# Detecting the Population Structure and Scanning for Signatures of Selection in Horses (*Equus caballus*) From Whole-Genome Sequencing Data

Evolutionary Bioinformatics
Volume 14: 1–9
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1176934318775106

Cheng Zhang[1,2], Pan Ni[1], Hafiz Ishfaq Ahmad[1], M Gemingguli[3], A Baizilaitibei[3], D Gulibaheti[3], Yaping Fang[2], Haiyang Wang[1,2], Akhtar Rasool Asif[1], Changyi Xiao[2], Jianhai Chen[1], Yunlong Ma[1], Xiangdong Liu[1], Xiaoyong Du[1,2] and Shuhong Zhao[1]

[1]Key Laboratory of Animal Genetics, Breeding and Reproduction of Ministry of Education, College of Animal Sciences & Technology, Huazhong Agricultural University, Wuhan, People's Republic of China. [2]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China. [3]College of Animal Science, Tarim University, Alar, China.

**ABSTRACT:** Animal domestication gives rise to gradual changes at the genomic level through selection in populations. Selective sweeps have been traced in the genomes of many animal species, including humans, cattle, and dogs. However, little is known regarding positional candidate genes and genomic regions that exhibit signatures of selection in domestic horses. In addition, an understanding of the genetic processes underlying horse domestication, especially the origin of Chinese native populations, is still lacking. In our study, we generated whole genome sequences from 4 Chinese native horses and combined them with 48 publicly available full genome sequences, from which 15 341 213 high-quality unique single-nucleotide polymorphism variants were identified. Kazakh and Lichuan horses are 2 typical Asian native breeds that were formed in Kazakh or Northwest China and South China, respectively. We detected 1390 loss-of-function (LoF) variants in protein-coding genes, and gene ontology (GO) enrichment analysis revealed that some LoF-affected genes were overrepresented in GO terms related to the immune response. Bayesian clustering, distance analysis, and principal component analysis demonstrated that the population structure of these breeds largely reflected weak geographic patterns. Kazakh and Lichuan horses were assigned to the same lineage with other Asian native breeds, in agreement with previous studies on the genetic origin of Chinese domestic horses. We applied the composite likelihood ratio method to scan for genomic regions showing signals of recent selection in the horse genome. A total of 1052 genomic windows of 10 kB, corresponding to 933 distinct core regions, significantly exceeded neutral simulations. The GO enrichment analysis revealed that the genes under selective sweeps were overrepresented with GO terms, including "negative regulation of canonical Wnt signaling pathway," "muscle contraction," and "axon guidance." Frequent exercise training in domestic horses may have resulted in changes in the expression of genes related to metabolism, muscle structure, and the nervous system.

**KEYWORDS:** horse, selective sweep, population genetic structure, single-nucleotide polymorphisms

**RECEIVED:** January 16, 2018. **ACCEPTED:** April 14, 2018.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Nature Scientific Foundation of China (nos 31501076 and 31560616), the National Natural Science Foundation of Hubei Province of China (no. 2015CFB244), and the Fundamental Research Funds for the Central Universities (nos 2662017JC027 and 2662015JC010). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Xiaoyong Du, Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, China. Email: duxiaoyong@mail.hzau.edu.cn

## Introduction

The genetic diversity of domesticated animals changes gradually through selective processes in populations. Archaeological and genetic evidence suggested that the initial domestication of horses (*Equus caballus*) began 5000 to 6000 years ago and possible multiple horse domestication events occurred across Eurasia.[1,2] For the past 400 years, the establishment of formal breed registries and continuous breed specialization has focused on the preservation and improvement of traits related to riding, draft, aesthetics, and performance. The intensive selection for these traits has led to high athletic quality and high skeletal muscle mass.[3] Chinese native horses were formed under different ecological conditions and exhibit high levels of morphological and genetic diversity.[4,5] Kazakh horses are believed to have been developed by crossing native horses with Mongolian, Middle Asian, and European breeds raised together in herds.[6] They are generally small, rugged horses with different coat colors. The Kazakh horse is used for riding and pack as well as for milk and meat. The Lichuan breed is geographically distributed in mountainous areas of Southern China and is broadly used in light draft and riding for human usage. In this study, we selected only Kazakh horses and Lichuan breeds, with distinct variations in numerous phenotypes, to test whether they represent potential Northern type and Southern type of Chinese horse, respectively. Previous studies demonstrated that Northern/Southern Chinese distinct groups can be used to discriminate the Han Chinese[7] as well the domestic pig genetically.[8]

Recent advancements in high-throughput sequencing technology and the assembly of the first horse genome sequence have allowed for in-depth analysis of the genetic variations present in horses.[9] Phenotypic changes associated with mutations

**Table 1.** Sampling and data source information for horses in the study.

| BREED | NO. | BIOSAMPLE ID[11,12,15,16] |
|---|---|---|
| Arabian | 6 | SAMN02179860, SAMN02439777, SAMEA3475296, SAMN05616420, SAMN05616421, SAMN05616422 |
| Duelmener | 1 | SAMN02422919 |
| Mongolian | 2 | SAMEA3498582, SAMEA3504070 |
| Franches-Montagnes | 12 | SAMEA3498888-SAMEA3498899 |
| Hanoverian | 2 | SAMN02439779, SAMN02439782 |
| Icelandic | 2 | SAMN02179857, SAMEA3355589 |
| Jeju pony | 4 | SAMN01057170-SAMN01057173 |
| Morgan | 1 | SAMEA3499838 |
| Norwegian Fjord | 1 | SAMN02179856 |
| Quarter | 4 | SAMEA3499834-SAMEA3499836; SAMEA34998378 |
| Sorraia | 2 | SAMN02439778, SAMN03955413 |
| Standardbred | 4 | SAMN02179441, SAMEA3499831-SAMEA3499833 |
| Przewalski | 3 | SAMN02179442,SAMN03009555, SAMN03009556 |
| Kazakh | 3 | SAMN07604084,SAMN07604083, SAMN07604081 |
| Lichuan | 1 | SAMN07604082 |

accompany domestication as a result of the joint impact of natural selection and human-controlled selective breeding. Due to a strong selection of beneficial alleles, a selective sweep leads to a single genomic background, which results in a large reduction in genetic variations in local region of the genome.[10] Selective sweeps have been found in domestic and wild horse populations using genome scans.[11–13] Selective sweeps have been found in 33 domestic horse breeds using genome scans of nearly 50 000 markers in Equine SNP50 BeadChip (Illumina, San Diego, CA, USA).[13] However, the limited markers cannot give single base resolution of the entire genome. In this study, we describe the whole genome sequencing of 4 horses from 2 Chinese native breed and 44 horses with publicly available genome data from completely different breeds to find and annotate distinct genetic disparities in horses using extremely precise SNP calling approaches. The aim of this study was to highlight candidate genes and to trace footprints of horse selection at the genome level. Genes were selected in proximity to genomic regions and positions displaying signatures of selection by the composite likelihood ratio (CLR) method using sequencing data.[14] Moreover, the functions of genes under selection were studied by gene ontology (GO) annotation analysis.

## Materials and Methods

### *Experimental animals and genome sequencing*

We sequenced 4 Chinese native horses including one individual from the Lichuan breed and 3 from the Kazakh breed. Genomic DNA was extracted from blood tissue using a standard phenol-chloroform protocol. This study was conducted according to the regulations approved by the ethical committee of Huazhong Agricultural University and also the standing committee of Hubei People's Congress, P. R. China.

High-quality DNA for genome sequencing was processed to construct short-insert DNA libraries according to the manufacturer's specifications (Illumina). The qualified libraries with appropriate insert size (500 bp [base pairs]) and concentration (>2 nM) were sequenced using the Illumina HiSeq X Ten platform with 150-bp paired-end reads (Illumina). Overall, we produced approximately 400.97 million raw reads (totaling 117 Gb of raw data) from the 4 samples (Table 1). These raw read data were deposited into the NCBI SRA Database with Bioproject accession numbers PRJNA401382/SRP117064. The genome sequence data of 44 other horse samples were retrieved from the NCBI SRA database[11,12,15,16] (Table 1). Low-quality reads were trimmed using Trimmomatic Version 0.36 with the following options: ILLUMINACLIP:TruSeq3-SE:2:30:10, LEADING:3, TRAILING:3, SLIDINGWINDOW:4:15, and MINLEN:75.[17]

### *SNP calling*

Horse gene sequences were retrieved from the NCBI database. Next-generation sequencing reads from the 48 samples were mapped individually to the reference genome using bwa-mem[18] with default parameters. BAM files for all animals were sorted, and duplicates were filtered using Picard-version-1.108 (http://broadinstitute.github.io/picard/). RealignerTargetCreator and IndelRealigner were applied for local realignment, and individual sample SNP calling was performed using HaplotypeCaller (with the following parameters: -pairHMM VECTOR_LOGLESS_CACHING –emitRefConfidence GVCF –variant_index_type LINEAR –variant_index_parameter 128000) in GATK version 3.50.[19] Multisample SNP calling was performed to merge the GVCFs using GenotypeGVCFs with default settings from GATK version 3.50.[19] The identified SNPs in the VCF were filtered for downstream analysis by requiring a minimum depth of 5 and a minimum root mean square mapping quality score of 20. We also removed SNPs with a minor allele frequency <0.05 and with >10% missing genotypes among the 48 samples in the population.

The 15 341 213 high-quality SNPs identified in the horse were categorized according to their genomic locations, including in exons, introns, untranslated regions (UTRs), and intergenic regions, and SNPs located in exons were further divided into synonymous and nonsynonymous SNPs. The high-quality SNPs recovered here were used in the GO, SNP summary, target gene, and selective sweep analyses. The putative SNPs were functionally annotated primarily based on the SnpEff software system analysis[20] with standard settings. For each SNP, the position (5′ UTR, 3′ UTR, exonic, intronic, intergenic, splice acceptor or donor site, upstream or downstream) and also the useful annotation (nonsynonymous, synonymous) were identified based on the gene annotation of the horse reference genome from the NCBI database (EquCab2.0, annotation date November 20, 2015).[9]

**Table 2.** Summary of single-nucleotide polymorphisms in horses.

| CATEGORY | COUNT | PERCENTAGE | NOTE |
|---|---|---|---|
| Sample size | N = 48 | — | "Splice_region" means that a variant is within 2 bp of a splice junction. "Splice_acceptor" means that the variant hits a splice acceptor site (defined as 2 bases before the exon start site, except for the first exon). "Splice_donor" means that the variant hits a splice donor site (defined as 2 bases after the end of the coding exon, except for the last exon). "Upstream/downstream" means that a variant overlaps with the 1-kb region upstream/downstream of the gene end site. The number of effects is larger than the number of SNPs because a variant can be annotated for 2 or more effects. |
| SNP | 15 341 213 | — | |
| 3_prime_UTR | 172 656 | 0.46 | |
| 5_prime_UTR | 46 446 | 0.13 | |
| Downstream | 2 488 567 | 6.70 | |
| Intergenic_region | 10 032 903 | 27.02 | |
| Intragenic | 794 | 0.00 | |
| Intron | 10 441 311 | 28.12 | |
| Missense | 103 427 | 0.28 | |
| Non_coding_transcript_exon | 96 876 | 0.26 | |
| Non_coding_transcript | 10 793 318 | 29.07 | |
| Splice_acceptor | 384 | 0.00 | |
| Splice_donor | 577 | 0.00 | |
| Splice_region | 25 231 | 0.07 | |
| Start_lost | 283 | 0.00 | |
| Stop_gained | 1054 | 0.00 | |
| Stop_lost | 224 | 0.00 | |
| Stop_retained | 82 | 0.00 | |
| Synonymous | 129 097 | 0.35 | |
| Upstream | 2 474 202 | 6.66 | |
| Exon | 327 360 | 0.88 | |
| No. of effects | 37 134 791 | 100 | |

### *Population genetics analysis*

We also conducted an individual-scale principal component analysis (PCA) for the 48 horses. Here, genotype likelihoods were estimated assuming Hardy-Weinberg equilibrium using the ANGSD package (with parameters: angsd -nInd 48 -doMajorMinor 1 -doMaf 1 -doPost 1 -doGeno 32 -doSaf 1).[21] Genotype likelihoods were used to compare horses via PCA using the function "ngsCovar" in ngsTools,[22] which implements a probabilistic approach to estimate the genotype covariance matrix (with parameters -nind 48 -nsites 12363534 -block_size 20000 -call 0 -minmaf 0.05).

The population structure among various domestic populations was inferred using admixture.[23] We removed the Przewalski horses because they are the only wild population. After excluding the 3 Przewalski horses, we choose 422 881 high-quality autosomal SNPs randomly to infer the genetic structure of the domestic horses, using the program admixture, which includes a cross-validation procedure that allows identifying the value of *K* for which the model has best predictive accuracy.

### *Detection of positive selection*

The method for detection of positive selection for horse is mainly similar to the previous one that we have used for ducks.[24] To identify sweeps, the CLR test was performed using allele frequencies, and the site frequency spectrum (SFS) of the complete chromosome was considered as the background SFS to calculate the combined likelihood of a recent selective sweep in each window.[14] The empirical CLR distributions were obtained by 1000 simulations of a neutral sequence equal to each chromosome in length using ms software[25] under a previously selected demographic model named "PSMC #1," which represents the basic model with a constant population size from t2 to t0.[11] In this PSMC profile, a mutation rate of $7.242 \times 10^{-9}$ per site per generation and a generation time of 8 years were used.[16] The CLR value and the corresponding *A* value in each alignment positions where the SweeD score is calculated for the horse and simulated genotypes were obtained from a SweeD run with a grid size of 10 kB. The corresponding $\alpha$ value of a CLR value is a function of the selection coefficient, the recombination rate, and the effective population size. The results from the simulated genotypes were used to calculate a critical threshold of CLR values for observed data in each chromosome. We identified candidate sites for selective sweep on the basis of *P* values after correction of multiple testing using a false discovery rate (FDR) of 0.05 (using "BH" method in the "p.adjust" function of R). We then identified annotated sequences overlapping with the selected regions as determined by SweeD.

### *Functional enrichment*

Functional enrichment was performed on the list of loss-of-function (LoF) genes and the genes under selection in horses as detected by the genome-wide selective sweep scans, respectively. The GO terms were obtained using the Databank for Annotation, Visualization and Integrated Discovery (DAVID).[26] DAVID was used to evaluate enrichment in the GO terms using known annotations of horse genes with *Equus caballus* selected as background. For further GO term analysis, a *P* value of .05 and a FDR of 25% were set.

## Results and Discussion

### *Genomic variants*

Overall, 15 341 213 high-quality SNPs were detected using GATK, representing the distinctive variants of 48 samples. The SNPs were functionally classified using SnpEff annotations with the boundaries of all neighboring genes (Table 2). Most SNPs were located in intergenic regions, and intronic SNPs were most common in genic regions. SnpEff also found

**Table 3.** Enriched biological processes from GO analysis of the loss-of-function genes.

| TERM | GENE COUNTS (GENES WITHOUT ANNOTATION NOT SHOWN) | *P* VALUE | BENJAMINI |
|---|---|---|---|
| GO:0007186~G protein–coupled receptor signaling pathway | 55 genes including *GPR143, OPN3* | 4.56E–12 | 4.64E–09 |
| GO:0002504~antigen processing and presentation of peptide or polysaccharide antigen via MHC class II | *DRB, DQA, DQB, LOC100060531* | 8.51E–09 | 4.32E–06 |
| GO:0050907~detection of chemical stimulus involved in sensory perception | 14 genes | 1.28E–04 | 4.25E–02 |
| GO:0006955~immune response | 17 genes including *DRB, DQA, DQB, LST1, GZMB, OAS1, EQMHCC1, FAS, EQMHCB2,* | 1.77E–04 | 4.39E–02 |
| GO:0002474~antigen processing and presentation of peptide antigen via MHC class I | 5 genes including *EQMHCC1, EQMHCB2* | 7.49E–04 | 1.41E–01 |
| GO:0007165~signal transduction | 21 genes including *PLPP5,PPP2R5C, CD83, GABRR1, SMOC1, GUCY1A3, GUCY1B3, UNC5D* | 4.33E–03 | 5.20E–01 |
| GO:0007608~sensory perception of smell | 13 genes | 1.15E–02 | 8.12E–01 |
| GO:0035518~histone H2A monoubiquitination | *KDM2B, CUL4B, PCGF1* | 3.75E–02 | 9.92E–01 |
| GO:0090179~planar cell polarity pathway involved in neural tube closure | *WNT5A, DVL3, SFRP1* | 4.49E–02 | 9.94E–01 |
| GO:0019882~antigen processing and presentation | 4 genes | 5.16E–02 | 9.95E–01 |
| GO:0006952~defense response | *CD83, DEFA5L, DEFA3, DEFA31L, DEFA16* | 5.95E–02 | 9.97E–01 |
| GO:0046322~negative regulation of fatty acid oxidation | *ACADVL, SIRT4* | 6.19E–02 | 9.96E–01 |
| GO:0033182~regulation of histone ubiquitination | *UBE2N, LOC102150222* | 6.19E–02 | 9.96E–01 |
| GO:0031058~positive regulation of histone modification | *UBE2N, LOC102150222* | 6.19E–02 | 9.96E–01 |
| GO:0055123~digestive system development | *WDR19, WDPCP* | 6.19E–02 | 9.96E–01 |
| GO:0071346~cellular response to interferon gamma | *WNT5A, CCL19, NOS2, DAPK1* | 8.41E–02 | 9.99E–01 |
| GO:0050821~protein stabilization | *DVL3, LAMP2, PFN2, WFS1, STXBP4, CSN3, USP13* | 8.49E–02 | 9.98E–01 |
| GO:0051443~positive regulation of ubiquitin-protein transferase activity | *UBE2N, LOC102150222, LOC100058286* | 8.85E–02 | 9.98E–01 |
| GO:2000360~negative regulation of binding of sperm to zona pellucida | *OVGP1, ASTL* | 9.14E–02 | 9.98E–01 |
| GO:0042991~transcription factor import into nucleus | *IPO9, SYK* | 9.14E–02 | 9.98E–01 |
| GO:0016485~protein processing | *APH1A, LOC100050554, GZMB, LOC100061896, CPN1* | 9.40E–02 | 9.97E–01 |
| GO:0035556~intracellular signal transduction | *SRPK2, DVL3, ARHGEF3, MKNK2, ASB11, SOCS7, MYO9A, DAPK1, HUNK, RPS6KA3, AIDA, GUCY1A3, GUCY1B3* | 9.91E–02 | 9.97E–01 |

Abbreviation: GO, gene ontology.

172 656 SNPs in the 3′ UTRs and 46 446 SNPs in 5′ UTRs. The SNPs in the upstream/5′ UTRs and downstream/3′ UTRs might affect transcription and translation, respectively, but actual functional effects must be confirmed case by case. With respect to SNPs found in splice regions, 384 were found in splice acceptor sites and 577 were found in splice donor sites. A total of 247 513 SNPs were found to have an effect on coding regions: 1054 SNPs may cause a premature stop codon, 224 may abrogate a termination codon, and 103 427 (0.28%) may cause nonsynonymous substitutions. The remaining SNPs within the coding regions were expected to be either synonymous (129 097) or in noncoding exons (96 876). As a measure of the quality of our SNP data, we recovered a transition-transversion (ts/tv) ratio of 2.05 across the horse genome, which closely mirrors the global ts/tv ratio of 2.0-2.1 for the human genome.[27]

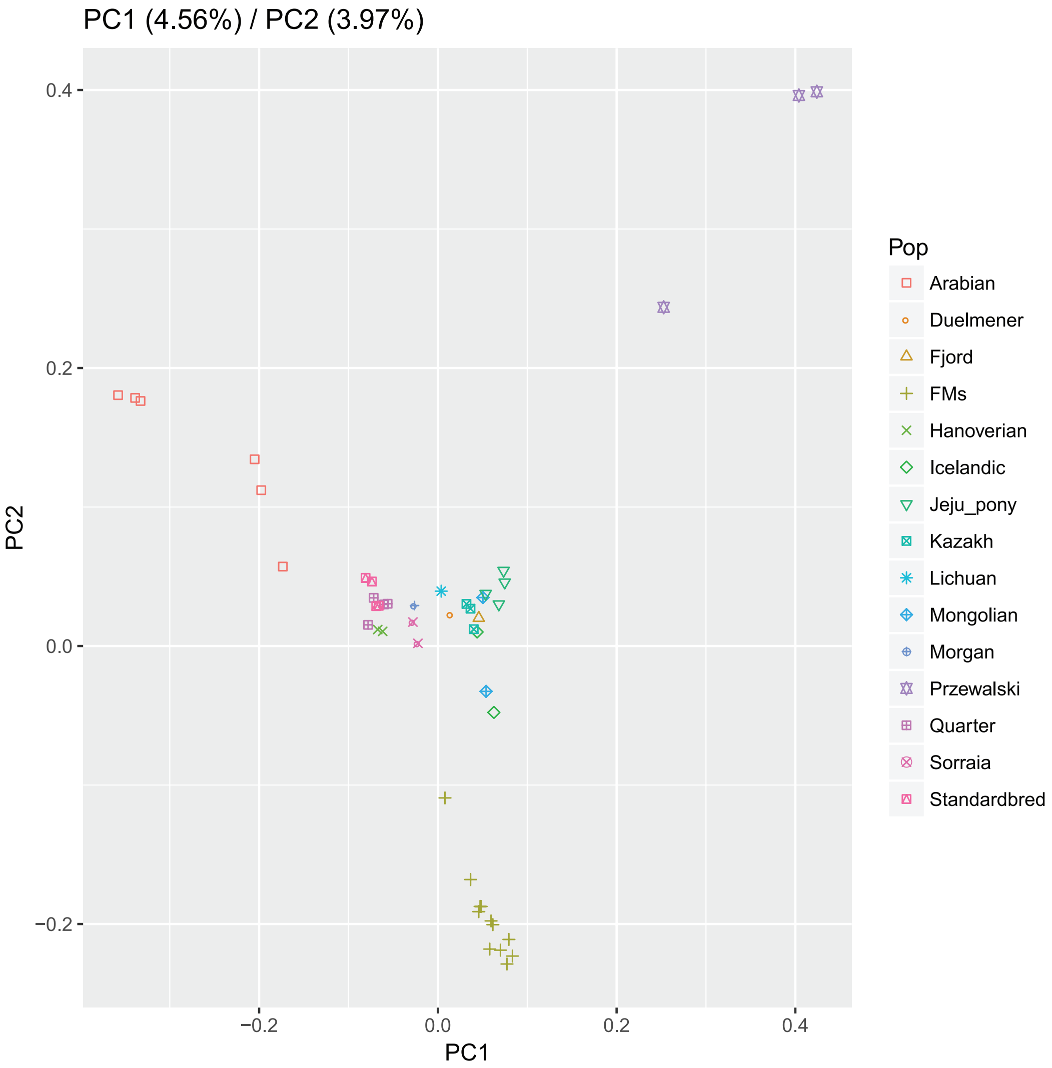

**Figure 1.** Principal component analysis results of all 48 horses. The *x-axis* denotes the value of PC1, whereas the *y-axis* denotes the value of PC2. Each dot in the figure represents one individual.

To gain insights into the SNPs underlying recessive traits, we clustered 5 high-impact severity types of SNPs into LoF variants, including splice acceptor, splice donor, stop gain, stop loss, and start loss SNPs in protein-coding genes. We detected 1390 LoF SNPs in protein-coding genes, all of which were heterozygous in at least one of the sampled animals. There were 981 genes with more than 1 LoF SNP. The GO enrichment analysis revealed that genes associated with "G protein–coupled receptor (GPCR) communication pathway," "antigen processing and presentation of peptide or polysaccharide antigen via major histocompatibility complex (MHC) class II," "detection of chemical stimulus involved in sensory perception," "immune response," and "signal transduction" were overrepresented in LoF-affected genes (Table 3). G protein–coupled receptors have been found to be linked with signaling pathways such as transmembrane receptor activity and neurotransmission.[28] Two similar GO enrichment results for LoF-containing genes pointing to GPCR receptor activity and sensory perception were also found in humans[29] and cattle.[30] Interestingly, genes associated with GO terms "antigen processing and presentation of peptide or polysaccharide substance via MHC class II," "immune response," "defense response," and "cellular response to interferon-gamma" were significantly ($P < .05$) overrepresented in the LoF-affected gene set (Table 3). These GO terms are involved in immune responses. *DRB*, *DQA*, and *DQB* are classical MHC class II molecules involved in the development of adaptive immune responses.

### *Population genetic structure*

Principal component analysis was used to explore the individual relationships within and among breeds. The first 2 principal components explained 8.53% of the variation of the studied horses (Figure 1). The position of Przewalski clearly deviates from the domestic horse in PC1. PC1 also allows visualizing the proximity among Duelmener, Fjord, Icelandic, Kazakh, Lichuan, and Mongolian (where we defined group A) and the proximity among Hanoverian, Morgan, Quarter, Sorraia, and Standardbred (where we defined group B).

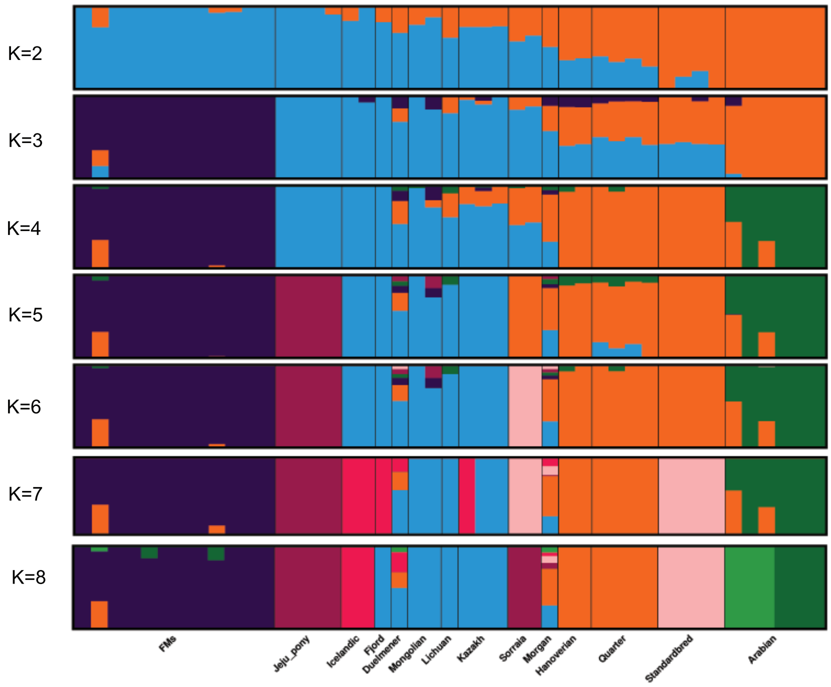

**Figure 2.** Bayesian clustering output for 5 *K* values from $K=2$ to $K=8$ in 45 domestic horses. Each individual is represented by a vertical line, which is partitioned into colored segments that represent the proportion of the inferred *K* clusters.

Admixture proportions were assessed without user-defined population information to infer the presence of distinct populations among the samples (Figure 2). At $K=3$ or $K=4$, Franches-Montagnes and Arabian forms one unique cluster; at $K=5$, Jeju pony forms one unique cluster. For other breeds, comparatively strong population structure exists among breeds, and they can be assigned to 2 (or 3) alternate clusters from $K=3$ to $K=5$ including group A (Duelmener, Fjord, Icelandic, Kazakh, Lichuan, and Mongolian) and group B (Hanoverian, Morgan, Quarter, Sorraia, and Standardbred). For group A, geographically this was unexpected, where Nordic breeds (Norwegian Fjord, Icelandic, and Duelmener) clustered with Asian breeds including the Mongolian. Previous results of mitochondrial DNA have revealed links between the Mongolian horse and breeds in Iceland, Scandinavia, Central Europe, and the British Isles. The Mongol horses are believed to have been originally imported from Russia subsequently became the basis for the Norwegian Fjord horse.[31] At $K=6$, Sorraia forms one unique cluster. The Sorraia horse has no long history as a domestic breed but is considered to be of a nearly ancestral type in the southern part of the Iberian Peninsula.[32] However, our result did not support Sorraia as an independent ancestral type based on result from $K=2$ to $K=5$, and the unique cluster in $K=6$ may be explained by the small population size and recently inbreeding programs. Genetic admixture of Morgan reveals that these breeds are currently or traditionally continually crossed with other breeds from $K=2$ to $K=8$. The Morgan horse has been a largely closed breed for 200 years or more but there has been some unreported crossbreeding in recent times.[33]

Bayesian clustering and PCA demonstrated the relationships among the horse breeds with weak geographic patterns. The tight grouping within most native breeds and looser grouping of individuals in admixed breeds have been reported previously in modern horses using data from a 54K SNP chip.[33,34] Cluster analysis reveals that Arabian or Franches-Montagnes forms one unique cluster with relatively low *K* value, which is consistent with former study using 50K SNP chip.[33,34] Interestingly, Standardbred forms a unique cluster with relatively high *K* value in this study, different from previous study.[33] To date, no footprints are available to describe how the earliest domestic horses spread into China in ancient times. Our study found that Kazakh and Lichuan were assigned to the same lineage as other native Asian breeds, in agreement with previous studies on the origin of Chinese domestic horses.[4,5,35,36] The strong genetic relationship between Asian native breeds and European native breeds have made it more difficult to understand the population history of the horse across Eurasia. Low levels of population differentiation observed between breeds might be explained by historical admixture. Unlike the domestic pig in China,[8] we suggest that in China, Northern/Southern distinct groups could not be used to genetically distinct native Chinese horse breeds. We consider that during domestication process of horse, gene flow continued among Chinese-domesticated horses.

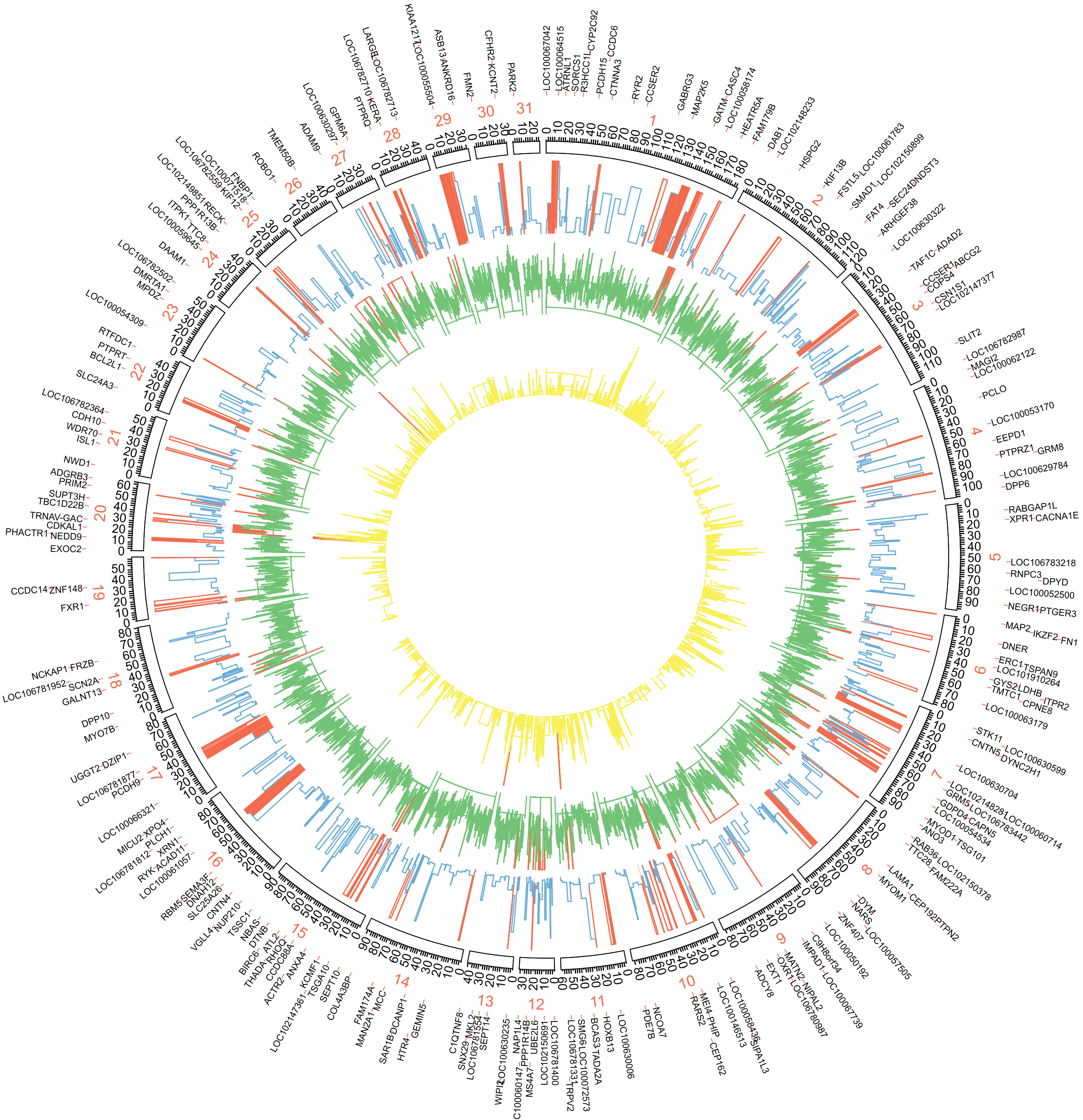

**Figure 3.** Circos plot of the global distribution of genes, SNP variants, and signature of selective sweeps along the 31 autosomes. The circles, from outside to inside, illustrate gene density (yellow), SNP density (green), and CLR values (blue). The genes located in regions with significant strong sweep signatures are presented as outliers. High values in each layout (gene density: number per 100 kB; SNP density: number per 200 kB, and CLR value >15) are marked in red. CLR indicates composite likelihood ratio; SNP, single-nucleotide polymorphism.

### *Genome-wide selection signatures*

The CLR test scanned genomic regions to find signals of positive selection in the horse genome. Composite likelihood ratio is calculated by comparing the regional SFS with the background SFS that specifies the probability of a signal at each window of 10 kB in length across the horse genome. A total of 1052 windows obtained a *P* value less than .01, indicating that the CLR of those windows surpassed the CLRs achieved from the distribution of 1000 neutral simulations. The SNP and gene density values for these regions are presented in Figure 3. Some of the 1052 windows were adjacent to one another, which led to the designation of 933 distinct core regions. The regions with the largest CLR in the horse genome are located on chromosome 8 (Figure 3), including 9 adjacent 10-KB windows. The top-scoring 90-Kb region was chr8: 1.916 to 2.006, and this region contains 227 SNPs, all of which are intron variants of the *LOC100065759* pseudogene.

**Table 4.** Enriched biological processes from GO analysis of the genes in the CLR region.

| TERM | GENES | *P* VALUE | BENJAMINI |
|---|---|---|---|
| GO:0090090~negative regulation of canonical Wnt signaling pathway | *RGS20, MCC, PARK2, ISL1, FRZB, SDHAF2, STK3* | 2.93E−03 | 9.20E−01 |
| GO:0006936~muscle contraction | *MYOM2, COL4A3BP, MYOM1, TMOD1* | 1.22E−02 | 9.95E−01 |
| GO:0007411~axon guidance | *MATN2, ROBO1, CNTN4, TTC8, EXT1, CDH4* | 1.31E−02 | 9.77E−01 |
| GO:0001736~establishment of planar polarity | *WDPCP, PTK7, TTC8* | 1.48E−02 | 9.59E−01 |
| GO:0008286~insulin receptor signaling pathway | *PHIP, PTPN2, GPLD1, RHOQ* | 2.22E−02 | 9.79E−01 |
| GO:0016192~vesicle-mediated transport | *FMN2, FNBP1, AP2S1, SAR1B, BCAS3* | 2.22E−02 | 9.60E−01 |
| GO:0045773~positive regulation of axon extension | *TRPV2, CDH4, FN1* | 2.70E−02 | 9.65E−01 |
| GO:0033539~fatty acid β-oxidation using acyl-CoA dehydrogenase | *ETFDH, ACAD11, ETFA* | 4.20E−02 | 9.90E−01 |
| GO:0050885~neuromuscular process controlling balance | *SLC1A3, NRXN1, HERC1, PTPRQ* | 4.26E−02 | 9.84E−01 |
| GO:0060541~respiratory system development | *WDPCP, SPEF2* | 5.10E−02 | 9.89E−01 |
| GO:0097118~neuroligin clustering involved in postsynaptic membrane assembly | *MAGI2, NRXN1* | 5.10E−02 | 9.89E−01 |
| GO:0007156~homophilic cell adhesion via plasma membrane adhesion molecules | *ROBO1, FAT4, PCDH9, CDH4, CDH10* | 5.36E−02 | 9.87E−01 |
| GO:0051645~Golgi localization | *DAB1, STK11* | 6.74E−02 | 9.93E−01 |
| GO:0042384~cilium assembly | *ACTR2, CEP162, WDPCP, DZIP1, TTC8* | 7.52E−02 | 9.94E−01 |
| GO:0090630~activation of GTPase activity | *RALGAPA1, RABGAP1L, BCAS3, TBC1D22B* | 7.89E−02 | 9.94E−01 |
| GO:0055088~lipid homeostasis | *COL4A3BP, ACAD11, ETFA* | 8.40E−02 | 9.93E−01 |
| GO:0060070~canonical Wnt signaling pathway | *STK11, PTK7, RYR2, FRZB* | 8.53E−02 | 9.92E−01 |
| GO:0086010~membrane depolarization during action potential | *SCN2A, CACNA1E, NALCN* | 8.92E−02 | 9.91E−01 |
| GO:0045184~establishment of protein localization | *WDPCP, DZIP1, MCC* | 8.92E−02 | 9.91E−01 |
| GO:0042297~vocal learning | *CNTNAP2, NRXN1* | 9.94E−02 | 9.93E−01 |

Abbreviation: GO, gene ontology.

### *Genes under selection involved in biological processes*

We explored the functions of genes associated with various biological processes. The $P$ value of .05 was considered significant for GO annotation. We recovered 21 enriched GO groups ($P$ < .05; Table 4). The top 3 GO terms with the lowest $P$ values were "negative regulation of canonical Wnt signaling pathway" (GO:0090090, $P$ = 2.93E−03), "muscle contraction" (GO:0006936, $P$ = 1.22E−02) and "axon guidance" (GO:0007411, $P$ = 1.31E−02). The GO term "muscle contraction" was defined as a change in muscle geometry induced by the force generated within muscle tissue. Force generation requires a chemo-mechanical energy conversion process that is mediated by actin/myosin complex activity, which produces force through adenosine triphosphate hydrolysis. Myomesin genes (*MYOM1* and *MYOM2*) are expressed in muscle cells to stabilize the 3-dimensional conformation of the thick filament, and tropomodulin1 (*TMOD1*) protein binds and restrains the minus end of actin, fine-tuning the length of actin filaments in muscle and nonmuscle cells.[37] A previous study in horses also found that genomic regions with the most genetic differentiation in domesticated horses were enriched in genes involved in metabolism, muscle contraction, reproduction, signaling pathways, and behavior,[11] which supports our result. The GO term "axon guidance" is a neural development subdomain in relation to this process, through which neurons send out axons to achieve the correct targets. In this GO term, the axon guidance receptor gene (*ROBO1*) is a candidate gene for developmental dyslexia, and *ROBO1* polymorphisms are associated with functioning in the language acquisition system.[38,39] Abnormal expression of Contactin-4 (*CNTN4*) has been implicated in some cases of autism. The cadherin gene (*CDH4*) has a predictive pivotal role during brain segmentation and neuronal outgrowth. Therefore, genes in this GO term may be related to the trait of racing in domestic horses, which is supported by other genes related to exercise-induced stress that were found in selective sweep tests in thoroughbreds.[15] Previous studies have shown that intensive exercise training in thoroughbreds leads to gene expression changes related to metabolism and muscle structure.[40,41]

## Author Contributions

SZ and XD conceived of and designed the experiments. CZ, JC, YM, and CX analyzed the data. MG, AB, DG, PN, and XL contributed to sampling. HIA, CZ, YF, and HW contributed to the writing of the manuscript. All the authors reviewed and approved the final manuscript.

## REFERENCES

1. Vila C, Leonard JA, Gotherstrom A, et al. Widespread origins of domestic horse lineages. *Science*. 2001;291:474–477.
2. Warmuth V, Eriksson A, Bower MA, et al. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. *Proc Natl Acad Sci U S A*. 2012;109:8202–8206.
3. Kayar SR, Hoppeler H, Lindstedt SL, et al. Total muscle mitochondrial volume in relation to aerobic capacity of horses and steers. *Pflugers Arch*. 1989;413:343–347.
4. Lei CZ, Su R, Bower MA, et al. Multiple maternal origins of native modern and ancient horse populations in China. *Anim Genet*. 2009;40:933–944.
5. Yang Y, Zhu Q, Liu S, Zhao C, Wu C. The origin of Chinese domestic horses revealed with novel mtDNA variants. *Anim Sci J*. 2017;88:19–26.
6. Gemingguli M, Iskhan KR, Li Y, et al. Genetic diversity and population structure of Kazakh horses *(Equus caballus)* inferred from mtDNA sequences. *Genet Mol Res*. 2016;15.
7. Zhao YB, Zhang Y, Zhang QC, et al. Ancient DNA reveals that the genetic structure of the northern Han Chinese was shaped prior to 3,000 years ago. *PLoS ONE*. 2015;10:e0125676.
8. Ai H, Fang X, Yang B, et al. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. *Nat Genet*. 2015;47:217–225.
9. Wade CM, Giulotto E, Sigurdsson S, et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*. 2009;326:865–867.
10. Andersson L. How selective sweeps in domestic animals provide new insight into biological mechanisms. *J Intern Med*. 2012;271:1–14.
11. Der Sarkissian C, Ermini L, Schubert M, et al. Evolutionary genomics and conservation of the endangered Przewalski's horse. *Curr Biol*. 2015;25:2577–2583.
12. Schubert M, Jonsson H, Chang D, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication. *Proc Natl Acad Sci U S A*. 2014;111:E5661–1569.
13. Petersen JL, Mickelson JR, Rendahl AK, et al. Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet*. 2013;9:e1003211.
14. Pavlidis P, Zivkovic D, Stamatakis A, Alachiotis N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol*. 2013;30:2224–2234.
15. Park W, Kim J, Kim HJ, et al. Investigation of de novo unique differentially expressed genes related to evolution in exercise response during domestication in Thoroughbred race horses. *PLoS ONE*. 2014;9:e91418.
16. Orlando L, Ginolhac A, Zhang G, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*. 2013;499:74–78.
17. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–2120.
18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–1760.
19. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–1303.
20. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6:80–92.
21. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*. 2014;15:356.
22. Fumagalli M, Vieira FG, Linderoth T, Nielsen R. ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*. 2014;30:1486–1487.
23. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19:1655–1664.
24. Lin R, Du X, Peng S, et al. Discovering all transcriptome single-nucleotide polymorphisms and scanning for selection signatures in ducks (*Anas platyrhynchos*). *Evol Bioinform Online*. 2015;11:67–76.
25. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*. 2002;18:337–338.
26. Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*. 2003;4:P3.
27. Marth GT, Yu F, Indap AR, et al. The functional spectrum of low-frequency coding variation. *Genome Biol*. 2011;12:R84.
28. Rohrer DK, Kobilka BK. G protein-coupled receptors: functional and mechanistic insights through altered gene expression. *Physiol Rev*. 1998;78:35–52.
29. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335:823–828.
30. Das A, Panitz F, Gregersen VR, Bendixen C, Holm LE. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. *BMC Genomics*. 2015;16:1043.
31. Jansen T, Forster P, Levine MA, et al. Mitochondrial DNA and the origins of the domestic horse. *Proc Natl Acad Sci U S A*. 2002;99:10905–10910.
32. Luis C, Cothran EG, Oom Mdo M. Inbreeding and genetic structure in the endangered Sorraia horse breed: implications for its conservation and management. *J Hered*. 2007;98:232–237.
33. Petersen JL, Mickelson JR, Cothran EG, et al. Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PLoS ONE*. 2013;8:e54997.
34. McCue ME, Bannasch DL, Petersen JL, et al. A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet*. 2012;8:e1002451.
35. Huang J, Zhao Y, Shiraigol W, et al. Analysis of horse genomes provides insight into the diversification and adaptive evolution of karyotype. *Sci Rep*. 2014;4:4958.
36. Ahmad HI, Ahmad MJ, Adeel MM, Asif AR, Du X. Positive selection drives the evolution of endocrine regulatory bone morphogenetic protein system in mammals. *Oncotarget*. 2018;9:18435–18445.
37. Rao JN, Madasu Y, Dominguez R. Mechanism of actin filament pointed-end capping by tropomodulin. *Science*. 2014;345:463–467.
38. Bates TC, Luciano M, Medland SE, Montgomery GW, Wright MJ, Martin NG. Genetic variance in a component of the language acquisition device: ROBO1 polymorphisms associated with phonological buffer deficits. *Behav Genet*. 2011;41:50–57.
39. Hannula-Jouppi K, Kaminen-Ahola N, Taipale M, et al. The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia. *PLoS Genet*. 2005;1:e50.
40. McGivney BA, McGettigan PA, Browne JA, et al. Characterization of the equine skeletal muscle transcriptome identifies novel functional responses to exercise training. *BMC Genomics*. 2010;11:398.
41. McGivney BA, Eivers SS, MacHugh DE, et al. Transcriptional adaptations following exercise in thoroughbred horse skeletal muscle highlights molecular mechanisms that lead to muscle hypertrophy. *BMC Genomics*. 2009;10:638.