ABSTRACT: We are using genome-wide genotype data to identify segments that have undergone selection during breed formation. One region encompasses the *GYS1* gene, where a gain of function mutation shows historical evidence of positive selection in draft horses, but is now associated with exercise intolerance. Another region contains the *MSTN* gene where a haplotype containing several variants is present in 93% of Quarter Horses and 50% of Thoroughbreds, and is associated with altered muscle fiber type proportions. A third region contains the *DMRT3* gene where a mutation permissive to performing alternative gaits in many breeds has been found. Lastly, conserved haplotypes underlying signals of selection in the draft breeds and Miniature horse suggest the presence of a locus important in the determination of size. These are some of the first steps towards the identification of genes important in the specialization of modern horse breeds.

Keywords: Domestic Horses, FST, Haplotypes

Introduction

Since domestication, selective pressures on the horse genome have been directed toward use in agriculture, transportation, and warfare. More recently breed registries, and continued breed specialization, have focused more upon improving traits related to aesthetics, performance, and the ability to do work. The result is wide variation in phenotypes across breeds, and the fixation or near-fixation of some of the desired traits within many breeds, but with considerable variation among breeds. Genomic segments and the functional alleles underlying highly desirable phenotypes also become fixed during breed formation.

We have used whole genome microsatellite and SNP genotype data collected from 33 breeds to identify putative genomic regions under selection in the modern horse. The variants and processes that have contributed to desired phenotypes can then be more readily defined. Several examples are provided in this presentation.

Materials and Methods

Identification of a *GYS1* mutation and its history of selection. A genome scan with ~100 microsatellite markers was performed in a population of 96 PSSM Quarter Horse cases and 96 Quarter Horse (McCue et al, 2008). Sanger sequencing identified a non-synonymous SNP in the *GYS1* gene, as well as 10 additional SNPs within the neighboring 5.2 kb. In total 51 SNPs from the 1.97 Mb surrounding *GYS1* were available for haplotype analysis (McCoy et al, 2013).

Identification of genomic signatures of selection. 744 horses from 33 breeds were genotyped for ~54,000 SNPs on Illumina 54K arrays (20 – 24 horses/breed) (McCue et al., 2012). Samples and genotypes were contributed by the international equine genome research community through an Equine Genetic Diversity Consortium (Petersen et al., 2013). Putative loci under selection were identified using a FST-based statistic (d) calculated in sliding 500 kb windows (Akey et al., 2010). The statistic detects locus specific deviation in allele frequencies for each breed relative to the genome-wide average of pairwise FST summed across breeds. A minimal density of 4 SNPs/window was required, resulting in evaluation of a total 23,401 SNPs within 3,229 windows. This resulted in 68.7% coverage of the autosomes with an average SNP density of 7.25 SNPs per window (range 4-20). 33 windows within each breed fell into the upper 99th percentile of the empirical distribution and were considered putative signatures of selection (Figure 1).

Figure 1: Example output of the d1 calculation for a single breed. The d1 value is plotted on the y axis and each autosome is shown in the x axis in alternating colors. Each dot represents one 500kb window. The dashed horizontal line represents the 99th percentile of the empirical distribution of d1 for this breed.

Loci for follow-up investigation were prioritized by the following criteria: windows containing the highest d1 value within a breed; windows that contained consecutive segments of significant d1 values within a breed; windows that were shared across breeds experiencing selective pressure for similar phenotypes; and windows that were near candidate genes with known functional significance.

Results

Selection at the *GYS1* locus and its consequences. Quarter Horses with the condition known as Polysaccharide Storage Myopathy (PSSM) recurrently develop muscle pain and muscle cell damage upon light exercise. PSSM is also characterized by excessive glycogen (2 – 4 times the level in normal horses) and abnormal appearing polysaccharide in muscle. Many draft breeds, including Belgians and Percherons, have similarly abnormal muscle histopathology but with fewer clinical signs of acute muscle damage. A genome scan with microsatellite markers identi-
fied significant association of PSSM in Quarter Horses to ECA10 (McCue et al., 2008). The GYS1 gene, encoding the skeletal muscle glycogen synthase was sequenced and an exonic G to A SNP that causes an Arg308His mutation was identified. Arg308 is a highly conserved residue in glycogen synthase enzymes across diverse organisms. Glycogen synthase enzyme activity measurements were consistent with gain of function and altered regulation of enzyme activity, which could explain increased and abnormally-structured glycogen. Heterozygosity for the mutation is sufficient to cause clinical signs of disease.

Genotyping many thousands of horses for the GYS1 mutation has demonstrated allele frequencies of ~4% in Quarter Horses and Paint Horses, 25 - 33% in Belgians and Percherons (McCue et al., 2010), and > 67% in several continental European draft breeds (Baird et al., 2010). This observation begs the question as to why a mutation associated with a deleterious muscle disease could have such high allele frequencies in many related breeds. Fine-scale DNA sequencing in 80 horses and genotype assays in 279 horses revealed a paucity of haplotypes carrying the mutant allele when compared to haplotypes containing the wild-type allele (McCoy et al, 2013). Additionally, we found increased linkage disequilibrium, measured by relative extended haplotype homozygosity, in haplotypes carrying the mutation compared to haplotypes carrying the wild-type allele. Coalescent simulations of Belgian horse populations demonstrated that the high frequency and extended haplotype associated with the GYS1 mutation were unlikely to have arisen under neutrality or due to population demography. These data suggest that the GYS1 mutation underwent historical selection in the Belgian, and by inference, many related draft breeds.

Genomic signatures of selection from whole genome SNP data. Numerous potential targets of selection were identified from analysis of the 54,000 SNP genotype data from 33 breeds. 695 (2.7%) of the 3,229 500 kb genomic windows were identified by the d4 statistic as a potentially selected segment in at least one breed (Petersen et al., 2013).

Analysis of breeds with a high frequency of the MC1R chestnut coat color mutation demonstrated the utility of this method. Phasing the SNP data from ECA3 revealed an extended conserved haplotype that covers the MC1R locus in our sample of the Morgan breed which was comprised largely of chestnut horses. An identical, but shorter, haplotype was found in North American Belgians, which have chestnut coat color, as well as a number of other breeds in our study.

ECA18, MSTN and athletic performance. A striking feature of these genome scans was a 6 Mb region on ECA18 with a highly significant d4 value in the Quarter Horse and Paint Horse (Petersen et al., 2013). The minimal shared haplotype within these breeds was 0.78 Mb long and occurred at a frequency of 0.91 – 1.00. Twelve genes are in this region including MSTN. Sequencing of MSTN in Quarter Horses identified a promoter variant due to a SINE insertion, and a polymorphism in intron 1, that are correlated at > 0.95 in the selected haplotype. The identical haplotype is within a 2 Mb segment in Thoroughbreds, where it occurs at a frequency of 0.53. The intronic polymorphism in particular has been studied by several groups and found to be associated with performance in Thoroughbreds; specifically the racing distance that a horse is best suited for (Hill et al., 2010a; Hill et al., 2010b; Binns et al., 2010; Tozaki et al., 2012).

Histological characterization of gluteal muscle biopsies from 79 Quarter Horses representing all three haplotypes demonstrated a significant association of muscle fiber type proportions, but not fiber diameters, with the MSTN genotypes (Petersen et al., 2013). Specifically, Type 2B fiber content was increased and Type 1 fiber content was decreased in horses containing the SINE insertion or the C allele of the intron 1 SNP. Each copy of the SINE allele (on C allele) increases the Type 2B fiber percentage, and decreases the Type 1 fiber percentage by approximately 4% (Table 1).

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Type 1 %</th>
<th>Type 2A %</th>
<th>Type 2B %</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>20.4 a</td>
<td>27.2 a</td>
<td>52.5 a</td>
</tr>
<tr>
<td>NS</td>
<td>15.9 b</td>
<td>26.3 b</td>
<td>57.8 ab</td>
</tr>
<tr>
<td>SS</td>
<td>15.7 b</td>
<td>24.0 b</td>
<td>60.3 b</td>
</tr>
</tbody>
</table>

Table 1: Effect of the MSTN SINE insertion on Gluteal Muscle Fiber Type Proportions. Gluteal muscle fiber type proportions in 79 Quarter Horses based upon myostatin genotypes. Genotypes: NN, homozygous for the normal allele; SS, homozygous for the SINE insertion; NS, heterozygous. The SINE (as well as the intron 1 SNP, not shown) are each significantly associated with a lower proportion of Type 1 and higher proportion of Type 2B muscle fibers. Values significantly different at 0.05 are designated with different superscripts.

ECA23, DMRT3 and alternative gaits. The three naturally-occurring gaits in all equids are walk, trot and canter/gallop. Some horses can use alternate gaits, typically at an intermediate speed, and “gaitedness” is a trait that is the basis for development of many breeds. The gaited breeds analyzed here, including the Standardbred, Icelandic, Peruvian Paso, Paso Fino, and others, shared a highly conserved haplotype on ECA23 under a strong signal of selection (Petersen et al., 2013). The shared ECA23 haplo-
type within a breed was 429 – 759 kb long and occurred at frequencies of 0.54 – 1.00. Three genes encoding members of the DMRT family reside in this region.

Pace is a two-beat gait in which the horse moves ipsilateral fore and hind legs (the legs on the same side of the body) in synchrony, in contrast to the trot, where the diagonal front and hind legs move forward and backward together. Concurrently with our study, a genome wide association study in Icelandic Horses with the ability to perform the pace identified significant association of a SNP from the same region of ECA23 (Andersson et al., 2012). Subsequent sequencing identified a mutation in the DMRT3 gene that leads to a premature stop codon and a truncated protein. The identical mutation was present in the gaited breeds examined in our $d_I$ study, as well as breeds involved with trotting and pacing harness racing. A recent study that genotyped 4,396 horses from 141 breeds worldwide confirmed the high frequency of the stop codon mutation in breeds classified as gaited or bred for harness racing (Promerova et al., 2014).

Examination of wild-type and DMRT3 null mice demonstrated that this gene is expressed in the d16 subdivision of spinal cord neurons, takes part in neuronal specification within this subdivision, and is critical for the development of a coordinated locomotor network controlling fast-paced limb movements (Andersson et al., 2012).

**Loci potentially associated with size.** Conserved haplotypes underlying signals of selection on ECA11 in the Belgian, Percheron, Shire, Clydesdale, and Miniature horse suggested the presence of a locus important in the determination of size (Petersen et al., 2013). This region is approximately 0.5 – 1.5 Mb long with haplotype frequencies of 0.74 – 0.85 in the draft breeds. An alternative haplotype of ~ 0.5 Mb is found in the Miniature and Shetlands. Thirteen genes are in this region. Whether there are different alleles of the same gene having opposite phenotypic effects, or two independent loci at this region cannot be determined at this time.

That haplotypes can be constructed across any region, accompanied by measurements of haplotype frequency and sharing across breeds, allowed us to investigate the LCORL/NCAPG locus on ECA3 as a candidate region for selection for size. Here we could demonstrate a shared haplotype of 0.6 – 1 Mb in length that was close to fixation in draft breeds and present at frequencies of ~ 0.67 in Swiss Warmbloods (Petersen et al., 2013).

**Discussion**

Mapping signatures of selection is the first step in the identification of genes important in the domestication and specialization of modern horse breeds. This consideration of 33 diverse breeds contributed by an international consortium has demonstrated the utility of whole-genome approaches to identify genes important in the creation of modern horse breeds. Loci apparently selected for performance and coincidently with disease (GYSI), coat color (MCIR), performance and muscling (MSTN), gait (DMRT3), and size (LCORL) have already been identified. Many more potential loci remain to be investigated further.

We hypothesize that the gain of function $GYSI$ mutation arose many hundreds of years ago. We further hypothesize that it was once considered beneficial and was likely selected for due to increased muscle glycogen deposition that could enhance performance on the poorer diets available at the time. The identical mutation has since been carried on into many of the modern breeds where it is now detrimental, probably due to altered regulation of glycogen metabolism and cellular energy balance, coupled with excessive soluble carbohydrate diets and limited exercise. The means by which these factors result in the clinical manifestation of muscle pain with PSSM is not yet completely understood. Genotyping for the $GYSI$ mutation is currently being used to diagnose PSSM in affected animals, predict susceptibility, and select animals for breeding to reduce the overall incidence of this condition.

A reasonable hypothesis at this time is that the MSTN SINE insertion partially disrupts the promoter function and alters transcription. The effect on fiber type ratios described here, if also true in Thoroughbred horses, could contribute to the functional effect of the MSTN locus on racing performance at shorter vs longer distances. While the MSTN haplotype containing the SINE and intronic SNP C allele is segregating in Thoroughbreds it is essentially fixed in Quarter Horse, consistent with selection for muscling in this latter breed. Lastly, the haplotype with the SINE is only present at a significant frequency in Thoroughbreds, Quarter Horses and Paint Horses, whereas haplotypes with the intronic C allele are present in a wide variety of breeds. This suggests to us that the SINE insertion is likely the functional allele responsible for enhancing certain muscle performance phenotypes in the three breeds where it is found. MSTN genotypes are currently being used in the Thoroughbred industry to place horses into races of an optimal length for success.

The DMRT3 mutation clearly has had a major impact on the diversification of the domestic horse, as this mutation is strongly associated with the altered gait characteristic of a number of breeds. Although this mutation appears permissive for the ability to perform alternate gaits, which can be either pace, or four-beat ambling gaits such as the running walk and fox trot, it does not appear sufficient by itself to explain all aspects of alternative 4-beat ambling gaits in many breeds. In other words, there are likely many more gait-specific loci to be discovered. The $DMRT3$ mutation also has a favourable effect on harness racing performance, where it might help prevent breaking into a gallop during high-speed trotting. $DMRT3$ genotyping can now be considered in breeds for which the mutation is segregating with the aim of increasing the likelihood for performing alternative gaits.

Caveats to the $d_I$ scanning approach include: the number of loci potentially worthy of follow-up investigation is huge; important loci can be in regions that are not
included in the current analysis due to low SNP density; the same window may have a hit in different breeds for different reasons; the approach is blinded to phenotype; and identification of functional alleles is challenging.

We are currently using whole genome sequencing to attempt to identify functional alleles within selected regions. As a first screen we are looking for alleles in genic regions that are at high frequency in the selected haplotype pool vs the alternate haplotype pool. We will then move on to genotype larger populations for selected variants.

References
1University of Minnesota, College of Veterinary Medicine, St Paul, MN, USA
2University of Minnesota, School of Statistics, Minneapolis, MN, USA
3Swedish University of Agricultural Sciences, Department of Animal Breeding and Genetics, Uppsala, Sweden
4University of Kentucky, Department of Veterinary Science, Lexington, KY, USA
5University of California Davis, School of Veterinary Medicine, Davis, CA, USA
6Equine Analysis, Midway, KY, USA
7University Estadual Paulista, Department of Veterinary Clinical Science, Botucatu-SP, Brazil
8University College Dublin, School of Veterinary Medicine, Dublin, Ireland
9University of Azores, Department of Agriculture, Angra do Heroismo, Portugal
10University of Perugia, Faculty of Veterinary Medicine, Perugia, Italy
11Texas A&M University, College of Veterinary Medicine and Biomedical Science, College Station, TX, USA
12University of Veterinary Medicine Hannover, Institute for Animal Breeding and Genetics, Hannover, Germany
13Animal Health Trust, Lanwades Park, Newmarket, Suffolk, United Kingdom
14French National Institute for Agricultural Research-Animal Genetics and Integrative Biology Unit, Jouy en Josas, France
15University of Sydney, Veterinary Science, New South Wales, Australia
16Nihon Bioresource College, Koga, Ibaraki, Japan
17University of Helsinki, Faculty of Veterinary Medicine, Helsinki, Finland
18University College Dublin, College of Agriculture, Food Science and Veterinary Medicine, Belfield, Dublin, Ireland
19University of Bern, Institute of Genetics, Bern, Switzerland
20Institute of Cancer Research, Breakthrough Breast Cancer Research Centre, London, United Kingdom
21Royal Veterinary College, Comparative Neuromuscular Diseases Laboratory, London, United Kingdom
22Swiss National Stud Farm SNSTF, Agroscope Liebefeld-Posieux Research Station ALP-Haras, Avenches, Switzerland
23Norwegian School of Veterinary Science, Department of Basic Sciences and Aquatic Medicine, Oslo, Norway
24Animal DNA Diagnostics Ltd, Cambridge, United Kingdom
25Laboratory of Racing Chemistry, Department of Molecular Genetics, Usunomiya, Tochigi, Japan