493-1505.

62. Petersen JL, Mickelson JR, Rendahl AK, Valberg SJ, Andersson LS, Axelsson J, Bailey E, Bannasch D, Binns MM, Borges AS, Brama P, da Camara MA,

Capomaccio S, Cappelli K, Cothran EG, Distl O, Fox-Clipsham L, Graves KT, Guerin G, Haase B, Hasegawa T, Hemmann K, Hill EW, Leeb T, Lindgren G, Lohi H, Lopes MS, McGivney BA, Mikko S, Orr N, Penedo MC, Piercy RJ, Raekallio M, Rieder S, Roed KH, Swinburne J, Tozaki T, Vaudin M, Wade CM, McCue ME (2013), Genome-wide analysis reveals selection for important traits in domestic horse breeds, *PLoS.Genet.* 9: e1003211.

63. Teyssedre S, Dupuis MC, Guerin G, Schibler L, Denoix JM, Elsen JM, Ricard A (2012), Genome-wide association studies for osteochondrosis in French Trotter horses, *J.Anim Sci.* 90: 45-53.

64. Orr N, Hill EW, Gu J, Govindarajan P, Conroy J, van Grevenhof EM, Ducro BJ, van Arendonk JA, Knaap JH, van Weeren PR, Machugh DE, Ennis S, Brama PA (2013), Genome-wide association study of osteochondrosis in the tarsocrural joint of Dutch Warmblood horses identifies susceptibility loci on chromosomes 3 and 10, *Anim Genet.* 44: 408-412.

65. Lykkjen S, Dolvik NI, McCue ME, Rendahl AK, Mickelson JR, Roed KH (2010), Genome-wide association analysis of osteochondrosis of the tibiotarsal joint in Norwegian Standardbred trotters, *Anim Genet.* 41 Suppl 2: 111-120.

66. Corbin LJ, Blott SC, Swinburne JE, Sibbons C, Fox-Clipsham LY, Helwegen M, Parkin TD, Newton JR, Bramlage LR, McIlwraith CW, Bishop SC, Woolliams JA, Vaudin M (2012), A genome-wide association study of osteochondritis dissecans in the Thoroughbred, *Mamm.Genome* 23: 294-303.

67. Wellcome Trust Case Control Consortium (2007), Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, *Nature* 447: 661-678.

68. Li M, Boehnke M, Abecasis GR (2006), Efficient study designs for test of genetic association using sibship data and unrelated cases and controls, *Am.J.Hum.Genet.* 78: 778-792.

69. Jiang D, McPeek MS (2014), Robust rare variant association testing for quantitative traits in samples with related individuals, *Genet.Epidemiol.* 38: 10-20.

70. Fardo DW, Druen AR, Liu J, Mirea L, Infante-Rivard C, Breheny P (2011), Exploration and comparison of methods for combining population- and family-based genetic association using the Genetic Analysis Workshop 17 mini-exome, *BMC.Proc.* 5 Suppl 9: S28.

71. Kazma R, Bailey JN (2011), Population-based and family-based designs to analyze rare variants in complex diseases, *Genet.Epidemiol.* 35 Suppl 1: S41-S47.

72. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blocker H, Distl O, Edgar RC, Garber M,

Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della VG, Fryc S, Guerin G, Hasegawa T, Hill EW, Jurka J, Kiialainen A, Lindgren G, Liu J, Magnani E, Mickelson JR, Murray J, Nergadze SG, Onofrio R, Pedroni S, Piras MF, Raudsepp T, Rocchi M, Roed KH, Ryder OA, Searle S, Skow L, Swinburne JE, Syvanen AC, Tozaki T, Valberg SJ, Vaudin M, White JR, Zody MC, Lander ES, Lindblad-Toh K (2009), Genome sequence, comparative analysis, and population genetics of the domestic horse, *Science* 326: 865-867.

73. Cirulli ET, Goldstein DB (2010), Uncovering the roles of rare variants in common disease through whole-genome sequencing, *Nat.Rev.Genet.* 11: 415-425.

74. Buerkle CA, Gompert Z (2013), Population genomics based on low coverage sequencing: how low should we go?, *Mol.Ecol.* 22: 3028-3035.

75. Coleman SJ, Zeng Z, Wang K, Luo S, Khrebtukova I, Mienaltowski MJ, Schroth GP, Liu J, MacLeod JN (2010), Structural annotation of equine protein-coding genes determined by mRNA sequencing, *Anim Genet.* 41 Suppl 2: 121-130.

76. Hurtig MB, Pool RR. Pathogenesis of equine osteochondrosis. In: McIlwraith CW and Trotter GW, eds. *Joint Disease in the Horse.* St. Louis, MO:Saunders Elsevier, 1996; 335-358.

77. Valentino LW, Lillich JD, Gaughan EM, Biller DR, Raub RH (1999), Radiographic prevalence of osteochondrosis in yearling feral horses, *Vet.Comp Orthop.Traumatol.* 12: 151-155.

78. Compston PC, Phillips CR, Payne RJ, Newton JR. Racehorse performance as an epidemiological outcome measure. Poster presented at *Society for Veterinary Epidemiology and Preventive Medicine Annual Conference.* Madrid, Spain, 2013.

79. Hill EW, Fonseca RG, McGivney BA, Gu J, Machugh DE, Katz LM (2012), MSTN genotype (g.66493737C/T) association with speed indices in Thoroughbred racehorses, *J.Appl.Physiol (1985.)* 112: 86-90.

80. Contino EK, Park RD, McIlwraith CW (2012), Prevalence of radiographic changes in yearling and 2-year-old Quarter Horses intended for cutting, *Equine Vet.J.* 44: 185-195.

81. Preston SA, Zimmel DN, Chmielewski TL, Trumble TN, Brown MP, Boneau JC, Hernandez JA (2010), Prevalence of various presale radiographic findings and association of findings with sales price in Thoroughbred yearlings sold in Kentucky, *J.Am.Vet.Med.Assoc.* 236: 440-445.

82. der Kinderen L (2005), *Heritability of osteochondrosis in Dutch Warmblood stallions from the second stallion inspection*. Unpublished dissertation, Wageningen University, Wageningen, The Netherlands.

83. Stock KF, Hamann H, Distl O (2005), Estimation of genetic parameters for the prevalence of osseous fragments in limb joints of Hanoverian Warmblood horses, *J.Anim Breed.Genet.* 122: 271-280.

84. van Grevenhof EM, Schurink A, Ducro BJ, van Weeren PR, Van Tartwijk JM, Bijma P, van Arendonk JA (2009), Genetic variables of various manifestations of osteochondrosis and their correlations between and within joints in Dutch warmblood horses, *J.Anim Sci.* 87: 1906-1912.

85. Stock KF, Distl O (2006), Genetic correlations between osseous fragments in fetlock and hock joints, deforming arthropathy in hock joints and pathologic changes in the navicular bones of Warmblood riding horses, *Livestock Sci.* 105: 35-43.

86. Hilla D, Distl O (2014), Heritabilities and genetic correlations between fetlock, hock and stifle osteochondrosis and fetlock osteochondral fragments in Hanoverian Warmblood horses, *J.Anim Breed.Genet.* 131: 71-81.

87. Vos NJ (2008), Incidence of osteochondrosis (dissecans) in Dutch warmblood horses presented for pre-purchase examination, *Ir.Vet.J.* 61: 33-37.

88. Wittwer C, Hamann H, Rosenberger E, Distl O (2006), Prevalence of osteochondrosis in the limb joints of South German Coldblood horses, *J.Vet.Med.A Physiol Pathol.Clin.Med.* 53: 531-539.

89. Riley CB, Scott WM, Caron JP, Fretz PB, Bailey JV, Barber SM (1998), Osteochondritis dessicans and subchondral cystic lesions in draft horses: a retrospective study, *Can.Vet.J.* 39: 627-633.

90. Pieramati C, Pepe M, Silvestrelli M, Bolla A (2003), Heritability of osteochondrosis dissecans in Maremmano horses, *Livestock Prod.Sci.* 79: 249-255.

91. Yonetani Y, Nakamura N, Natsuume T, Shiozaki Y, Tanaka Y, Horibe S (2010), Histological evaluation of juvenile osteochondritis dissecans of the knee: a case series, *Knee.Surg.Sports Traumatol.Arthrosc.* 18: 723-730.

92. Ytrehus B, Carlson CS, Ekman S (2007), Etiology and pathogenesis of osteochondrosis, *Vet.Pathol.* 44: 429-448.

93. Olstad K, Ytrehus B, Ekman S, Carlson CS, Dolvik NI (2007), Early lesions of osteochondrosis in the distal tibia of foals, *J.Orthop.Res.* 25: 1094-1105.

94. Olstad K, Ytrehus B, Ekman S, Carlson CS, Dolvik NI (2011), Early lesions of articular osteochondrosis in the distal femur of foals, *Vet.Pathol.* 48: 1165-1175.

95. Ytrehus B, Grindflek E, Teige J, Stubsjoen E, Grondalen T, Carlson CS, Ekman S (2004), The effect of parentage on the prevalence, severity and location of lesions of osteochondrosis in swine, *J.Vet.Med.A Physiol Pathol.Clin.Med.* 51: 188-195.

96. Wagoner G, Cohn BNE (1931), Osteochondritis dissecans: a resume of the theories of etiology and the consideration of heredity as an etiologic factor, *Arch.Surg.* 23: 1-25.

97. Edmonds EW, Polousky J (2013), A review of knowledge in osteochondritis dissecans: 123 years of minimal evolution from Konig to the ROCK study group, *Clin.Orthop.Relat Res.* 471: 1118-1126.

98. Howald H (1942), Zur kenntnis der osteochondrosis dissecans (osteochondrotos dissecans), *Archiv fur orthopadische und Unfall-Chirurgie* 41: 730-788.

99. Atanda A, Jr., Shah SA, O'Brien K (2011), Osteochondrosis: common causes of pain in growing bones, *Am.Fam.Physician* 83: 285-291.

100. Pappas AM (1981), Osteochondrosis dissecans, *Clin.Orthop.Relat Res.* 59-69.

101. Lindholm TS, Osterman K, Vankka E (1980), Osteochondritis dissecans of elbow, ankle and hip: a comparison survey, *Clin.Orthop.Relat Res.* 245-253.

102. Schenck RC, Jr., Goodnight JM (1996), Osteochondritis dissecans, *J.Bone Joint Surg.Am.* 78: 439-456.

103. Omer GE, Jr. (1981), Primary articular osteochondroses, *Clin.Orthop.Relat Res.* 33-40.

104. Kocher MS, Tucker R, Ganley TJ, Flynn JM (2006), Management of osteochondritis dissecans of the knee: current concepts review, *Am.J.Sports Med.* 34: 1181-1191.

105. Dipaola JD, Nelson DW, Colville MR (1991), Characterizing osteochondral lesions by magnetic resonance imaging, *Arthroscopy* 7: 101-104.

106. Doyle SM, Monahan A (2010), Osteochondroses: a clinical review for the pediatrician, *Curr.Opin.Pediatr.* 22: 41-46.

107. Kocher MS, Czarnecki JJ, Andersen JS, Micheli LJ (2007), Internal fixation of juvenile osteochondritis dissecans lesions of the knee, *Am.J.Sports Med.* 35: 712-718.

108. Wall E, Von SD (2003), Juvenile osteochondritis dissecans, *Orthop.Clin.North Am.* 34: 341-353.

109. Foland JW, McIlwraith CW, Trotter GW (1992), Arthroscopic surgery for osteochondritis dissecans of the femoropatellar joint of the horse, *Equine Vet.J.* 24: 419-423.

110. Grondalen T (1974), Osteochondrosis and arthrosis in pigs. I. Incidence in animals up to 120 kg live weight, *Acta Vet.Scand.* 15: 1-25.

111. Jorgensen B (2000), Longevity of breeding sows in relation to leg weakness symptoms at six months of age, *Acta Vet.Scand.* 41: 105-121.

112. Jorgensen B (2000), Osteochondrosis/osteoarthrosis and claw disorders in sows, associated with leg weakness, *Acta Vet.Scand.* 41: 123-138.

113. Carlson CS, Meuten DJ, Richardson DC (1991), Ischemic necrosis of cartilage in spontaneous and experimental lesions of osteochondrosis, *J.Orthop.Res.* 9: 317-329.

114. Dewey C. Diseases of the nervois and locomotor systems. In: Straw BE, D'Allaire S, Mengeling WL et al., eds. *Diseases of Swine.* 8th Ed, Oxford: Blackwell Science Ltd, 1999; 861-863.

115. Devine DV, VanPelt SR, Boileau MJ (2011), What is your diagnosis? Osteochondrosis dissecans, *J.Am.Vet.Med.Assoc.* 238: 39-40.

116. Bullough PG *Atlas of Orthopedic Pathology with Clinical and Radiological Correlations,* New York: Gower Medical Publishing, 1992.

117. Jansson N, Ducharme NG (2005), Angular limb deformities in foals: treatment and prognosis, *Comp.Cont.Educ.Pract.* 27: 134-146.

118. Nielsen NA (1933), Osteochondritis dissecans capitulim humeri, *Acta Orthop.Scand.* 4: 307-418.

119. Linden B (1976), The incidence of osteochondritis dissecans in the condyles of the femur, *Acta Orthop.Scand.* 47: 664-667.

120. Duthie RB, Houghton GR (1981), Constitutional aspects of the osteochondroses, *Clin.Orthop.Relat Res.* 19-27.

121. Kroll A, Hertsch B, Hoppner S (2001), Entwicklung osteochondraler veranderungen in den fessel- und talokruralgelenken im rontgenbild beim fohlen, *Pferdeheilkunde* 17: 489-500.

122. Carlson CS, Hilley HD, Meuten DJ, Hagan JM, Moser RL (1988), Effect of reduced growth rate on the prevalence and severity of osteochondrosis in gilts, *Am.J.Vet.Res.* 49: 396-402.

123. Barrie HJ (1987), Osteochondritis dissecans 1887-1987. A centennial look at Konig's memorable phrase, *J.Bone Joint Surg.Br.* 69: 693-695.

124. Uozumi H, Sugita T, Aizawa T, Takahashi A, Ohnuma M, Itoi E (2009), Histologic findings and possible causes of osteochondritis dissecans of the knee, *Am.J.Sports Med.* 37: 2003-2008.

125. Cahill BR (1995), Osteochondritis Dissecans of the Knee: Treatment of Juvenile and Adult Forms, *J.Am.Acad.Orthop.Surg.* 3: 237-247.

126. Skagen PS, Horn T, Kruse HA, Staergaard B, Rapport MM, Nicolaisen T (2011), Osteochondritis dissecans (OCD), an endoplasmic reticulum storage disease?: a morphological and molecular study of OCD fragments, *Scand.J.Med.Sci.Sports* 21: e17-e33.

127. van de Lest CH, Brama PA, van EB, DeGroot J, van Weeren PR (2004), Extracellular matrix changes in early osteochondrotic defects in foals: a key role for collagen?, *Biochim.Biophys.Acta* 1690: 54-62.

128. Lecocq M, Girard CA, Fogarty U, Beauchamp G, Richard H, Laverty S (2008), Cartilage matrix changes in the developing epiphysis: early events on the pathway to equine osteochondrosis?, *Equine Vet.J.* 40: 442-454.

129. Ytrehus B, Ekman S, Carlson CS, Teige J, Reinholt FP (2004), Focal changes in blood supply during normal epiphyseal growth are central in the pathogenesis of osteochondrosis in pigs, *Bone* 35: 1294-1306.

130. Olstad K, Hendrickson EH, Carlson CS, Ekman S, Dolvik NI (2013), Transection of vessels in epiphyseal cartilage canals leads to osteochondrosis and osteochondrosis dissecans in the femoro-patellar joint of foals; a potential model of juvenile osteochondritis dissecans, *Osteoarthritis.Cartilage.* 21: 730-738.

131. Ytrehus B, Andreas HH, Mellum CN, Mathisen L, Carlson CS, Ekman S, Teige J, Reinholt FP (2004), Experimental ischemia of porcine growth cartilage produces lesions of osteochondrosis, *J.Orthop.Res.* 22: 1201-1209.

132. van Weeren PR, Barneveld A (1999), The effect of exercise on the distribution and manifestation of osteochondrotic lesions in the Warmblood foal, *Equine Vet.J.Suppl* 16-25.

133. Lepeule J, Bareille N, Robert C, Ezanno P, Valette JP, Jacquet S, Blanchard G, Denoix JM, Seegers H (2009), Association of growth, feeding practices and exercise conditions with the prevalence of Developmental Orthopaedic Disease in limbs of French foals at weaning, *Prev.Vet.Med.* 89: 167-177.

134. Douglas G, Rang M (1981), The role of trauma in the pathogenesis of the osteochondroses, *Clin.Orthop.Relat Res.* 158: 28-32.

135. Campbell CJ, Ranawat CS (1966), Osteochondritis dissecans: the question of etiology, *J.Trauma* 6: 201-221.

136. Carlson CS, Cullins LD, Meuten DJ (1995), Osteochondrosis of the articular-epiphyseal cartilage complex in young horses: evidence for a defect in cartilage canal blood supply, *Vet.Pathol.* 32: 641-647.

137. Hintz HF (1987), Factors which influence developmental orthopedic disease, *Am.Assoc.Equine Practr.Proc.* 33: 159-162.

138. McIlwraith CW, American Quarter Horse Association. Summary of panel findings. In: McIlwraith CW, ed. *Proceedings Panel on Developmental Orthopedic Disease, AQHA Developmental Orthopedic Symposium.* Dallas, TX: American Quarter Horse Association, 1986; 55-61.

139. Cymbaluk NF, Smart ME (1993), A review of possible metabolic relationships of copper to equine bone disease, *Equine Vet.J.Suppl* 16: 19-26.

140. Hurtig M, Green SL, Dobson H, Mikuni-Takagaki Y, Choi J (1993), Correlative study of defective cartilage and bone growth in foals fed a low copper diet, *Equine Vet.J.Suppl* 16: 66-73.

141. Savage CJ, McCarthy RN, Jeffcott LB (1993), Effects of dietary phosphorus and calcium on induction of dyschondroplasia in foals, *Equine Vet.J.Suppl* 16: 80-83.

142. Savage CJ, McCarthy RN, Jeffcott LB (1993), Effects of dietary energy and protein on induction of dyschondroplasia in foals, *Equine Vet.J.Suppl* 16: 74-79.

143. Frantz NZ, Andrews GA, Tokach MD, Nelssen JL, Goodband RD, Derouchey JM, Dritz SS (2008), Effect of dietary nutrients on osteochondrosis lesions and cartilage properties in pigs, *Am.J.Vet.Res.* 69: 617-624.

144. Gabel AA, Knight DA, Reed SM (1987), Comparison of incidence and severity of developmental orthopedic disease on 17 farms before and after adjustment of ration, *Am.Assoc.Equine Practr.Proc.* 33: 163-170.

145. van Grevenhof EM, Heuven HCM, van Weeren PR, Bijma P (2012), The relationship between growth and osteochondrosis in specific joints in pigs, *Livestock Sci.* 143: 85-90.

146. Novotny H (1952), Osteochondrosis dissecans in two brothers; the pre- and developed state, *Acta Radiol.* 37: 493-497.

147. Pick MP (1955), Familial osteochondritis dissecans, *J.Bone Joint Surg.Br.* 37-B: 142-145.

148. Mubarak SJ, Carroll NC (1979), Familial osteochondritis dissecans of the knee, *Clin.Orthop.Relat Res.* 140: 131-136.

149. Phillips HO, Grubb SA (1985), Familial multiple osteochondritis dissecans. Report of a kindred, *J.Bone Joint Surg.Am.* 67: 155-156.

150. Stattin EL, Wiklund F, Lindblom K, Onnerfjord P, Jonsson BA, Tegner Y, Sasaki T, Struglics A, Lohmander S, Dahl N, Heinegard D, Aspberg A (2010), A missense mutation in the aggrecan C-type lectin domain disrupts extracellular matrix interactions and causes dominant familial osteochondritis dissecans, *Am.J.Hum.Genet.* 86: 126-137.

151. Tsirikos AI, Riddle EC, Kruse R (2003), Bilateral Kohler's disease in identical twins, *Clin.Orthop.Relat Res.* 409: 195-198.

152. Mackie T, Wilkins RM (2010), Case report: Osteochondritis dissecans in twins: treatment with fresh osteochondral grafts, *Clin.Orthop.Relat Res.* 468: 893-897.

153. Woods K, Harris I (1995), Osteochondritis dissecans of the talus in identical twins, *J.Bone Joint Surg.Br.* 77: 331.

154. Reiland S, Ordell N, Lundeheim N, Olsson SE (1978), Heredity of osteochondrosis, body constitution and leg weakness in the pig. A correlative investigation using progeny testing, *Acta Radiol.Suppl* 358: 123-137.

155. Fan B, Onteru SK, Mote BE, Serenius T, Stalder KJ, Rothschild MF (2009), Large-scale association study for structural soundness and leg locomotion traits in the pig, *Genet.Sel Evol.* 41: 14.

156. Wittwer C, Hamann H, Distl O (2009), The candidate gene XIRP2 at a quantitative gene locus on equine chromosome 18 associated with osteochondrosis in fetlock and hock joints of South German Coldblood horses, *J.Hered.* 100: 481-486.

157. Andersson-Eklund L, Uhlhorn H, Lundeheim N, Dalin G, Andersson L (2000), Mapping quantitative trait loci for principal components of bone measurements and osteochondrosis scores in a wild boar x large white intercross, *Genet.Res.* 75: 223-230.

158. Olstad K, Ytrehus B, Ekman S, Carlson CS, Dolvik NI (2008), Epiphyseal cartilage canal blood supply to the distal femur of foals, *Equine Vet.J.* 40: 433-439.

159. Ytrehus B, Carlson CS, Lundeheim N, Mathisen L, Reinholt FP, Teige J, Ekman S (2004), Vascularisation and osteochondrosis of the epiphyseal growth cartilage of the distal femur in pigs--development with age, growth rate, weight and joint shape, *Bone* 34: 454-465.

160. Barnewolt CE, Shapiro F, Jaramillo D (1997), Normal gadolinium-enhanced MR images of the developing appendicular skeleton: Part I. Cartilaginous epiphysis and physis, *AJR Am.J.Roentgenol.* 169: 183-189.

161. Blumer MJ, Longato S, Richter E, Perez MT, Konakci KZ, Fritsch H (2005), The role of cartilage canals in endochondral and perichondral bone formation: are there similarities between these two processes?, *J.Anat.* 206: 359-372.

162. Visco DM, Van S, Hill MA, Kincaid SA (1989), The vascular supply of the chondro-epiphyses of the elbow joint in young swine, *J.Anat.* 163: 215-229.

163. Olstad K, Ytrehus B, Ekman S, Carlson CS, Dolvik NI (2008), Epiphyseal cartilage canal blood supply to the tarsus of foals and relationship to osteochondrosis, *Equine Vet.J.* 40: 30-39.

164. Jans LB, Jaremko JL, Ditchfield M, Verstraete KL (2011), Evolution of femoral condylar ossification at MR imaging: frequency and patient age distribution, *Radiology* 258: 880-888.

165. Jans LB, Jaremko JL, Ditchfield M, Huysse WC, Verstraete KL (2011), MRI differentiates femoral condylar ossification evolution from osteochondritis dissecans. A new sign, *Eur.Radiol.* 21: 1170-1179.

166. Baker CL, III, Baker CL, Jr., Romeo AA (2010), Osteochondritis dissecans of the capitellum, *J.Shoulder.Elbow.Surg.* 19: 76-82.

167. Hixon AL, Gibbs LM (2000), Osteochondritis dissecans: a diagnosis not to miss, *Am.Fam.Physician* 61: 151-156.

168. Theiss F, Hilbe M, Furst A, Klein K, von Rechenberg B (2010), Histological evaluation of intraarticular osteochondral fragments, *Pferdeheilkunde* 26: 541-552.

169. Fisher DM, De Smet AA (1993), Radiologic analysis of osteochondritis dissecans and related osteochondral lesions, *Contemp.Diag.Rad.* 16: 1-5.

170. Thacker MM, Dabney KW, Mackenzie WG (2012), Osteochondritis dissecans of the talar head: natural history and review of literature, *J.Pediatr.Orthop.B* 21: 373-376.

171. Bui-Mansfield LT, Kline M, Chew FS, Rogers LF, Lenchik L (2000), Osteochondritis dissecans of the tibial plafond: imaging characteristics and a review of the literature, *AJR Am.J.Roentgenol.* 175: 1305-1308.

172. van Weeren PR. Osteochondrosis. In: Stick JA and Auer JA eds. *Equine Surgery.* 4th Ed. St. Louis, MO: Saunders Elsevier, 2012; 1239-1255.

173. Relave F, Meulyzer M, Alexander K, Beauchamp G, Marcoux M (2009), Comparison of radiography and ultrasonography to detect osteochondrosis lesions in the tarsocrural joint: a prospective study, *Equine Vet.J.* 41: 34-40.

174. Crijns CP, Gielen IM, Van Bree HJ, Bergman EH (2010), The use of CT and CT arthrography in diagnosing equine stifle injury in a Rheinlander gelding, *Equine Vet.J.* 42: 367-371.

175. Bojanic I, Smoljanovic T, Kubat O (2011), Osteochondritis dissecans of the first metatarsophalangeal joint: arthroscopy and microfracture technique, *J.Foot Ankle Surg.* 50: 623-625.

176. Hermanson E, Ferkel RD (2009), Bilateral osteochondral lesions of the talus, *Foot Ankle Int.* 30: 723-727.

177. Bravo C, Kawamura H, Yamaguchi T, Hotokebuchi T, Sugioka Y (1996), Experimental osteochondritis dissecans--the role of cartilage canals in chondral fractures of young rabbits, *Fukuoka Igaku Zasshi* 87: 133-141.

178. Hoppe F (1984), Radiological investigations of osteochondrosis dissecans in Standardbred Trotters and Swedish Warmblood horses, *Equine Vet.J.* 16: 425-429.

179. Siffert RS (1981), Classification of the osteochondroses, *Clin.Orthop.Relat Res.* 158: 10-18.

180. Rejno S, Stromberg B (1978), Osteochondrosis in the horse. II. Pathology, *Acta Radiol.Suppl* 358: 153-178.

181. McIlwraith CW. Clinical aspects of osteochondritis dissecans. In: McIlwraith CW and Trotter GW, eds. *Joint Disease in the Horse.* St. Louis, MO: Saunders Elsevier, 1996; 360-383.

182. McIlwraith CW, Foerner JJ, Davis DM (1991), Osteochondritis dissecans of the tarsocrural joint: results of treatment with arthroscopic surgery, *Equine Vet.J.* 23: 155-162.

183. Hoppe F, Philipsson J (1984), Tavlingsresultat hos travhastar med ostochondros i hasleden, *Svensk Vet.Tidn.* 36: 285-288.

184. Peremans K, Verschooten F (1997), Results of conservative treatment of osteochondrosis of the tibiotarsal joint in the horse, *J.Equine Vet.Sci.* 17: 322-326.

185. Physick-Sheard PW (1986), Career profile of the Canadian Standardbred. I. Influence of age, gait and sex upon chances of racing, *Can.J.Vet.Res.* 50: 449-456.

186. Physick-Sheard PW, Russell M (1986), Career profile of the Canadian Standardbred. III. Influence of temporary absence from racing and season, *Can.J.Vet.Res.* 50: 471-478.

187. R Development Core Team. R: A language and environment for statistical computing [Computer Software]. 2011. Vienna, Austria, R Foundation for Statistical computing.

188. Fox J, Weisberg S. *An R companion to applied regression*, Thousand Oaks, CA: Sage, 2011.

189. Venables WN, Ripley BD. *Modern applied statistics with R*, New York: Springer, 2002.

190. Elias I, Jung JW, Raikin SM, Schweitzer MW, Carrino JA, Morrison WB (2006), Osteochondral lesions of the talus: change in MRI findings over time in talar lesions without operative intervention and implications for staging systems, *Foot Ankle Int.* 27: 157-166.

191. Smith MM, Vasseur PB, Morgan JP (1985), Clinical evaluation of dogs after surgical and nonsurgical management of osteochondritis dissecans of the talus, *J.Am.Vet.Med.Assoc.* 187: 31-35.

192. Jacquet S, Robert C, Valette JP, Denoix JM (2013), Evolution of radiological findings detected in the limbs of 321 young horses between the ages of 6 and 18 months, *Vet.J.* 197: 58-64.

193. Dik KJ, Enzerink E, van Weeren PR (1999), Radiographic development of osteochondral abnormalities, in the hock and stifle of Dutch Warmblood foals, from age 1 to 11 months, *Equine Vet.J.Suppl* 9-15.

194. Brink P, Dolvik NI, Tverdal A (2009), Lameness and effusion of the tarsocrural joints after arthroscopy of osteochondritis dissecans in horses, *Vet.Rec.* 165: 709-712.

195. McIlwraith CW (2013), Surgical versus conservative management of osteochondrosis, *Vet.J.* 197: 19-28.

196. Hernandez J, Hawkins DL (2001), Training failure among yearling horses, *Am.J.Vet.Res.* 62: 1418-1422.

197. Drevemo S, Dalin G, Fredricson I, Hjerten G (1980), Equine locomotion; 1. The analysis of linear and temporal stride characteristics of trotting standardbreds, *Equine Vet.J.* 12: 60-65.

198. Wilson BD, Neal RJ, Howard A, Groenendyk S (1988), The gait of pacers. 1: kinematics of the racing stride, *Equine Vet.J.* 20: 341-346.

199. Wilson BD, Neal RJ, Howard A, Groenedyk S (1988), The gait of pacers. 2: factors influencing pacing speed, *Equine Vet.J.* 20: 347-351.

200. Robilliard JJ, Pfau T, Wilson AM (2007), Gait characterisation and classification in horses, *J.Exp.Biol.* 210: 187-197.

201. Pool RR. Pathogenesis of developmental lesions. In: McIlwraith CW ed. *Proceedings Panel on Developmental Orthopedic Disease, AQHA Developmental Orthopedic Symposium.* Dallas, TX: American Quarter Horse Association, 1986; 22-25.

202. Lepeule J, Bareille N, Robert C, Valette JP, Jacquet S, Blanchard G, Denoix JM, Seegers H (2013), Association of growth, feeding practices and exercise conditions with the severity of the osteoarticular status of limbs in French foals, *Vet.J.* 197: 65-71.

203. van Weeren PR, Denoix JM (2013), The Normandy field study on juvenile osteochondral conditions: conclusions regarding the influence of genetics, environmental conditions and management, and the effect on performance, *Vet.J.* 197: 90-95.

204. Wittwer C, Hamann H, Rosenberger E, Distl O (2007), Genetic parameters for the prevalence of osteochondrosis in the limb joints of South German Coldblood horses, *J.Anim Breed.Genet.* 124: 302-307.

205. Jonsson L, Dalin G, Egenvall A, Nasholm A, Roepstorff L, Philipsson J (2011), Equine hospital data as a source for study of prevalence and heritability of osteochondrosis and palmar/plantar osseous fragments of Swedish Warmblood horses, *Equine Vet.J.* 43: 695-700.

206. McCue ME, Valberg SJ, Miller MB, Wade C, DiMauro S, Akman HO, Mickelson JR (2008), Glycogen synthase (GYS1) mutation causes a novel skeletal muscle glycogenosis, *Genomics* 91: 458-466.

207. McCue ME, Valberg SJ, Lucio M, Mickelson JR (2008), Glycogen synthase 1 (GYS1) mutation in diverse breeds with polysaccharide storage myopathy, *J.Vet.Intern.Med.* 22: 1228-1233.

208. Wittwer C, Lohring K, Drogemuller C, Hamann H, Rosenberger E, Distl O (2007), Mapping quantitative trait loci for osteochondrosis in fetlock and hock joints and palmar/plantar osseus fragments in fetlock joints of South German Coldblood horses, *Anim Genet.* 38: 350-357.

209. Dierks C, Komm K, Lampe V, Distl O (2010), Fine mapping of a quantitative trait locus for osteochondrosis on horse chromosome 2, *Anim Genet.* 41 Suppl 2: 87-90.

210. Lampe V, Dierks C, Distl O (2009), Refinement of a quantitative gene locus on equine chromosome 16 responsible for osteochondrosis in Hanoverian warmblood horses, *Animal.* 3: 1224-1231.

211. Browning SR, Browning BL (2007), Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering, *Am.J.Hum.Genet.* 81: 1084-1097.

212. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007), PLINK: a tool set for whole-genome association and population-based linkage analyses, *Am.J.Hum.Genet.* 81: 559-575.

213. Thornton T, McPeek MS (2010), ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure, *Am.J.Hum.Genet.* 86: 172-184.

214. Zhou X, Stephens M (2012), Genome-wide efficient mixed-model analysis for association studies, *Nat.Genet.* 44: 821-824.

215. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010), Variance component model to account for sample structure in genome-wide association studies, *Nat.Genet.* 42: 348-354.

216. Yu J, Pressoir G, Briggs WH, Vroh B, I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006), A unified mixed-model method for association mapping that accounts for multiple levels of relatedness, *Nat.Genet.* 38: 203-208.

217. Zhang Z, Todhunter RJ, Buckler ES, Van Vleck LD (2007), Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood, *J.Anim Sci.* 85: 881-885.

218. Romer P, Weingartner J, Desaga B, Kubein-Meesenburg D, Reicheneder C, Proff P (2012), Effect of excessive methionine on the development of the cranial growth plate in newborn rats, *Arch.Oral Biol.* 57: 1225-1230.

219. Blewett HJH (2008), Exploring the mechanisms behind S-adenosylmethionine (SAMe) on the treatment of osteoarthritis, *Crit.Rev.Food Sci.Nutr.* 48: 458-463.

220. Dowthwaite GP, Edwards JC, Pitsillides AA (1998), An essential role for the interaction between hyaluronan and hyaluronan binding proteins during joint development, *J.Histochem.Cytochem.* 46: 641-651.

221. Aikawa T, Segre GV, Lee K (2001), Fibroblast growth factor inhibits chondrocytic growth through induction of p21 and subsequent inactivation of cyclin E-Cdk2, *J.Biol.Chem.* 276: 29347-29352

222. Wu G, Chen W, Fan H, Zheng C, Chu J, Lin R, Ye J, Xu H, Li X, Huang Y, Ye H, Liu X, Wu M (2013), Duhuo Jisheng Decoction promotes chondrocyte proliferation through accelerated G1/S transition in osteoarthritis, *Int.J.Mol.Med.* 32: 1001-1010.

223. Lowenheim H, Reichl J, Winter H, Hahn H, Simon C, Gultig K, Muller A, Zenner HP, Zimmermann U, Knipper M (2005), In vitro expansion of human nasoseptal chondrocytes reveals distinct expression profiles of G1 cell cycle inhibitors for replicative, quiescent, and senescent culture stages, *Tissue Eng* 11: 64-75.

224. Duarte C, Kobayashi Y, Kawamoto T, Moriyama K (2014), Relaxin receptors 1 and 2 and nuclear receptor subfamily 3, group C, member 1 (glucocorticoid receptor) mRNAs are expressed in oral components of developing mice, *Arch.Oral Biol.* 59: 111-118.

225. Beier F, Loeser RF (2010), Biology and pathology of Rho GTPase, PI-3 kinase-Akt, and MAP kinase signaling pathways in chondrocytes, *J.Cell Biochem.* 110: 573-580.

226. Krejci P, Krakow D, Mekikian PB, Wilcox WR (2007), Fibroblast growth factors 1, 2, 17, and 19 are the predominant FGF ligands expressed in human fetal growth plate cartilage, *Pediatr.Res.* 61: 267-272.

227. Razidlo DF, Whitney TJ, Casper ME, Gee-Lawrence ME, Stensgard BA, Li X, Secreto FJ, Knutson SK, Hiebert SW, Westendorf JJ (2010), Histone deacetylase 3 depletion in osteo/chondroprogenitor cells decreases bone density and increases marrow fat, *PLoS.One.* 5: e11492.

228. Bradley EW, Carpio LR, Westendorf JJ (2013), Histone deacetylase 3 suppression increases PH domain and leucine-rich repeat phosphatase (Phlpp)1 expression in chondrocytes to suppress Akt signaling and matrix secretion, *J.Biol.Chem.* 288: 9572-9582.

229. Lauder RM, Huckerby TN, Nieduszynski IA (1996), The structure of the keratan sulphate chains attached to fibromodulin isolated from articular cartilage, *Eur.J.Biochem.* 242: 402-409.

230. Edvardson S, Ashikov A, Jalas C, Sturiale L, Shaag A, Fedick A, Treff NR, Garozzo D, Gerardy-Schahn R, Elpeleg O (2013), Mutations in SLC35A3 cause autism spectrum disorder, epilepsy and arthrogryposis, *J.Med.Genet.* 50: 733-739.

231. Thomsen B, Horn P, Panitz F, Bendixen E, Petersen AH, Holm LE, Nielsen VH, Agerholm JS, Arnbjerg J, Bendixen C (2006), A missense mutation in the bovine SLC35A3 gene, encoding a UDP-N-acetylglucosamine transporter, causes complex vertebral malformation, *Genome Res.* 16: 97-105.

232. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P (2009), Genotype-imputation accuracy across worldwide human populations, *Am.J.Hum.Genet.* 84: 235-250.

233. Ventura R, Schenkel F, Sargolzaei M et al. Accuracy of imputation to high density SNP data in multibreed beef cattle. *PAG XXI.* 2013; P0529.

234. Howie B, Marchini J, Stephens M (2011), Genotype imputation with thousands of genomes, *G3.(Bethesda.)* 1: 457-470.

235. Petersen JL, Mickelson JR, Cothran EG, Andersson LS, Axelsson J, Bailey E, Bannasch D, Binns MM, Borges AS, Brama P, da Camara MA, Distl O, Felicetti M, Fox-Clipsham L, Graves KT, Guerin G, Haase B, Hasegawa T, Hemmann K, Hill EW, Leeb T, Lindgren G, Lohi H, Lopes MS, McGivney BA, Mikko S, Orr N, Penedo MC, Piercy RJ, Raekallio M, Rieder S, Roed KH, Silvestrelli M, Swinburne J, Tozaki T, Vaudin M, Wade M, McCue ME (2013), Genetic diversity in the modern horse illustrated from genome-wide SNP data, *PLoS.One.* 8: e54997.

236. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I (2009), Whole population, genome-wide mapping of hidden relatedness, *Genome Res.* 19: 318-326.

237. Li L, Li Y, Browning SR, Browning BL, Slater AJ, Kong X, Aponte JL, Mooser VE, Chissoe SL, Whittaker JC, Nelson MR, Ehm MG (2011), Performance of genotype imputation for rare variants identified in exons and flanking regions of genes, *PLoS.One.* 6: e24945.

238. Brunberg E, Andersson L, Cothran G, Sandberg K, Mikko S, Lindgren G (2006), A missense mutation in PMEL17 is associated with the Silver coat color in the horse, *BMC.Genet.* 7: 46.

239. Locke MM, Penedo MC, Bricker SJ, Millon LV, Murray JD (2002), Linkage of the grey coat colour locus to microsatellites on horse chromosome 25, *Anim Genet.* 33: 329-337.

240. Locke MM, Ruth LS, Millon LV, Penedo MC, Murray JD, Bowling AT (2001), The cream dilution gene, responsible for the palomino and buckskin coat colours, maps to horse chromosome 21, *Anim Genet.* 32: 340-343.

241. Swinburne JE, Hopkins A, Binns MM (2002), Assignment of the horse grey coat colour gene to ECA25 using whole genome scanning, *Anim Genet.* 33: 338-342.

242. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011), Low-coverage sequencing: implications for design of complex trait association studies, *Genome Res.* 21: 940-951.

243. Kim SY, Li Y, Guo Y, Li R, Holmkvist J, Hansen T, Pedersen O, Wang J, Nielsen R (2010), Design of association studies with pooled or un-pooled next-generation sequencing data, *Genet.Epidemiol.* 34: 479-491.

244. Menelaou A, Marchini J (2013), Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold, *Bioinformatics.* 29: 84-91.

245. Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV (2012), Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare, *BMC.Genomics* 13: 78.

246. Towers RE, Murgiano L, Millar DS, Glen E, Topf A, Jagannathan V, Drogemuller C, Goodship JA, Clarke AJ, Leeb T (2013), A Nonsense Mutation in the IKBKG Gene in Mares with Incontinentia Pigmenti, *PLoS.One.* 8: e81625.

247. Scheet P, Stephens M (2006), A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, *Am.J.Hum.Genet.* 78: 629-644.

248. Barrett JC, Fry B, Maller J, Daly MJ (2005), Haploview: analysis and visualization of LD and haplotype maps, *Bioinformatics.* 21: 263-265.

249. Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del AG, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011), A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat.Genet.* 43: 491-498.

250. The ENCODE Project Consortium (2012), An integrated encyclopedia of DNA elements in the human genome, *Nature* 489: 57-74.

251. Fukuoka Y, Hagihara M, Nagatsu T, Kaneda T (1993), The relationship between collagen metabolism and temporomandibular joint osteoarthrosis in mice, *J.Oral Maxillofac.Surg.* 51: 288-291.

252. Niedermeyer J, Garin-Chesa P, Kriz M, Hilberg F, Mueller E, Bamberger U, Rettig WJ, Schnapp A (2001), Expression of the fibroblast activation protein during mouse embryo development, *Int.J.Dev.Biol.* 45: 445-447.

253. Sadanandam A, Rosenbaugh EG, Singh S, Varney M, Singh RK (2010), Semaphorin 5A promotes angiogenesis by increasing endothelial cell proliferation, migration, and decreasing apoptosis, *Microvasc.Res.* 79: 1-9.

254. McCoy AM, Toth F, Dolvik NI, Ekman S, Ellermann J, Olstad K, Ytrehus B, Carlson CS (2013), Articular osteochondrosis: a comparison of naturally-occurring human and animal disease, *Osteoarthritis.Cartilage.* 21: 1638-1647.

255. Macsai CE, Georgiou KR, Foster BK, Zannettino AC, Xian CJ (2012), Microarray expression analysis of genes and pathways involved in growth plate cartilage injury responses and bony repair, *Bone* 50: 1081-1091.

256. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005), Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proc.Natl.Acad.Sci.U.S.A* 102: 15545-15550.

257. Lee D, Lee GK, Yoon KA, Lee JS (2013), Pathway-based analysis using genome-wide association data from a Korean non-small cell lung cancer study, *PLoS.One.* 8: e65396.

258. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van EP (2005), Identifying SNPs predictive of phenotype using random forests, *Genet.Epidemiol.* 28: 171-182.

259. Pan Q, Hu T, Malley JD, Andrew AS, Karagas MR, Moore JH (2014), A system-level pathway-phenotype association analysis using synthetic feature random forest, *Genet.Epidemiol.* 38: 209-219.

260. Yao C, Spurlock DM, Armentano LE, Page CD, Jr., VandeHaar MJ, Bickhart DM, Weigel KA (2013), Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle, *J.Dairy Sci.* 96: 6716-6729.

261. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham DL, Bryant J, Wolmark N (2004), A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer, *N.Engl.J.Med.* 351: 2817-2826.

262. Hendricks BL. *International Encyclopedia of Horse Breeds*. Norman, OK: University of Oklahoma Press, 2007; 208-210.

263. Albertsdottir E, Eriksson S, Sigurdsson A, Arnason T (2011), Genetic analysis of 'breeding field test status' in Icelandic horses, *J.Anim Breed.Genet.* 128: 124-132.

264. Kim S, Kettlewell JR, Anderson RC, Bardwell VJ, Zarkower D (2003), Sexually dimorphic expression of multiple doublesex-related genes in the embryonic mouse gonad, *Gene Expr.Patterns.* 3: 77-82.

265. Promerova M, Andersson LS, Juras R, Penedo MC, Reissmann M, Tozaki T, Bellone R, Dunner S, Horin P, Imsland F, Imsland P, Mikko S, Modry D, Roed KH, Schwochow D, Vega-Pla JL, Mehrabani-Yeganeh H, Yousefi-Mashouf N, Cothran G, Lindgren G, Andersson L (2014), Worldwide frequency distribution of the 'Gait keeper' mutation in the DMRT3 gene, *Anim Genet.* 45: 274-282.

266. Anney R, Klei L, Pinto D, Regan R, Conroy J, Magalhaes TR, Correia C, Abrahams BS, Sykes N, Pagnamenta AT, Almeida J, Bacchelli E, Bailey AJ, Baird G, Battaglia A, Berney T, Bolshakova N, Bolte S, Bolton PF, Bourgeron T, Brennan S, Brian J, Carson AR, Casallo G, Casey J, Chu SH, Cochrane L, Corsello C, Crawford EL, Crossett A, Dawson G, de JM, Delorme R, Drmic I, Duketis E, Duque F, Estes A, Farrar P, Fernandez BA, Folstein SE, Fombonne E, Freitag CM, Gilbert J, Gillberg C, Glessner JT, Goldberg J, Green J, Guter SJ, Hakonarson H, Heron EA, Hill M, Holt R, Howe JL, Hughes G, Hus V, Igliozzi R, Kim C, Klauck SM, Kolevzon A, Korvatska O, Kustanovich V, Lajonchere CM, Lamb JA, Laskawiec M, Leboyer M, Le CA, Leventhal BL, Lionel AC, Liu

XQ, Lord C, Lotspeich L, Lund SC, Maestrini E, Mahoney W, Mantoulan C, Marshall CR, McConachie H, McDougle CJ, McGrath J, McMahon WM, Melhem NM, Merikangas A, Migita O, Minshew NJ, Mirza GK, Munson J, Nelson SF, Noakes C, Noor A, Nygren G, Oliveira G, Papanikolaou K, Parr JR, Parrini B, Paton T, Pickles A, Piven J, Posey DJ, Poustka A, Poustka F, Prasad A, Ragoussis J, Renshaw K, Rickaby J, Roberts W, Roeder K, Roge B, Rutter ML, Bierut LJ, Rice JP, Salt J, Sansom K, Sato D, Segurado R, Senman L, Shah N, Sheffield VC, Soorya L, Sousa I, Stoppioni V, Strawbridge C, Tancredi R, Tansey K, Thiruvahindrapduram B, Thompson AP, Thomson S, Tryfon A, Tsiantis J, Van EH, Vincent JB, Volkmar F, Wallace S, Wang K, Wang Z, Wassink TH, Wing K, Wittemeyer K, Wood S, Yaspan BL, Zurawiecki D, Zwaigenbaum L, Betancur C, Buxbaum JD, Cantor RM, Cook EH, Coon H, Cuccaro ML, Gallagher L, Geschwind DH, Gill M, Haines JL, Miller J, Monaco AP, Nurnberger JI, Jr., Paterson AD, Pericak-Vance MA, Schellenberg GD, Scherer SW, Sutcliffe JS, Szatmari P, Vicente AM, Vieland VJ, Wijsman EM, Devlin B, Ennis S, Hallmayer J (2010), A genome-wide scan for common alleles affecting risk for autism, *Hum.Mol.Genet.* 19: 4072-4082.

267. Oedegaard KJ, Greenwood TA, Johansson S, Jacobsen KK, Halmoy A, Fasmer OB, Akiskal HS, Haavik J, Kelsoe JR (2010), A genome-wide association study of bipolar disorder and comorbid migraine, *Genes Brain Behav.* 9: 673-680.

268. Asahina H, Masuba A, Hirano S, Yuri K (2012), Distribution of protocadherin 9 protein in the developing mouse nervous system, *Neuroscience* 225: 88-104.

269. Janisch KM, Vock VM, Fleming MS, Shrestha A, Grimsley-Myers CM, Rasoul BA, Neale SA, Cupp TD, Kinchen JM, Liem KF, Jr., Dwyer ND (2013), The vertebrate-specific Kinesin-6, Kif20b, is required for normal cytokinesis of polarized cortical stem cells and cerebral cortex size, *Development* 140: 4672-4682.

270. Sapir T, Levy T, Sakakibara A, Rabinkov A, Miyata T, Reiner O (2013), Shootin1 acts in concert with KIF20B to promote polarization of migrating neurons, *J.Neurosci.* 33: 11932-11948.

271. Demonbreun AR, Lapidos KA, Heretis K, Levin S, Dale R, Pytel P, Svensson EC, McNally EM (2010), Myoferlin regulation by NFAT in muscle injury, regeneration and repair, *J.Cell Sci.* 123: 2413-2422.

272. Li R, Ackerman WE, Mihai C, Volakis LI, Ghadiali S, Kniss DA (2012), Myoferlin depletion in breast cancer cells promotes mesenchymal to epithelial shape change and stalls invasion, *PLoS.One.* 7: e39766.

273. Eisenberg MC, Kim Y, Li R, Ackerman WE, Kniss DA, Friedman A (2011), Mechanistic modeling of the effects of myoferlin on tumor cell invasion, *Proc.Natl.Acad.Sci.U.S.A* 108: 20078-20083.

274. Zhang W, Hui KY, Gusev A, Warner N, Ng SM, Ferguson J, Choi M, Burberry A, Abraham C, Mayer L, Desnick RJ, Cardinale CJ, Hakonarson H, Waterman M, Chowers Y, Karban A, Brant SR, Silverberg MS, Gregersen PK, Katz S, Lifton RP, Zhao H, Nunez G, Pe'er I, Peter I, Cho JH (2013), Extended haplotype association study in Crohn's disease identifies a novel, Ashkenazi Jewish-specific missense mutation in the NF-kappaB pathway gene, HEATR3, *Genes Immun.* 14: 310-316.

275. Grupe A, Li Y, Rowland C, Nowotny P, Hinrichs AL, Smemo S, Kauwe JS, Maxwell TJ, Cherny S, Doil L, Tacey K, van LR, Myers A, Wavrant-De VF, Kaleem M, Hollingworth P, Jehu L, Foy C, Archer N, Hamilton G, Holmans P, Morris CM, Catanese J, Sninsky J, White TJ, Powell J, Hardy J, O'Donovan M, Lovestone S, Jones L, Morris JC, Thal L, Owen M, Williams J, Goate A (2006), A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease, *Am.J.Hum.Genet.* 78: 78-88.

276. Kuniba H, Yoshiura K, Kondoh T, Ohashi H, Kurosawa K, Tonoki H, Nagai T, Okamoto N, Kato M, Fukushima Y, Kaname T, Naritomi K, Matsumoto T, Moriuchi H, Kishino T, Kinoshita A, Miyake N, Matsumoto N, Niikawa N (2009), Molecular karyotyping in 17 patients and mutation screening in 41 patients with Kabuki syndrome, *J.Hum.Genet.* 54: 304-309.

277. Snel B, Lehmann G, Bork P, Huynen MA (2000), STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene, *Nucleic Acids Res.* 28: 3442-3444.

278. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von MC, Jensen LJ (2013), STRING v9.1: protein-protein interaction networks, with increased coverage and integration, *Nucleic Acids Res.* 41: D808-D815.

279. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J (2009), ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, *Bioinformatics.* 25: 1091-1093.

280. Raychaudhuri S, Plenge RM, Rossin EJ, Ng AC, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ (2009), Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions, *PLoS.Genet.* 5: e1000534.

281. Foulkes AS. *Applied Statistical Genetics with R*. New York: Springer, 2009; 157-198.

282. Hamond C, Martins G, Lilenbaum W (2012), Subclinical leptospirosis may impair athletic performance in racing horses, *Trop.Anim Health Prod.* 44: 1927-1930.

283. Young LE, Rogers K, Wood JL (2008), Heart murmurs and valvular regurgitation in thoroughbred racehorses: epidemiology and associations with athletic performance, *J.Vet.Intern.Med.* 22: 418-426.

284. Witte TH, Mohammed HO, Radcliffe CH, Hackett RP, Ducharme NG (2009), Racing performance after combined prosthetic laryngoplasty and ipsilateral ventriculocordectomy or partial arytenoidectomy: 135 Thoroughbred racehorses competing at less than 2400 m (1997-2007), *Equine Vet.J.* 41: 70-75.

285. Reardon RJ, Fraser BS, Heller J, Lischer C, Parkin T, Bladon BM (2008), The use of race winnings, ratings and a performance index to assess the effect of thermocautery of the soft palate for treatment of horses with suspected intermittent dorsal displacement. A case-control study in 110 racing Thoroughbreds, *Equine Vet.J.* 40: 508-513.

286. Woodie JB, Ducharme NG, Kanter P, Hackett RP, Erb HN (2005), Surgical advancement of the larynx (laryngeal tie-forward) as a treatment for dorsal displacement of the soft palate in horses: a prospective study 2001-2004, *Equine Vet.J.* 37: 418-423.

287. Barnes AJ, Slone DE, Lynch TM (2004), Performance after partial arytenoidectomy without mucosal closure in 27 Thoroughbred racehorses, *Vet.Surg.* 33: 398-403.

288. Strand E, Martin GS, Haynes PF, McClure JR, Vice JD (2000), Career racing performance in Thoroughbreds treated with prosthetic laryngoplasty for laryngeal neuropathy: 52 cases (1981-1989), *J.Am.Vet.Med.Assoc.* 217: 1689-1696.

289. Tetens J, Ross MW, Lloyd JW (1997), Comparison of racing performance before and after treatment of incomplete, midsagittal fractures of the proximal phalanx in standardbreds: 49 cases (1986-1992), *J.Am.Vet.Med.Assoc.* 210: 82-86.

290. Schnabel LV, Bramlage LR, Mohammed HO, Embertson RM, Ruggles AJ, Hopper SA (2007), Racing performance after arthroscopic removal of apical sesamoid fracture fragments in Thoroughbred horses age < 2 years: 151 cases (1989--2002), *Equine Vet.J.* 39: 64-68.

291. Martin GS, Strand E, Kearney MT (1996), Use of statistical models to evaluate racing performance in thoroughbreds, *J.Am.Vet.Med.Assoc.* 209: 1900-1906.

292. Martin GS, Strand E, Kearney MT (1997), Validation of a regression model for standardizing lifetime racing performances of thoroughbreds, *J.Am.Vet.Med.Assoc.* 210: 1641-1645.

293. Timeform. How Timeform handicaps horses. Available at: https://www.timeform.com/Racing/Articles/How_Timeform_handicaps_horses. Accessed 20 Mar 2014.

294. Stock KF, Reents R (2013), Genomic selection: Status in different species and challenges for breeding, *Reprod.Domest.Anim* 48 Suppl 1: 2-10.

295. Vandenplas J, Janssens S, Buys N, Gengler N (2013), An integration of external information for foreign stallions into the Belgian genetic evaluation for jumping horses, *J.Anim Breed.Genet.* 130: 209-217.

296. Viklund A, Nasholm A, Strandberg E, Philipsson J (2010), Effects of long-time series of data on genetic evaluations for performance of Swedish Warmblood riding horses, *Animal.* 4: 1823-1831.

297. Viklund A, Braam A, Nasholm A, Strandberg E, Philipsson J (2010), Genetic variation in competition traits at different ages and time periods and correlations with traits at field tests of 4-year-old Swedish Warmblood horses, *Animal.* 4: 682-691.

298. Correa MJ, da M (2007), Genetic evaluation of performance traits in Brazilian Quarter Horse, *J.Appl.Genet.* 48: 145-151.

299. Mota MD, Abrahao AR, Oliveira HN (2005), Genetic and environmental parameters for racing time at different distances in Brazilian Thoroughbreds, *J.Anim Breed.Genet.* 122: 393-399.

300. Pallerla SR, Lawrence R, Lewejohann L, Pan Y, Fischer T, Schlomann U, Zhang X, Esko JD, Grobe K (2008), Altered heparan sulfate structure in mice with deleted NDST3 gene function, *J.Biol.Chem.* 283: 16885-16894.

301. Teplyuk NM, Haupt LM, Ling L, Dombrowski C, Mun FK, Nathan SS, Lian JB, Stein JL, Stein GS, Cool SM, van Wijnen AJ (2009), The osteogenic transcription factor Runx2 regulates components of the fibroblast growth factor/proteoglycan signaling axis in osteoblasts, *J.Cell Biochem.* 107: 144-154.

302. Failly M, Bartoloni L, Letourneau A, Munoz A, Falconnet E, Rossier C, de Santi MM, Santamaria F, Sacco O, Lozier-Blanchet CD, Lazor R, Blouin JL (2009), Mutations in DNAH5 account for only 15% of a non-preselected cohort of patients with primary ciliary dyskinesia, *J.Med.Genet.* 46: 281-286.

303. Singh S, Canseco DC, Manda SM, Shelton JM, Chirumamilla RR, Goetsch SC, Ye Q, Gerard RD, Schneider JW, Richardson JA, Rothermel BA, Mammen PP (2014), Cytoglobin modulates myogenic progenitor cell viability and muscle regeneration, *Proc.Natl.Acad.Sci.U.S.A* 111: E129-E138.

304. Di GC, Zara S, De CM, Ruffini R, Porzionato A, Macchi V, De CR, Cataldi A (2013), Cytoglobin and neuroglobin in the human brainstem and carotid body, *Adv.Exp.Med.Biol.* 788: 59-64.

305. Tian SF, Yang HH, Xiao DP, Huang YJ, He GY, Ma HR, Xia F, Shi XC (2013), Mechanisms of neuroprotection from hypoxia-ischemia (HI) brain injury by up-

regulation of cytoglobin (CYGB) in a neonatal rat model, *J.Biol.Chem.* 288: 15988-16003.

306. Hill EW, McGivney BA, Gu J, Whiston R, Machugh DE (2010), A genome-wide SNP-association study confirms a sequence variant (g.66493737C>T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses, *BMC.Genomics* 11: 552.

307. Dhar SS, Liang HL, Wong-Riley MT (2009), Nuclear respiratory factor 1 co-regulates AMPA glutamate receptor subunit 2 and cytochrome c oxidase: tight coupling of glutamatergic transmission and energy metabolism in neurons, *J.Neurochem.* 108: 1595-1606.

308. Mendez-Villanueva A, Buchheit M, Simpson B, Peltola E, Bourdon P (2011), Does on-field sprinting performance in young soccer players depend on how fast they can run or how fast they do run?, *J.Strength.Cond.Res.* 25: 2634-2638.

309. Wilson RS, Niehaus AC, David G, Hunter A, Smith M (2014), Does individual quality mask the detection of performance trade-offs? A test using analyses of human physical performance, *J.Exp.Biol.* 217: 545-551.

310. Buchheit M, Mendez-Villanueva A (2013), Reliability and stability of anthropometric and performance measures in highly-trained young soccer players: effect of age and maturation, *J.Sports Sci.* 31: 1332-1343.

311. Hayes M, Smith D, Castle PC, Watt PW, Ross EZ, Maxwell NS (2013), Peak power output provides the most reliable measure of performance in prolonged intermittent-sprint cycling, *J.Sports Sci.* 31: 565-572.

312. Ziogas GG, Patras KN, Stergiou N, Georgoulis AD (2011), Velocity at lactate threshold and running economy must also be considered along with maximal oxygen uptake when testing elite soccer players during preseason, *J.Strength.Cond.Res.* 25: 414-419.

313. Castagna C, Impellizzeri FM, Chaouachi A, Manzi V (2013), Preseason variations in aerobic fitness and performance in elite-standard soccer players: a team study, *J.Strength.Cond.Res.* 27: 2959-2965.

314. Munsters CC, van den BJ, van WR, Sloet van Oldruitenborgh-Oosterbaan MM (2013), Young Friesian horses show familial aggregation in fitness response to a 7-week performance test, *Vet.J.* 198: 193-199.

315. Bitschnau C, Wiestner T, Trachsel DS, Auer JA, Weishaupt MA (2010), Performance parameters and post exercise heart rate recovery in Warmblood sports horses of different performance levels, *Equine Vet.J.Suppl* 42(S38): 17-22.

316. Courouce A, Chatard JC, Auvinet B (1997), Estimation of performance potential of standardbred trotters from blood lactate concentrations measured in field conditions, *Equine Vet.J.* 29: 365-369.

317. Rudolph JA, Spier SJ, Byrns G, Rojas CV, Bernoco D, Hoffman EP (1992), Periodic paralysis in quarter horses: a sodium channel mutation disseminated by selective breeding, *Nat.Genet.* 2: 144-147.

318. Rieder S, Taourit S, Mariat D, Langlois B, Guerin G (2001), Mutations in the agouti (ASIP), the extension (MC1R), and the brown (TYRP1) loci and their association to coat color phenotypes in horses (Equus caballus), *Mamm.Genome* 12: 450-455.

319. Matsumoto H, Sasaki K, Bessho T, Kobayashi E, Abe T, Sasazaki S, Oyama K, Mannen H (2012), The SNPs in the ACACA gene are effective on fatty acid composition in Holstein milk, *Mol.Biol.Rep.* 39: 8637-8644.

320. Estany J, Tor M, Villalba D, Bosch L, Gallardo D, Jimenez N, Altet L, Noguera JL, Reixach J, Amills M, Sanchez A (2007), Association of CA repeat polymorphism at intron 1 of insulin-like growth factor (IGF-I) gene with circulating IGF-I concentration, growth, and fatness in swine, *Physiol Genomics* 31: 236-243.

321. Wang K, Li M, Hakonarson H (2010), Analysing biological pathways in genome-wide association studies, *Nat.Rev.Genet.* 11: 843-854.

322. Song S, Black MA (2008), Microarray-based gene set analysis: a comparison of current methods, *BMC.Bioinformatics.* 9: 502.

323. Hedegaard J, Arce C, Bicciato S, Bonnet A, Buitenhuis B, Collado-Romero M, Conley LN, Sancristobal M, Ferrari F, Garrido JJ, Groenen MA, Hornshoj H, Hulsegge I, Jiang L, Jimenez-Marin A, Kommadath A, Lagarrigue S, Leunissen JA, Liaubet L, Neerincx PB, Nie H, van der PJ, Prickett D, Ramirez-Boo M, Rebel JM, Robert-Granie C, Skarman A, Smits MA, Sorensen P, Tosser-Klopp G, Watson M (2009), Methods for interpreting lists of affected genes obtained in a DNA microarray experiment, *BMC.Proc.* 3 Suppl 4: S5.

324. Wang K, Edmondson AC, Li M, Gao F, Qasim AN, Devaney JM, Burnett MS, Waterworth DM, Mooser V, Grant SF, Epstein SE, Reilly MP, Hakonarson H, Rader DJ (2011), Pathway-Wide Association Study Implicates Multiple Sterol Transport and Metabolism Genes in HDL Cholesterol Regulation, *Front Genet.* 2: 41.

325. Sharma A, Gulbahce N, Pevzner SJ, Menche J, Ladenvall C, Folkersen L, Eriksson P, Orho-Melander M, Barabasi AL (2013), Network-based analysis of genome wide association data provides novel candidate genes for lipid and lipoprotein traits, *Mol.Cell Proteomics.* 12: 3398-3408.

326. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM (2014), Ensembl 2014, *Nucleic Acids Res.* 42: D749-D755.

327. Kalbfleisch T, Rebolledo-Mendez J, Orlando L et al. Resources, and progress towards a fully annotated EquCab3. *Plant & Animal Genome XXII.* 2014;W279.

328. Li Y, Willer C, Sanna S, Abecasis G (2009), Genotype imputation, *Annu.Rev.Genomics Hum.Genet.* 10: 387-406.

329. McCoy AM, McCue ME (2014), Validation of imputation between equine genotyping arrays, *Anim Genet.* 45: 153.

330. Browning BL, Browning SR (2013), Improving the accuracy and efficiency of identity-by-descent detection in population data, *Genetics* 194: 459-471.

331. Haberland M, Montgomery RL, Olson EN (2009), The many roles of histone deacetylases in development and physiology: implications for disease and therapy, *Nat.Rev.Genet.* 10: 32-42.

332. Vega RB, Matsuda K, Oh J, Barbosa AC, Yang X, Meadows E, McAnally J, Pomajzl C, Shelton JM, Richardson JA, Karsenty G, Olson EN (2004), Histone deacetylase 4 controls chondrocyte hypertrophy during skeletogenesis, *Cell* 119: 555-566.

333. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009), The Sequence Alignment/Map format and SAMtools, *Bioinformatics.* 25: 2078-2079.

Appendix 1

Sequencing Histone Deacetylase 4 (*HDAC4*)

Background

As reported in **Chapter 4**, in the genome-wide association study (GWAS) for osteochondrosis (OC) performed in the initial cohort of 94 horses, a region of association was identified on ECA6 from ~24.3-24.7Mb. This region contained two named genes, neither of which seemed to be an obvious candidate for contributing to OC risk. However, an excellent candidate gene, *HDAC4*, was located just outside of this region, at ~24.8-25.1Mb.

The HDAC superfamily is composed of 11 highly conserved proteins broken into 4 classes. Class IIa HDACs, of which *HDAC4* is an example, are characterized by relatively restricted expression patterns and have large N-terminal extensions with conserved binding sites for transcription factors (e.g. MEF2 [myocyte enhancer factor 2] and the chaperone protein 14-3-3).[331] *HDAC4* is the only member of its class to be expressed in cartilage and has been demonstrated to play a crucial role in normal endochondral ossification.[331;332] *HDAC4* knockout mice are markedly smaller than their wild-type littermates with stunted appendage growth due to global premature ossification, and in fact, die in the neonatal period because their ribcages ossify to the point that they cannot breathe. In contrast, exogenous *HDAC4* overexpression leads to a failure of normal ossification. Interestingly, the flat bones of knockout mice (e.g. skull), which do not undergo endochondral ossification, developed normally in this model. Expression of *HDAC4* was highest in the prehypertrophic and hypertrophic cartilage cells, co-localizing

with *RUNX2*, another key regulator in chondrocyte hypertrophy and ossification.[332] Although the phenotype in individuals affected with OC does not approach the severity of the knockout model, it is not unreasonable that one or more variants in this gene could contribute to disease risk.

Materials and Methods

Candidate Gene Sanger Sequencing: The human sequence for *HDAC4* was aligned to the horse reference using the Basic Local Alignment Search Tool (BLAST) of the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/Blast.cgi). Primers for each annotated exon were designed using Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi). Primers for annotated exons in the human sequence that did not align to the horse reference were designed based on the human sequence. Primer optimization and Sanger sequencing were performed in five individuals, three affected with OC and two unaffected. The master mix for each PCR reaction included 1.5µl 10x PCR buffer (QIAGEN, Valencia, CA), 1.0µl each of [20µM] forward and reverse primer, 1.5µl [300µM] dNTPs, 0.15µl HotStarTaq® DNA polymerase (QIAGEN, Valencia, CA), 8.25µl water, and 3.0µl DNA (2.5ng/µl). Primer pairs are detailed in **Table 1**. PCR reactions were carried out under the following thermocycler conditions: 20 min at 95°C; 35 cycles of 30 sec at 94°C, 30 sec at 60°C, 30 sec at 72°C; 15 min at 72°C. PCR products were visualized on 2% agarose gels with ethidium bromide prior to submission for sequencing. To prepare the PCR products for Sanger sequencing, 1µl USB® ExoSAP-IT® (Affymetrix, Santa Clara, CA) was added to 4µl PCR product and incubated in the thermocycler for 15 min at 37°C for PCR

product cleanup followed by 15 min at 80°C for enzyme inactivation. Subsequently, 1µl [20µM] primer and 6µl water were added and the sample (12µl total volume) submitted to the University of Minnesota BioMedical Genomics Center (UMGC) for sequencing. Sequences were analyzed using Sequencher software (Gene Codes Corporation, Ann Arbor, MI) and were aligned to both equine and human reference sequences for each annotated exon.

Sanger Sequencing of cDNA: To capture exons that could not be amplified from genomic DNA, RNA was isolated from cartilage and subchondral bone (SCB) and reverse transcribed to cDNA. Frozen samples were placed in RLT buffer (QIAGEN, Valencia, CA) and homogenized using a Polytron instrument, then decanted into clean tubes. 20µl proteinase K was added to the sample, followed by incubation at 55°C for 25min and at room temperature for 20min. The sample was then spun down at 12,000rpm for 10min at room temperature and 5min at 4°C. The supernatant (800µl) was removed to a clean tube, and 80µl sodium acetate, 800µl phenol, and 160µl chloroform added. The sample was agitated for 15sec, followed by incubation on ice for 15min before being spun down at 12,000rpm for 20min at 4°C. This phenol/chloroform step was repeated before isopropanol and ethanol washes. After the ethanol wash, 25-40µl TE buffer was added to rehydrate the sample. Quantity and purity of extracted RNA were assessed using spectrophotometric readings at 260 and 280nm (NanoDrop 1000, Thermo Scientific, Wilmington, DE). Reverse transcription of RNA to cDNA was performed using the First-Strand Synthesis System (Invitrogen, Carlsbad, CA) per manufacturer instructions. Briefly, 1µl random hexamers (50ng/µl) and 1µl of a 10mM dNTP mix were added to 8µl of each RNA sample, followed by incubation at 65°C for 5min and on ice for 1min. After

the addition of the reverse transcriptase and 10min incubation at room temperature, the samples were placed in the thermocycler at 42°C for 50min, then 70°C for 15min. Samples were chilled and briefly centrifuged prior to addition of 1µl RNase H to remove any remaining RNA and a final incubation at 37°C for 20min. cDNA was stored at -20°C until use. Primers for cDNA samples were designed from the sequence of flanking exons generated from experimental samples (described above).

As an alternative approach to identify missing exons and 5'/3' UTR sequence, RNAseq data for equine *HDAC4* was obtained from Dr. James MacLeod (University of Kentucky). This data was compiled for visualization using the 'pileup' command in SAMtools.[333] Regions with a read depth of greater than 50 for a distance of at least 100bp were identified as putative "regions of interest" that could correspond to an exon or UTR. Primers pairs for these regions of interest are described in **Table 2**. PCR master mix components and reaction conditions were as described above. Preparation for sequencing was as described above. Sequence was analyzed in Sequencher and assembled together with the existing sequence data against the equine reference sequence and annotated human exons.

Results

Candidate Gene Sequencing: HDAC4 has 27 annotated exons in the human sequence, but not all of these have corresponding annotations in the horse sequence. Notably, exon 1 is missing from the equine annotation, and a large gap in the reference sequence (~20kb) spans the region expected to contain exons 13-15. Sequence corresponding to human exons 13 and 15 mapped to the equine "Chromosome

Unknown." Eight SNPs were identified within regions corresponding to 22 exons and the adjacent intronic sequence (**Table 1**). Six of these were intronic and two were exonic, located in exons 22 and 24 (equine annotation exons 19 and 21). The SNP at chr6: 24897009 was a synonymous thymine (T) to cytosine (C) base substitution. The SNP at chr6:24884889 was a cytosine (C) to thymine (T) base substitution that, based on the human sequence, may result in a alanine (A) to valine (V) amino acid substitution. Using genomic DNA, amplification of exons 1, 12, 14, 17, and 27 (based on the human annotation), as well as the 5' and 3' untranslated regions (UTR), was unsuccessful despite multiple attempts at primer design. cDNA was successfully obtained from cartilage and subchondral bone, but none of the tested primer pairs resulted in a PCR product of sufficient purity to warrant sequencing.

Sequence was successfully amplified and aligned from 6 of the 9 regions of interest identified in the RNAseq data. Based on alignment with horse and human sequence, these putatively corresponded to sections of the 3'UTR, exon 17, and either exon 1 or the 5'UTR (**Table 2**). One SNP was found within the putative exon 17 sequence (chr6:24908181), but it did not align with the human annotated sequence for this exon, so it could not be determined if it was exonic or intronic. Five SNPs were discovered within the putative 3'UTR sequence: chr6:24866490 (G/A), 24866332 (G/A), 24865511 (T/C), 24865403 (G/A), 24864640 (T/C).

Conclusions

Despite employing multiple approaches, including the use of genomic DNA, cDNA isolated from cartilage and subchondral bone, and RNAseq data, we were unable

to bridge a 20kb gap in the middle of *HDAC4* and sequence missing exons. Attempts to identify the first exon and 5'UTR were also minimally successful. Several novel variants were discovered, but their significance is unknown. Genotyping these variants in a large population of horses phenotyped for OC would be required to determine if they segregated with disease status, and this could be done in the future.

The difficulties we encountered with a traditional candidate gene sequencing approach are not unique to this gene, and are likely one reason why so few candidate genes for complex diseases have been thoroughly investigated in the horse. As the cost of next-generation sequencing technologies decrease, these offer a viable alternative to Sanger sequencing. The application of this alternative methodology to OC has been performed as a part of this thesis work, and is described in detail in **Chapter 6**.

Table 1: *HDAC4* primer pairs and the gene structures they encompass according to the published horse and human annotations. Primers for which there was no horse annotation were designed based on the human reference sequence. For each primer pair, the forward primer is listed first, then the reverse primer.

Sequencing Primers	**Horse Annotation**	**Human Annotation**	**SNPs [bases] bp**
5'-GCTAACAGATTCCAAGTGGTTTG-3' 5'-TTTGCACTTGGTGCTGTCAT-3'	none	exon 2	none
5'AGGGGCTTTCGTTTTCTGAT-3' 5'-GAGCTTCTGGTCCAGCTCAG-3'	exon 2	exon 3	intronic [C/T] 25035908
5'-ACTGGTTTTTGCGTTTGGAC-3' 5'-AGACGGTTATCGGGAGCAG-3'	exons 3/4	exon 4	none
5'-AATGCTGGCTGTGTAGACGA-3' 5'-CTCACCAGGCATCTGGTACA-3'	exon 5	exon 5	none
5'-TGCTGATTACACCTGCGTTT-3' 5'-CATTCAAAGACGAGCCCACT-3'	exon 6	exon 6	none
5'-CTTTCAGTCTTGCCCCAGAG-3' 5'-CACGCCACCAAAAGAAGAGT-3'	exon 7	exon 7	intronic [G/A] 24956750
5'-GCCAAAGGCATCAGGTATGT-3' 5'-AGCAGCTTTTAGCATCTGACG-3'	exon 8	exon 8	none
5'-GGGACTATGGTCGTGTGCTT-3' 5'-GGTTGGGAGCTGTTCTCTGA-3'	exon 9	exon 9	none
5'-CCCTTTCTCGTGGTACGTGT-3' 5'-AGCCTGGGTCTACAAACCTTC-3'	exons 10/11	exon 10/11	none
5'-GTGCCTGTTCGTCCGATAGT-3' 5'-GCAGGTGCATAAATGACCAC-3'	none	exon 13	none
5'-CGTGGATCTCTGGACCAAAC-3' 5'-GACAGCCTGGCTTCTTTGAG-3'	none	exon 15	intronic [C/T] unknown

5’-TGGGGGTTAAAGCATTGAAG-3’ 5’-CTGCAAAACCAGTGGCTCTC-3’	exon 13	exon 16	none
5’-AACCTTGGCATCATTGCTCT-3’ 5’-ACGCAACAGTGATCAACAGG-3’	exon 15	exon 18	none
5’-CCAGATGTCGCTGATGCTTA-3’ 5’-CAGAACCCGAAAGAGTCCAG-3’	exon 16	exon 19	none
5’-GAGAGGCCAGGGAGTGTTCT-3’ 5’-TGCATCTTTGCGGTAGTCTG-3’	exon 17	exon 20	intronic [G/A] 24899501
5’-TGGCTCACTTTTCACAGACG-3’ 5’-TGCCTGAGATCACAGTCTGC-3’	exon 18	exon 21	none
5’-CAGCCTCACTGGCAGATGTA-3’ 5’-ATGGCAGACAAAAGGGAAGA-3’	exon 19	exon 22	exonic [C/T] 24897009
5’-TAAAGTGAGCAGAGGCGTGA-3’ 5’-CTTGCTGGCACTGTCATGTT-3’	exon 20	exon 23	none
5’-TGAGATGATTTGCGGCATTA-3’ 5’-TCCTTGTGCTCACCTCCTTC-3’	exon 21	exon 24	intronic [G/A] 24885025 exonic [C/T] 24884889
5’-TCTCCCGGGAAAAATACCTT-3’ 5’-TGCTGTTCTAGCAGGTGGTG-3’	exon 22	exon 25	intronic [G/A] 24870589
5’-CCTGGGGAGGACTTGATTC-3’ 5’-CAAGGGTTTCCAGGGACTTT-3’	exon 23	exon 26	none

Table 2: *HDAC4* primer pairs based on RNAseq data and the physical position of the "regions of interest" they encompass. For primer pairs that amplified and aligned, the putative corresponding gene structure, based on human annotation, is listed. For each primer pair, the forward primer is listed first, then the reverse primer.

Sequencing Primers	Region bp start	Region bp stop	Putative corresponding gene structure
5'-AGTGGCTTGCAGACTCCTGT-3' 5'-TCTGAGAGCCTTGGTCCAGT-3'	24743584	24744165	did not amplify
5'-ACGTTCAGATGAGGGGACAG-3' 5'-GACCGTTTCCAACTTTCTCG-3'			3'UTR
5'-CTTCTATGGGCAAAGGGTGA-3' 5'-CTTTGCATCGAAAGGAAAGC-3'	24863829	24865050	
5'-AAATCTGCAACCCCACTGAG-3' 5'-GAAACTGCGCAGAATTCACA-3'			
5'-CCCGGTTCTCCATCAGAATA-3' 5'-AAGGGCCTCTTTGTCGAAGT-3'			
5'-GTGGGCCCAAAAGTCTACAA-3' 5'-ATGCGATAACGTGGACCTCT-3'	24865116	24866967	
5'-TACATGGTCACGCTCTCTGC-3' 5'-AGCACTTGAAGCCACCAGTT-3'			
5'-CAGGAAGGAGGAGATGGTCA-3' 5'-ATGGCAGTGAGCTGGGTAAC-3'			
5'-TCTAAAATCCAGGGCCTGCT-3' 5'-CACGGAAAGATCCCGAATAA-3'			
5'-ACCACCACGAAAGACCTCAG-3' 5'-GCCCTAGGCACACTTTCAGT-3'			
5'-ACCACCACGAAAGACCTCAG-3' 5'-GCCCTAGGCACACTTTCAGT-3'	24867171	24867419	

5’-ACACCTGGCAGCAATTATCC-3’ 5’-TCCCAAAAGATCCCTCTGTG-3’	24907952	24907998	Exon 17
5’-TGAAGGAGGGACTTGTTTGG-3’ 5’-TGGCACAGAACACCCATTTA-3’	24931244	24931459	did not amplify
5’-CGAGGAAGACAGATGGAAGG-3’ 5’-GCTTTGGGGAGAAAAGGAAA-3’	25011281	25011368	did not amplify
5’-TAGTTGTGGGATGTGGGACA-3’ 5’-AATTTGCAGCCCAAATTCAC-3’	25235253	25235339	5’UTR or Exon 1
5’-ACTCCCAAACAAACGGACAG-3’ 5’-AGCTCTGAGTCCACCCCTCT-3’	25361305	25361609	did not align

Appendix 2

Permissions for Inclusion of Published Work

Three of the chapters in this thesis have been published in their entirety, with modifications made only to conform to formatting requirements:

Chapter 2: Articular osteochondrosis: a comparison of naturally-occurring human and animal disease. (2013) *Osteoarthritis & Cartilage* 21:1638-1647.

Chapter 3: Short- and long-term racing performance of Standardbred pacers and trotters after early surgical intervention for tarsal osteochondrosis. (2014) *Equine Veterinary Journal* doi: 10.1111/evj.12297 [Published online 12 May 2014].

Chapter 5: Validation of imputation between equine genotyping arrays. (2014) *Anim Genet* 45:153. (Brief Note)

The publishers of these journals have the stated policy that authors retain the right to utilize published articles in theses, as long as appropriate acknowledgements are made. These guidelines can be found at:

Elsevier (Osteoarthritis & Cartilage):

http://www.elsevier.com/wps/find/authors.authors/postingpolicy

Wiley-Blackwell (Equine Veterinary Journal and Animal Genetics):

http://authorservices.wiley.com/bauthor/faqs_copyright.asp

Accessed May 2014