Genetic structure and genome-wide association study of the traditional Kazakh horses

A. Pozharskiy[1,2], A. Abdrakhmanova[1], I. Beishova[3], A. Shamshidin[3], A. Nametov[3], T. Ulyanova[3], G. Bekova[3], N. Kikebayev[3], A. Kovalchuk[3], V. Ulyanov[3], A. Turabayev[3], M. Khusnitdinova[1], K. Zhambakin[1], Z. Sapakhova[1], M. Shamekova[1]*, D. Gritsenko[1]*

[1] Institute of Plant Biology and Biotechnology, 050040 Almaty, Kazakhstan
[2] Al-Farabi Kazakh National University, Molecular Biology and Genetics, 050040 Almaty, Kazakhstan
[3] West Kazakhstan Agrarian and Technical University named after Zhangir khan, 090009, Uralsk, Kazakhstan,

* **Correspondence: d.kopytina@gmail.com; m.shamekova@gmail.com**

Abstract

Horses are traditionally used in Kazakhstan as a source of food and as working and saddle animals as well. Here, for the first time, microarray-based medium-density SNP genotyping of six traditionally defined types and breeds of indigenous Kazakh horses was conducted to reveal their genetic structure and find markers associated with animal size and weight. The results showed that the pre-defined separation between breeds and sampled populations was not supported by the molecular data. The lack of genetic variation between breeds and populations was revealed by the principal component analysis (PCA), ADMIXTURE, and distance based analyses, as well as the general population parameters expected and observed heterozygosity (H_e and H_o) and between group fixation index (F_{st}). The comparison with previously published data on global horse breed diversity revealed the relatively high level of individual diversity of Kazakh horses in comparison with the well-known foreign breeds. The Mongolian and Tuva breeds were identified as the closest horse landraces, demonstrating similar patterns of internal variability. The genome-wide association analysis was performed for animal size and weight as the traits directly related with meat productivity of horses. The analysis identified a set of 60 SNPs linked with horse genes involved in the regulation of processes of development of connective tissues and the bone system, neural system, immune system regulation, and other processes. The present study is novel and introduces Kazakh horses as a promising genetic source for horse breeding and selection.

Keywords

Equus caballus, Zhabe, Adai, Naiman, Mugalzar, Kushum, Kostanay, SNP, GWAS

Introduction

Horses (*Equus caballus* L.) have traditionally played a central role in the whole history of Kazakhstan and the Kazakh people. The eneolithic Botai culture (Northern Kazakhstan) contains arguably the earliest evidences of the use of horses by the local tribes (Levine, 1999), however, it remains disputable whether horses were domesticated or obtained by hunting (Outram et al., 2009, 2021; Taylor and Barrón-Ortiz, 2021). Genomic data revealed that Botai horses were closer to Przewalski's horses than to modern domestic lineages (Gaunitz et al., 2018), thus, even if the Neoilthic horse domestication had taken place at Botai, it had occurred independent of the main course of horse domestication (Kyselý and Peške, 2022). Nevertheless, horses became an important part of the steppe pastoralism and nomadism at the area of modern Kazakhstan and Central Asia as early as the Bronze Age (Frachetti and Benecke, 2009; Outram et al., 2012). From the Saka-Skythian tribes to three Kazakh Hordes, throughout the process of the Kazakh ethnogenesis, horses were not only an essential economical resource for the local Nomadic peoples (Chang, 2015), but also became an important part of the cultural legacy inherited by independent Kazakhstan, with rich symbolic connotations and influence on Kazakh language (Sarbassova, 2015).

Traditional Kazakh husbandry practices have not changed significantly for centuries and have been based on grazing and seasonal transhumance (migrations between *zhaylau* and *qystau*, summer and winter pastures, respectively). The conditions of horse pasturing have been kept close to natural: free grazing under the herdsman's control; in winter, horses find their fodder from under the snow—*tebindeu*. Herds of horses were the main measure of one's wellness; the owners selected the best animals or interchanged them with other breeders to keep and improve the horses' valuable qualities, such as strength, endurance, and the most important for nomads, meat and milk productivity. As a result of such folk selection over hundreds and thousands of years, the Kazakh horse was formed, with some traditional types distinguished based on their qualities and geographical distribution, the most known of which are Zhabe[1], Adai[2], and Naiman horses. Zhabe and Adai horses are the most common types. The Zhabe type is specialized for meat and milk production, and the Adai type has more expressed qualities of saddle horses; both these types originated in Western Kazakhstan and are used in all regions of the country (Dmitriev et al., 1989). The third type, Naiman, has been traditionally bred by the inhabitants of Dzhungar Alatau as the universal horses for use in mountainous conditions; this type is similar to Mongolian horses and considered the most subtle of Kazakh horses (Dyussegaliyev, 2022). All these types are well adapted for the traditional Kazakh methods of seasonal pasturing and transhumance (Barmintsev, 1958).

Traditional Kazakh horses became a progenitors of new breeds by crossing with the animals with desirable traits. Kostanay[3] breed was under development since 1887 year by crossing Kazakh mares with stallions of saddle breeds, Don, Astrakhan, Strelets, Thoroughbred, to combine their best qualities (Dmitriev et al., 1989). The breed was finally registered in 1951 (Nechayev et al., 2005), and work on its improvement has continued into the present. These horses have good saddle and draft characteristics and are suitable for both stable and steppe maintenance (Dmitriev et al., 1989).

The Kushum horses were bred between 1931 and 1976 by crossing Kazakh horses with Thoroughbreds, Don breeds, and their half-breeds, initially as military horses and later, after World War II, as steppe herd horses for meat and milk production. The resulting breed is versatile; the horses have high endurance and are capable of high live weight gai (Dmitriev et al., 1989).

The Mugalzhar breed was developed based on Kazakh horses between 1969 and 1998 to improve meat and milk productivity (Iskhan et al., 2019). Non-specific crossings and selection within the Zhabe breed allowed a significant increase of live weight (80–100 kg for mares, 100–120 kg for stallions, compared to the original horses), without changing technology of their maintenance (Dyussegaliyev, 2022).

A significant damage to the Kazakh horse population was caused by the Soviet policy of the involuntary collectivization and eradication of private animal ownership, and during the period between 1928 and 1958, the amount of Kazakh horses was reduced from 4,640,000 to 300,000 (Nechayev et al., 2005). According to the data of the Bureau of National statistics of the Agency for Strategic planning and reforms of the Republic of Kazakhstan (Bureau of National statistics, 2023), the total amount of horses (including foreign breeds) in the country was 1,666.4 thousands in 1991; after the drastic decrease in the 1990s, the number of horses was growing steadily and reached 3,489.8 thousands in 2021, due to the development of husbandry in Kazakhstan.

Kazakhstan is the second large horse meat producer in the world after China, however, it is mainly limited to the domestic market, as the country is not a significant exporter of horse meat (Jastrzębska et al., 2019). With the growing interest in horse meat as a safe and nutritious alternative to beef, despite a prejudicial attitude existing in many countries (Stanciu, 2015), Kazakhstan has a potential to become an important provider in the global horse meat market. This requires an extensive modernization of horse breeding to comply with the internationally acclaimed standards. An important aspect of such modernization is a wide implementation of the contemporary methods of molecular genetics and genomics in breeding practices to better understand the genetic structures of horse lines and breeds, improve classification and management of horse genotypes, assist selection

1 Variants of transliteration: Jabe, Dzhabe.
2 Variants of transliteration: Aday, Adaev.
3 Variant of transliteration: Kustanai.

using molecular markers associated with valuable traits, etc. The commercial microarray genotyping panels for animals contain tens or hundreds of thousands of SNP markers selected to reflect the total genetic variability, helping scan genomes for potentially important polymorphisms without expensive whole-genome sequencing. In Kazakhstan, SNP microarray genotyping was previously used to describe genetic structures of the local breeds of sheep, another animal of essential importance for the country (Pozharskiy et al., 2020; Zhumadillayev et al., 2022). For horses, the EquineSNP50 panel was developed and proven to be suitable for genome-wide association analysis and studies of horse diversity (McCue et al., 2012; Petersen et al., 2013a; b).

The scope of the present work was to explore genetic diversity of Kazakh horses using the Equine80k SNP microarray, the upgraded and expanded version of the EquineSNP50 array. The horses of three traditional types (Zhabe, Adai, and Naiman) and three derivative breeds (Kushum, Kostanay, and Mugalzhar) were sampled from herds in different regions of Kazakhstan to reveal their genetic variability between and within breeds. The previously published data of genetic diversity of horse breeds from all over the world was also used for comparison. Also, a genome-wide association study (GWAS) was performed to find SNPs associated with size and live weight of the horses. This is the first study involving high-density SNP array into an investigation of traditional Kazakh horses in Kazakhstan. The obtained results will increase understanding of the genetics of Kazakh horses and their place in the global diversity of horse breeds and lay a basis for further molecular and genomic studies.

Materials and methods

Sample collection and data acquisition

Genetic materials of three traditional types and three breeds of Kazakh horses (for convenience, we will further refer these as six breeds) were collected at 25 horse farms in different regions of Kazakhstan (Figure 1; Supplementary table 1). Hairs were sampled from horses' tails and/or manes and stored at +4°C until further use; the hair follicles were used for DNA extraction. DNA was isolated using the kit "DNK-Extran2" (Syntol, Russian Federation) following the manufacturer's protocol and quantified using Qubit 4 Fluoremeter with the Qubit dsDNA Broad Range reagent (Thermo Fisher Scientific, USA) for downstream SNP genotyping.

The measurements of horses were taken prior to hair sampling. The measurements included the height at the withers, oblique body length, chest circumference, cannon bone circumference, and body weight.

SNP genotyping data on the genetic diversity of foreign horse breeds (Petersen et al., 2013a) was retrieved from the Open Science Foundation repository (https://osf.io/gx42p/, accessed 20.07.2022), with the explicit permission from Dr. Jessica Petersen.

SNP genotyping and quality control

Genotyping was performed using Equine80k SNP array with the iScan system (Illumina, USA) according to the manufacturer's protocol. The assignments of genotypes and primary quality control were conducted using the Genotyping module of the GenomeStudio software (Illumina). The data was filtered using the following thresholds (the primary quality control): call rate ≥ 0.9, median GC score ≥ 0.8 for samples; call frequency ≥ 0.95 and GT score ≥ 0.7 for SNP. The indel markers presented in the array were also excluded. All data satisfying these criteria were exported and transformed to PLINK text input files (.ped + .map) using general data handling utilities of R (R Core Team, 2019). PLINK1.9 (Purcell et al., 2007; Chang et al., 2015) was further used to exclude SNPs with minor allele frequencies ≤ 0.05 and those deviating from Hardy-Weinberg equilibrium with a *p*-value threshold of $1 \cdot 10^{-10}$. The missing genotypes were imputed using BEAGLE (Browning et al., 2018).

Genetic structure analysis

All samples were tested for the internal and between-population structure using ADMIXTURE (Alexander and Lange, 2011) for K from 1 to 10 with 10 cross-validation replicates. Based on the results the outlier genotypes were identified and taken into account for any further analyses.

The general population genetic analysis was performed using PLINK1.9 and summarized using general R functions. It included the evaluation of linkage disequilibrium (LD), expected and observed heterozygosity (H_e, H_o), and between-population fixation index (F_{st}). The analysis was performed for the whole sample for breeds and populations. Only populations with more than 50 sampled individuals were considered for population-based analysis. The markers in the strong LD ($r^2 > 0.7$) were not considered for the population statistics.

The comparative analysis of the Kazakh horse genotypes with respect to the well-known foreign horse breeds was performed using data from (Petersen et al., 2013a). The dataset was transformed to fit the data format used throughout the present work using general R scripting tools and merged with our data. The merged dataset included 10 randomly selected individuals from each local population with more than 10 samples and all individuals from populations of the lower sample size. Additionally, the outlier genotypes were added to identify the possible sources of admixture. The merged dataset was finally filtered to include only SNPs with call frequency ≥ 0.95. A principal component analysis (PCA) was performed using PLINK and visualized using R with the 'ggplot2' package (Wickham, 2016). The ADMIXTURE analysis was run for K from 1 to 40, with 10 cross-validation replicates and visualized using CLUMPAK (Kopelman et al., 2015). The distance matrix was calculated using Manhattan distance in the 'dartR' package, and the neighbor joining tree was constructed using 'ape' package and visualized using FigTree (Rambaut, 2018) software.

Genome wide association analysis

The genome-wide association analysis was performed using PLINK1.9 supplemented with R scripting, including use of the 'ggplot2' package to create Manhattan plots.

The results of the genetic structure analysis were taken into account for the sample selection for GWAS. The association analysis was conducted for weight and size. The animals of the age below three years and the outliers were excluded from the analysis. Pearson's correlation test was used to ensure the independence of the phenotypic variables from age. The size variable was defined using measurements of the height at the withers (HW), oblique body length (OBL), chest circumference (CC), and cannon bone circumference (CBC). These parameters were normalized by subtracting the mean and dividing by the standard deviation, then PCA was performed, and the first component was selected as the new size variable.

The genome-wide association analysis was performed using a linear regression test with the adaptive Monte-Carlo permutation method of *p*-value correction for multiple comparisons (PLINK's '--linear perm' command). SNPs in strong linkage disequilibrium were excluded from the analysis ($r^2 > 0.7$).

The markers with the resulting corrected *p*-value below the threshold of 0.001 were annotated using the Variant Effect Predictor tool (VEP) (McLaren et al., 2016) and the DAVID web server (Huang et al., 2009; Sherman et al., 2022). The horse genome assembly EquCab3.0 (GCA_002863925.1) was used as a reference for annotation.

Results

Genotyping and genetic structure analysis

Samples of six Kazakh horse breeds were collected from 25 livestock farms in the North-Kazakhstan, West-Kazakhstan, East-Kazakhstan, Mangystau, Akmola, Kostanay, Zhambyl, Almaty, and Aktobe regions (Figure 1, Supplementary table 1). A total of 2,200 horses were sampled and processed for DNA isolation and SNP genotyping. As a result of selection based on quality control of extracted DNA samples and subsequently obtained SNP genotypes, 2,020 individuals were retained for the analysis. For 1,876 and 1,883 samples, respectively, measurements of size and weight were available.

A total of 74,116 SNPs for 2,020 samples satisfying the selected primary quality control criteria were obtained. Further filtering based on MAF and HWE left a total of 60,987 markers for further analyses.

All breeds and populations with 50 or more sampled individuals were analyzed using general population statistics, expected and observed heterozygosity, and pairwise F_{st} between breeds and populations (Table 1). All breeds and populations had very close values of heterozygosity; on average, 0.3462 and 0.3432, standard deviations 0.0051 and 0.0043, for expected and observed heterozygosity, respectively. All pairs of breeds and populations demonstrated F_{st} not exceeding 0.001, indicating a very low degree of differentiation between sampled groups. The lowest between-breed value was observed for Kostanay and Adai (F_{st} = 0.0003) and the highest value was between Kushsum and Adai (F_{st} = 0.005).

The analysis of the genetic structure within the whole sample of Kazakh horses with the ADMIXTURE algorithm confirmed the lack of differentiation between breeds and populations. Although the true *K* was not revealed by cross-validation, as the standard cross validation error did not reached the minimum value in *K* runs from 1 to 10, examination of the results allowed us to select *K* = 2 as the optimal structure (Figure 2, a). The results of *K* = 2 demonstrated that the sample was generally homogeneous, with the exception of some outlying genotypes. These outliers were taken into account for further analyses. The further results for *K* from 3 to 10 highlighted within-population variability and did not add information about genetic structure between populations, however the separation of the outliers was supported.

Data from Petersen et al. (2013) was used to put the Kazakh horse samples into a context of global horse diversity. As the number of individuals representing the foreign breeds was significantly lower than for Kazakh horse breeds, we achieved balanced data volumes by limiting the number of local horses to not exceed 10 per population; these individuals were sampled randomly. The outlier genotypes mentioned above were additionally selected to identify the probable source of the introgression. The total merged dataset included 35,419 SNPs for 1,176 animals, including 138 Zhabe horses, 74 Kostanay, 66 Mugalzhar, 60 Kushum, and 27 Aday horses. The 10-fold cross-validation test of the ADMIXTURE analysis identified *K* = 28 as an optimal number of clusters (Figure 2, b). The results demonstrated that Kazakh horses have had higher levels of individual variability compared to well-established foreign breeds. No clustering patterns to distinguished between Kazakh horse breeds were observed. Across the foreign breeds, the most similar to Kazakh horse patterns were observed in Tuva (TUVA) and Mongolian (MON) horses. The identified outlier specimens showed high similarity to the Thoroughbred breed (TB), indicated by orange throughout all *K*'s, however, with the optimal *K* = 28, the outliers of Kazakh horses displayed the high probability of the new cluster (shown pale yellow) with only a minor occurrence in Thoroughbred horses.

The results of the PCA (Figure 3) also demonstrated the low level of differentiation between Kazakh breeds. We separately considered the sample with outlier genotypes (Figure 3, b) and without them (Fig. 3, a). Figure 3, a, demonstrates that all Kazakh horses formed one group; the corresponding ellipses and central points were strongly overlapping. In Figure 3, b, the outlier genotypes were shifted towards Thoroughbred horses (populations from the UK and the USA). The breeds closest to Kazakh horses were Mongolian and Tuva horses, as well as the Caspian breed and the groups of Andalusian (AND, Spain) and South American horses. In the overall structure, Kazakh horses had a central position with respect to the directions of distribution of worldwide breeds.

In the neighbor joining tree, there were also no clear structures corresponding to Kazakh breeds; most samples were combined into a distinct heterogeneous group, including Mongolian and

Tuva horses (Figure 4; Supplementary figure 1). Unlike the results of other methods described above, only a few outlying genotypes diverged from the main group of genotypes. Four horses of the Kostanay, Mugalzhar and Aday breeds were placed close to Thoroughbred horses, and five horses of the Kushum, Mugalzhar, Zhabe, and Naiman breeds were placed between the American breed Morgan and a cluster of South American and Spanish horses.

Genome-wide association analysis

The genome-wide analysis of the association between SNP markers with horse body size and weight were performed for all animals with available phenotypic data, excluding horses less than three years of age and the outlying genotypes identified by the genetic structure analysis. The total number of animals was 1,533. To ensure that age did not have an impact as a covariate in the selected sample, all phenotypic variables were tested for Pearson's correlation against age. The standard body measurements, the height at the withers and oblique body length, displayed no significant correlation at a *p*-value threshold 0.05 (*p*-values of 0.8388 and 0.4211, respectively). Only weak correlations were revealed for chest circumference (0.0841, $p = 0.000507$), cannon bone circumference (0.1011, $p = 2.922 \cdot 10^{-5}$), and weight (0.1121, $p = 3.496 \cdot 10^{-6}$). Four body measurements were combined into a single variable using PCA after normalization; the first principal component describing 81% of total variation was selected as a new size variable.
The association analysis was performed using a linear regression algorithm implemented in PLINK software with the adaptive correction of *p*-values based on a Monte-Carlo permutation test. As a result, a total of 81 and 84 SNPs showed significant associations with size and weight, respectively, at a selected significance level 0.001 (Figure 5). Of these combined SNP sets, 60 variants were associated with the known horse genes using Ensembl's VEP server and with the corresponding biological processes using DAVID (Table 3, Supplementary table 3). Surprisingly, there was almost no overlap between two sets of markers associated with the respective traits. Only two variants, BIEC2_117960 and BIEC2-187196, showed significant association with both traits. While the former marker was linked with the gene *OR4C269P*, which has had no available gene ontology annotation, the latter was identified in relation to the gene of ecto-5'-nucleotidase (*NT5E*) involved into metabolism of adenosine phosphates. In general, the identified genes play regulatory or signaling roles in a wide range of processes, from the cellular level to the whole organism. The most essential or basic, in our opinion, gene ontology (GO) terms for biological processes (BP) are listed in the Table 4; the complete annotations produced by DAVID are available in Supplementary table 3. The genes *BMP6*, *DDR3*, and *CREB3L1* are involved in the development and metabolism of connective tissues, including bones. These genes were linked with SNPs significantly associated with weight. The *BMP6* gene had three linked SNPs, which was the highest number across all genes. Several genes, *DPF1*, *GNAT3*, *NEGR1*, etc., have been annotated as involved into the development of the neural system. More specifically, the *NEGR1* gene was associated with feeding and locomotory behavior, and the *GNAT3* gene was associated with taste perception. The genes *BMP6*, *RELA1*, *AIM2*, *PDE4D*, and *IGF1R* were associated with regulation of the immune system, in addition to other processes,. The *EIF2AK4* gene was associated with the cellular response to cold stress and protein starvation.

Discussion

This study of Kazakh horses involved three traditional types (Zhabe, Adai, and Naiman) and three relatively recent breeds (Kostanay, Kushum, Mugalzhar), which were derived from the older lines. Except for the Kostanay horses, which are subjected to stable maintenance, control of breeding, and certification of offspring, the other five breeds are being reproduced freely during perennial herd pasturing. However, the genetic data did not support this difference of breeding strategies. All six breeds, including Kostanay, form one group without a notable subdivision corresponding to breeds and populations. This observation was supported by the distance-based and ADMIXTURE

clustering methods as well as principal component analysis. Also, pairwise F_{st} values between breeds and populations were low, indicating almost free gene flow between populations and the factual absence of between-breed boundaries. Hence, we conclude that the studied traditional types and breeds cannot be discriminated from each other on the genetic level, and all Kazakh horses should be considered one 'breed'. Yet, we question the applicability of the term 'breed' to Kazakh horses because of the revealed high level of individual variability. We assume that the Kazakh horses represent a sum of relatively unspecialized genetic lines. In some sense, all Kazakh horses are close to the natural population. Indeed, the traditional Kazakh way of pasturing doesn't imply strict mating control in herds; gene flow between populations (herds) occurs through trading and exchange of genotypes by horse owners, and the means of modern transporting capabilities reduces the influence of geographical isolation. The close values of expected and observed heterozygosity in breeds and populations of sufficient sizes indicate the absence or low impact of possible selective factors.

Interestingly, the horses of the Tuva and Mongolian breeds (Petersen et al., 2013a)) display similar patterns of individual variability and close proximity to Kazakh horses. According to the published results (Petersen et al., 2013a), these breeds also had higher levels of expected heterozygosity in comparison to many breeds from around the world. Although the small volume of available data for these breeds limits the possible comparisons, we could speculate that Kazakh, Tuva and Mongolian horses could be grouped together. Historically, the nomadic peoples of Central Asia, Mongolia, and Siberia had close interconnections and shared some common ways in horse breeding. Thus, we suggest that the Kazakh, Mongolian, and Tuva horses, and possibly some other breeds not considered here, could be parts of a broader defined landrace, which we would designate as "Nomads' horses". In the results of the PCA, all these genotypes had a central position in respect to the main direction of the distribution of breeds. In the study of the Chinese populations of Mongolian horses, a similar PCA pattern was previously observed, with the exception of Kazakh horses (Han et al., 2019). The authors of that study discussed the high homogeneity and the lack of genetic structure between the considered populations and hypothesized that the Mongolian horses may represent the closest descents of the earliest domesticated horse lines. With our results revealing close proximity of Kazakh horses with the Mongolian and Tuva horses, and similar genetic patterns of Kazakh horses, we consider the mentioned study to be in line with our assumption about "Nomads' horses". The analysis of horse mitochondrial DNA revealed that, to a significant extent, horses domesticated in the Central Asian region have retained their genetic diversity (Cieslak et al., 2010). We assume that the traditional nomadic ways of horse husbandry played an important role in avoiding a bottleneck effect as they imply only the limited involvement of humans in selection and reproduction.

The analysis of genetic structure of Kazakh horses revealed the presence of significantly deviating genotypes, mainly in populations of Zhabe, Kushum, and Mugalzhar horses. The ADMIXTURE analysis and PCA helped attribute these horses to hybridization with Thoroughbred horses. Indeed, this breed has been popular among breeders in the country, as well as in the world, as the reference saddle and breed horse. Thoroughbred stallions have been imported since the beginning of the twentieth century and are used in crosses to improve the saddle quality of Kazakh horses. However, prior to sampling, all animals used in the work were attested as purebred Kazakh horses of respective breeds and types. This fact raises important questions about the present state of breeding control and certification in Kazakhstan. Our study revealed some significant issues. First, the hybridization of Kazakh horses with other horse breeds, including foreign ones, has been poorly documented, hence the difficulty of tracing back the admixed lineages and their correct identification. Second, the classification of the types and breeds of Kazakh is based on morphological traits and geographical distribution; however, these features do not provide unambiguous boundaries between breeds and, thus, their definition relies mostly on the subjective experience of the breeders.

The lack of genetic structure across the studied breeds and populations allowed us to combine all horses, except for hybrid genotypes discussed above, for association analysis. To date, most genome-wide association studies on horses have focused on racing performance (Littiere et al., 2020; Bailey et al., 2022) and health (Raudsepp et al., 2019). The traits related with food quality of horses have remained out of focus of horse genomics, as horse products, mainly meat, remain exotic or

even marginal in many countries (Cieslak et al., 2010). Thus, the study of genetics of such traits is novel, not only for Kazakhstan. Here we tested a set of SNP markers for association to the most general parameters related with meat productivity, live weight and size of an animal. Sixty SNPs were found to be associated with either of these two traits and linked with the functionally annotated horse genes. The set of identified genes included genes involved in various biological processes as regulatory and signal factors. The interesting notion was that almost all significant annotations were related to size or weight independently, despite the obvious correlation between these traits. Among all functionally annotated genes some certain aspects of biological processes potentially related with the traits of interest can be noted. First, the development of connective tissues and bone system, which are crucial for an animal to support its weight and size. Second, the development of the neural system; the more specific effects of the genes *GNAT3* and *NEGR1* on the fodder preferences of horses and so, indirectly, on their growth, could be an interesting topic of discussion in future studies. Third, the regulation of immune processes, which influence growth by affecting general health. However, it should be kept in mind that the gene annotations by gene ontology have been based mainly based on the data on humans and the model animals (mouse, rat, etc.); thus, the true physiological roles of the identified genes in horses may vary. Also, the possible associations of variants that remained unannotated require future clarification with the updated annotation data for horse genomes.

Conclusions

The Kazakh horses, with traditionally defined types and breeds, represent a single landrace with an absence of clearly expressed internal genetic structure. The traditional ways of horse breeding and husbandry in Kazakhstan have led to the formation of a relatively unspecialized landrace with genetic properties similar to a wild-living population. Along with the genetically similar Mongolian and Tuva breeds, the original genetic pool of Kazakh horses can potentially serve as a new source of genetic material for horse breeding, for the development of new breeds, or the improvement of existing lineages.

Acknowledgments

The study was funded within the framework of the research project AP14870614 «Genetic marking of productive traits of the Kazakh horse of the Dzhabe type based on genome wide coverage SNP genotyping» (the Ministry of science and higher education of the Republic of Kazakhstan) and the targeted funding program BR10764999 “Development of technologies for effective management of the breeding process and preservation of the gene pool in horse breeding” (the Ministry of Agriculture of the Republic of Kazakhstan).

References

Alexander, D.H., and Lange, K., 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12, 246.

Bailey, E., Petersen, J.L., and Kalbfleisch, T.S., 2022. Genetics of Thoroughbred Racehorse Performance. Annual Review of Animal Biosciences 10, 131–150.

Barmintsev, Yu.N., 1958. Evolution of horse breeds in Kazakhstan. Experience of zootechnical study of the poroblem of breed formation[Evoljucija konskih porod v Kazahstane. Opyt zootehnicheskogo issledovanija problemy porodoobrazovanija]. Kazgosizdat, Alma-Ata.

Browning, B.L., Zhou, Y., and Browning, S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. American Journal of Human Genetics 103, 338–348.

Bureau of National statistics, 2023. Statistics of agriculture, forestry, hunting and fisheries. Retrieved on 20 February 2023, from https://stat.gov.kz/official/industry/14/statistic/8.

Chang, C., 2015. The study of nomads in the republic of Kazakhstan. In The Ecology of Pastoralism, pp. 17–40. University Press of Colorado, USA.

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, s13742-015-0047–8.

Cieslak, M., Pruvost, M., Benecke, N., Hofreiter, M., Morales, A., Reissmann, M., and Ludwig, A., 2010. Origin and History of Mitochondrial DNA Lineages in Domestic Horses. PLOS ONE 5, e15311.

Dmitriev, N.G., Ėrnst, L.K., and Food and Agriculture Organization of the United Nations (Eds), 1989. Animal genetic resources of the USSR. Food and Agriculture Organization of the United Nations, Rome.

Dyussegaliyev, M.Zh., 2022. The genotypes of herd horses of the west region of Kazakhstan. The Agriculture and Ecosystems in Modern World: Regional and Inter countries' research 1, 33–43.

Frachetti, M., and Benecke, N., 2009. From sheep to (some) horses: 4500 years of herd structure at the pastoralist settlement of Begash (south-eastern Kazakhstan). Antiquity 83, 1023–1037.

Gaunitz, C., Fages, A., Hanghøj, K., Albrechtsen, A., Khan, N., Schubert, M., Seguin-Orlando, A., Owens, I.J., Felkel, S., Bignon-Lau, O., de Barros Damgaard, P., Mittnik, A., Mohaseb, A.F., Davoudi, H., Alquraishi, S., Alfarhan, A.H., Al-Rasheid, K.A.S., Crubézy, E., Benecke, N., Olsen, S., Brown, D., Anthony, D., Massy, K., Pitulko, V., Kasparov, A., Brem, G., Hofreiter, M., Mukhtarova, G., Baimukhanov, N., Lõugas, L., Onar, V., Stockhammer, P.W., Krause, J., Boldgiv, B., Undrakhbold, S., Erdenebaatar, D., Lepetz, S., Mashkour, M., Ludwig, A., Wallner, B., Merz, V., Merz, I., Zaibert, V., Willerslev, E., Librado, P., Outram, A.K., and Orlando, L., 2018. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science 360, 111–114.

Han, H., Bryan, K., Shiraigol, W., Bai, D., Zhao, Y., Bao, W., Yang, S., Zhang, W., MacHugh, D.E., Dugarjaviin, M., and Hill, E.W., 2019. Refinement of Global Domestic Horse Biogeography Using Historic Landrace Chinese Mongolian Populations. Journal of Heredity 110, 769–781.

Huang, D.W., Sherman, B.T., and Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4, 44–57.

Iskhan, K.Zh., Kalashnikov, V.V., Akimbekov, A.R., Mongush, S.D., Demin, V.A., Rzabayev, T.S., Nesipbaeva, A.K., Zhilkybaeva, M.M., and Zhikishev, Y.K., 2019. Zootechnic characteristics of modern populations of Mugalzhar horse breed. Bulletin of the National Academy of Sciences of the Republic of Kazakhstan. 6, 75–82.

Jastrzębska, E., Daszkiewicz, T., Górecka-Bruzda, A., and Feliś, D., 2019. Current situation and prospects for the horse meat market in Poland and the world. Medycyna Weterynaryjna 75, 6203–2019.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I., 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources 15, 1179–1191.

Kyselý, R., and Peške, L., 2022. New discoveries change existing views on the domestication of the horse and specify its role in human prehistory and history – a review. Archeologické rozhledy 74.

Levine, M.A., 1999. Botai and the Origins of Horse Domestication. Journal of Anthropological Archaeology 18, 29–78.

Littiere, T.O., Castro, G.H.F., Rodriguez, M. del P.R., Bonafé, C.M., Magalhães, A.F.B., Faleiros, R.R., Vieira, J.I.G., Santos, C.G., and Verardo, L.L., 2020. Identification and Functional Annotation of Genes Related to Horses' Performance: From GWAS to Post-GWAS. Animals 10, 1173.

McCue, M.E., Bannasch, D.L., Petersen, J.L., Gurr, J., Bailey, E., Binns, M.M., Distl, O., Guérin, G., Hasegawa, T., Hill, E.W., Leeb, T., Lindgren, G., Penedo, M.C.T., Røed, K.H., Ryder, O.A., Swinburne, J.E., Tozaki, T., Valberg, S.J., Vaudin, M., Lindblad-Toh, K., Wade, C.M.,

and Mickelson, J.R., 2012. A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. PLOS Genetics 8, e1002451.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biology 17, 122.

Nechayev, I.N., Zhumagul, A.E., Sizonov, G.V., Zhaytapov, T., Kikebayev, N.A., and Nurushev, M.Zh., 2005. Kazakh horse: past, present, future [Kazahskaja loshad': proshloe, nastojashhee, budushhee]. Edelveis, Almaty.

Outram, A., Bendrey, R., Evershed, R.P., Orlando, L., and Zaibert, V.F., 2021. Rebuttal of Taylor and Barrón-Ortiz 2021 Rethinking the evidence for early horse domestication at Botai.

Outram, A.K., Kasparov, A., Stear, N.A., Varfolomeev, V., Usmanova, E., and Evershed, R.P., 2012. Patterns of pastoralism in later Bronze Age Kazakhstan: new evidence from faunal and lipid residue analyses. Journal of Archaeological Science 39, 2424–2435.

Outram, A.K., Stear, N.A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., Thorpe, N., and Evershed, R.P., 2009. The Earliest Horse Harnessing and Milking. Science 323, 1332–1335.

Petersen, J.L., Mickelson, J.R., Cothran, E.G., Andersson, L.S., Axelsson, J., Bailey, E., Bannasch, D., Binns, M.M., Borges, A.S., Brama, P., Machado, A. da C., Distl, O., Felicetti, M., Fox-Clipsham, L., Graves, K.T., Guérin, G., Haase, B., Hasegawa, T., Hemmann, K., Hill, E.W., Leeb, T., Lindgren, G., Lohi, H., Lopes, M.S., McGivney, B.A., Mikko, S., Orr, N., Penedo, M.C.T., Piercy, R.J., Raekallio, M., Rieder, S., Røed, K.H., Silvestrelli, M., Swinburne, J., Tozaki, T., Vaudin, M., Wade, C.M., and McCue, M.E., 2013a. Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLOS ONE 8, e54997.

Petersen, J.L., Mickelson, J.R., Rendahl, A.K., Valberg, S.J., and Andersson, L.S., 2013b. Genome-wide analysis reveals selection for important traits in domestic horse breeds. PLoS Genet 9, 1003211.

Pozharskiy, A., Khamzina, A., Gritsenko, D., Khamzina, Z., Kassymbekova, S., Karimov, N., Karymsakov, T., and Tlevlesov, N., 2020. SNP genotyping and population analysis of five indigenous Kazakh sheep breeds. Livestock Science 241, 104252.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., and Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559–575.

R Core Team, 2019. R: A Language and Environment for Statistical Computing.

Rambaut, A., 2018. FigTree.

Raudsepp, T., Finno, C.J., Bellone, R.R., and Petersen, J.L., 2019. Ten years of the horse reference genome: insights into equine biology, domestication and population dynamics in the post-genome era. Animal Genetics 50, 569–597.

Sarbassova, G., 2015. Language and identity in Kazakh horse culture. bilig 75, 227–248.

Sherman, B.T., Hao, M., Qiu, J., Jiao, X., Baseler, M.W., Lane, H.C., Imamichi, T., and Chang, W., 2022. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research 50, W216–W221.

Stanciu, S., 2015. Horse Meat Consumption − Between Scandal and Reality. Procedia Economics and Finance 23, 697–703.

Taylor, W.T.T., and Barrón-Ortiz, C.I., 2021. Rethinking the evidence for early horse domestication at Botai. Scientific Reports 11, 1–11.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Zhumadillayev, N., Dossybayev, K., Khamzina, A., Kapasuly, T., Khamzina, Z., and Tlevlesov, N., 2022. SNP Genotyping Characterizes the Genome Composition of the New Baisary Fat-Tailed Sheep Breed. Animals 12, 1468.

Table 1. Population genetic statistics of horse breeds and populations. Population with less than 50 sampled individuals are not shown, but accounted for breeds. N – number of samples, H_e – expected heterozygosity, H_o – observed heterozygosity.

				Pairwise Wright's F_{st} between types/breeds															
Breed/Population	**N**	**H_e**	**H_o**	**Adai**			**Zhabe**						**Mugalzhar**				**Kushum**	**Naiman**	**Kostanay**
					Kuanysh	KozhyrAta		AgroDamu	Kalka	Zhaksylyk	Anar	Akzhar		Orazbay	Bakytbek	Karatomar		Alem	TokZhailau
Adai	**303**	**0.3438**	**0.3402**	**0.0000**															
Kuanysh	208	0.3442	0.3424		0.0000	0.0033													
KozhyrAta	95	0.3396	0.3355		0.0033	0.0000													
Zhabe	**631**	**0.3496**	**0.3453**	**0.0011**			**0.0000**												
AgroDamu	153	0.3438	0.3389		0.0019	0.0009		0.0000											
Kalka	143	0.3554	0.3444		0.0069	0.0086		0.0056	0.0000										
Zhaksylyk	86	0.3423	0.3412		0.0030	0.0018		0.0022	0.0081	0.0000									
Anar	75	0.3392	0.3410		0.0050	0.0012		0.0022	0.0099	0.0016	0.0000								
Akzhar	64	0.3446	0.3485		0.0025	0.0038		0.0031	0.0079	0.0011	0.0034	0.0000							
Mugalzhar	**584**	**0.3456**	**0.3402**	**0.0010**			**0.0005**						**0.0000**						
Orazbay	231	0.3428	0.3402		0.0023	0.0014		0.0007	0.0068	0.0009	0.0014	0.0022		0.0000					
Bakytbek	192	0.3493	0.3409		0.0029	0.0025		0.0015	0.0029	0.0024	0.0032	0.0027		0.0015	0.0000				
Karatomar	161	0.3414	0.3394		0.0023	0.0016		0.0014	0.0082	0.0025	0.0030	0.0039		0.0014	0.0030	0.0000			
Kushum	**226**	**0.3564**	**0.3514**	**0.0050**			**0.0029**						**0.0043**				**0.0000**		
Alem	226	0.3564	0.3514		0.0048	0.0070		0.0046	0.0015	0.0063	0.0088	0.0054		0.0056	0.0022	0.0065		0.0000	
Naiman	**158**	**0.3462**	**0.3472**	**0.0014**			**0.0016**						**0.0016**				**0.0046**	**0.0000**	
TokZhailau	158	0.3462	0.3472		0.0013	0.0037		0.0018	0.0066	0.0030	0.0046	0.0026		0.0023	0.0027	0.0017		0.0046	0.0000
Kostanay	**116**	**0.3451**	**0.3419**	**0.0003**			**0.0009**						**0.0011**				**0.0037**	**0.0016**	**0.0000**

Table 2. Foreign breeds used for comparative analysis (data by (Petersen et al., 2013a))

Abbreviation	Breed name	Country of origin*	Number of samples
AKTK	Akhal Teke	Turkmenistan	19
AND	Andalusian	Spain	18
ARR	Arabian	Arabia	24
BEL	Belgian	Belgium	30
CSP	Caspian	Iran	18
CLYD	Clydesdale	UK	24
EXMR	Exmoor	UK	24
FELL	Fell Pony	UK	21
FIN	Finnhorse	Finland	27
FLCR	Florida Cracker	USA	7
FM	Franches-Montagnes	Switzerland	19
FT	French Trotter	France	17
HAN	Hanoverian	Germany	15
ICE	Icelandic	Iceland	25
LUST	Lusitano	Portugal	24
MNGP	Mangalarga Paulista	Brazil	15
MARM	Maremmano	Italy	24
MINI	Miniature	Unknown	21
MON	Mongolian	Mongolia	19
MOR	Morgan	USA	40
NFST	New Forest Pony	UK	15
NSWE	North Swedish Horse	Sweden	19
NORF	Norwegian Fjord	Norway	21
PT	Paint	USA	25
PERC	Percheron	France	23
PERU	Peruvian Paso	Peru	20
PRPF	Puerto Rican Paso Fino	Puerto Rico	21
QH	Quarter Horse	USA	40
SB	Saddlebred	USA	25
SHET	Shetland	UK	27
SHR	Shire	UK	23
STBDNor	Standardbred (Norway)	USA	25
STBDUS	Standardbred (US)	USA	15
SZWB	Swiss Warmblood	Switzerland	14
TB_UK	Thoroughbred (UK/Ire)	UK	19
TB_US	Thoroughbred (US)	UK	17
TUVA	Tuva	Russia	15

* According to the original publication

Table 3. Annotated SNP associated with horse weight and size.

Variant ID	Location	Genes	Corrected *p*-value*		Gene onthology term**	Biological process (GO:BP)
			Weight	Size		
BIEC2_7947	1:17461309	*ABLIM1*	-	0.0003862	GO:0007010	cytoskeleton organization
Affx-101142060	31:8550289	*ACAT2*	0.0009455	-	GO:0006635	fatty acid beta-oxidation
AX-104155593	26:41616293	*ADARB1*	-	0.0009522	GO:0006382	adenosine to inosine editing
BIEC2_901539	5:34710217	*AIM2*	0.0007969	-	GO:0002218	activation of innate immune response
BIEC2_19530	1:41720677	*ASAH2*	-	0.0003041	GO:0042759	long-chain fatty acid biosynthetic process
BIEC2_516275	20:8331534		0.000298	-		
BIEC2_516282	20:8332816	*BMP6*	$8.805*10^{-5}$	-	GO:0000122	negative regulation of transcription from RNA polymerase II promoter
BIEC2_516343	20:8424937		0.0002289	-		
BIEC2_209409	13:9735624	*COL26A1*	0.000467	-	GO:0010811	positive regulation of cell-substrate adhesion
BIEC2_486667	2:69819186	*CPE*	0.0008825	-	GO:0006518	peptide metabolic process
BIEC2_174862	12:11000582	*CREB3L1*	0.0005222	-	GO:0006357	regulation of transcription from RNA polymerase II promoter
BIEC2_900752	5:31687044	*DDR2*	-	0.0004624	GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway
BIEC2_103610	10:9449473	*DPF1*	0.0002326	-	GO:0045944	positive regulation of transcription from RNA polymerase II promoter
Affx-102382610	1:149996219	*EIF2AK4*	-	0.0004106	GO:0000077	DNA damage checkpoint
BIEC2_1061054	8:68483265	*EPG5*	0.000296	-	GO:0006862	nucleotide transport
BIEC2_252279	14:7810156	*ERGIC1*	0.0009707	-	GO:0016192	vesicle-mediated transport
BIEC2_515625	20:6838695	*F13A1*	-	0.0003289	GO:0018149	peptide cross-linking
BIEC2_453812	2:4970163	*FAM151A*	0.0008247	-	-	-
TBIEC2_1007439	6:57510178	*FGD4*	0.0006988	-	GO:0007010	cytoskeleton organization
BIEC2_489911	2:75633527	*FNIP2*	-	0.0003669	GO:0001932	regulation of protein phosphorylation
BIEC2_645513	24:33029319	*FOXN3*	-	0.0004419	GO:0045892	negative regulation of transcription, DNA-templated
BIEC2-411146	18:38386190	*GALNT5*	-	0.0001543	GO:0006486	protein glycosylation
BIEC2_736048	28:25672835	*GAS2L3*	0.0001233	-	GO:0000226	microtubule cytoskeleton organization
BIEC2_842473	4:814418	*GNAT3*	-	0.0007865	GO:0007188	adenylate cyclase-modulating G-protein coupled receptor signaling pathway
BIEC2_124717	10:53976398	*GRIK2*	-	0.0006631	GO:0050804	modulation of synaptic transmission
BIEC2_275332	14:90328895	*HEXB* *GFM2*	0.0008033	-	GO:0001501 GO:0032543	skeletal system development mitochondrial translation
UKUL191	1:105140620	*IGF1R*	0.0001121	-	GO:0008284	positive regulation of cell proliferation
BIEC2_155534	11:45542428	*INPP5K* *PITPNA*	0.0007336	-	GO:0001933 -	negative regulation of protein phosphorylation -
BIEC2_908994	5:53136645	*KCND3*	-	0.000701	GO:0071805	potassium ion transmembrane transport
BIEC2-144059	11:22985500	*MED1*	0.0007038	-	GO:0000902	cell morphogenesis
AX-105000894	7:96769215	*METTL15*	-	0.0007262	GO:0032259	methylation
Affx-102902050	5:87153725	*NEGR1*	-	0.0003762	GO:0010976	positive regulation of neuron projection development
BIEC2_357099	16:68809463	*NEK11*	-	0.0008988	GO:0016572	histone phosphorylation
BIEC2_117960	10:39882171	*NT5E*	0.0003571	8.929×10^{-5}	GO:0009166	nucleotide catabolic process
BIEC2_554615	20:29466156	*OR12D2N*	-	0.0003629	GO:0050911	detection of chemical stimulus involved in sensory perception of smell
BIEC2-187196	12:17084040	*OR4C269P*	1.7×10^{-5}	0.0007977	-	-
BIEC2_461433	2:21106066	*OSCP1*	0.000956	-	GO:1990961	drug transmembrane export
BIEC2_553357	21:13990313	*PDE4D*	-	0.0003479	GO:0007165	signal transduction
BIEC2_952801	6:46670386	*PIK3C2G*	-	0.0003862	GO:0048015	phosphatidylinositol-mediated signaling
BIEC2_583163	22:15520818	*PLCB1*	0.0005666	0.000792	GO:0000086	G2/M transition of mitotic cell cycle
BIEC2_331272	16:18885332	*PPP4R2*	0.0007754	-	GO:0006470	protein dephosphorylation
BIEC2_651775	24:44695504	*RCOR1*	0.0008017	-	GO:0016575	histone deacetylation
BIEC2_207487	12:29515936	*RELA* *SIPA1*	-	0.0009667	GO:0006357 GO:0051056	regulation of transcription from RNA polymerase II promoter regulation of small GTPase mediated signal transduction
BIEC2_844810	4:4272449	*RELN*	-	0.0008289	GO:0001764	neuron migration
BIEC2_606629	22:5228365	*RIN2*	-	0.0008374	GO:0010595	positive regulation of endothelial cell migration
BIEC2_1006740	7:74196933	*RRM1*	0.0006299	-	GO:0006260	DNA replication
AX-104552893	9:19411659	*RRS1*	-	0.0009052	GO:0000027	ribosomal large subunit assembly
Affx-101191469	1:78170974	*SLC35F3*	0.0003044	-	GO:0015888	thiamine transpor
BIEC2_835239	31:8606400	*SOD2*	0.0002175	-	GO:0001315	age-dependent response to reactive oxygen species
BIEC2_878560	31:8606149		0.0001386	-		
BIEC2_154252	11:42866856	*SPAG5*	-	0.0008988	GO:0051301	cell division
BIEC2_736519	28:26656456	*SYCP3*	0.0005929	-	GO:0051321	meiotic cell cycle
BIEC2_1063214	8:76414093	*TCF4*	0.0003984	-	GO:0006357	regulation of transcription from RNA polymerase II promoter
BIEC2_504893	2:106480107	*TRPC3*	0.0008173	-	GO:0070588	calcium ion transmembrane transport
BIEC2_1028408	8:10248408	*TTC28*	-	0.0008527	GO:0007346	regulation of mitotic cell cycle
BIEC2_50291	1:115156781	*TUBGCP5*	-	0.0009209	GO:0007020	microtubule nucleation,
BIEC2_988205	7:26791905	*UBE4A*	-	0.0004775	GO:0000209	protein polyubiquitination
BIEC2_773787	28:26943895	*WASHC3*	-	0.0006032	-	-
BIEC2_1027876	8:9374074	*ZMAT5*	5.608×10^{-5}	-	-	-
BIEC2_1066438	8:92652089	*ZNF407*	-	0.0004628	GO:0010468	regulation of gene expression

* Only *p*-values below the selected significance threshold 0.001 are shown

** Only one term per gene is shown. See Supplementary file 1 for the complete DAVID annotation report, including Gene Onthology for cellular components (CC) and molecular processes (MP)

Figure 1. Geographical distribution of horse farms sampled in the study. 1 - Kuanysh, 2 - Kozhyr-Ata, 3 – Alem, 4 – Orazbay, 5 – Bakytbek, 6 – Kara-Tomar, 7 – Agro-Damu, 8 – Kalka, 9 – Zhaksylyk, 10 – Anar, 11 – Akzhar, 12 – Sayahat, 13 – Kyzylsok, 14 – Ernur, 15 – Orazaly, 16 – Tok-Zhailau, 17 – Kazakh Tulpary, 18 – ChV, 19 – Tasmukhambetov, 20 – Kostanay Tulpary, 21 – Stepnoye, 22 – Kabazhaev, 23 – Baranovskiy, 24 – Akhmetov, 25 – Zhussupov.

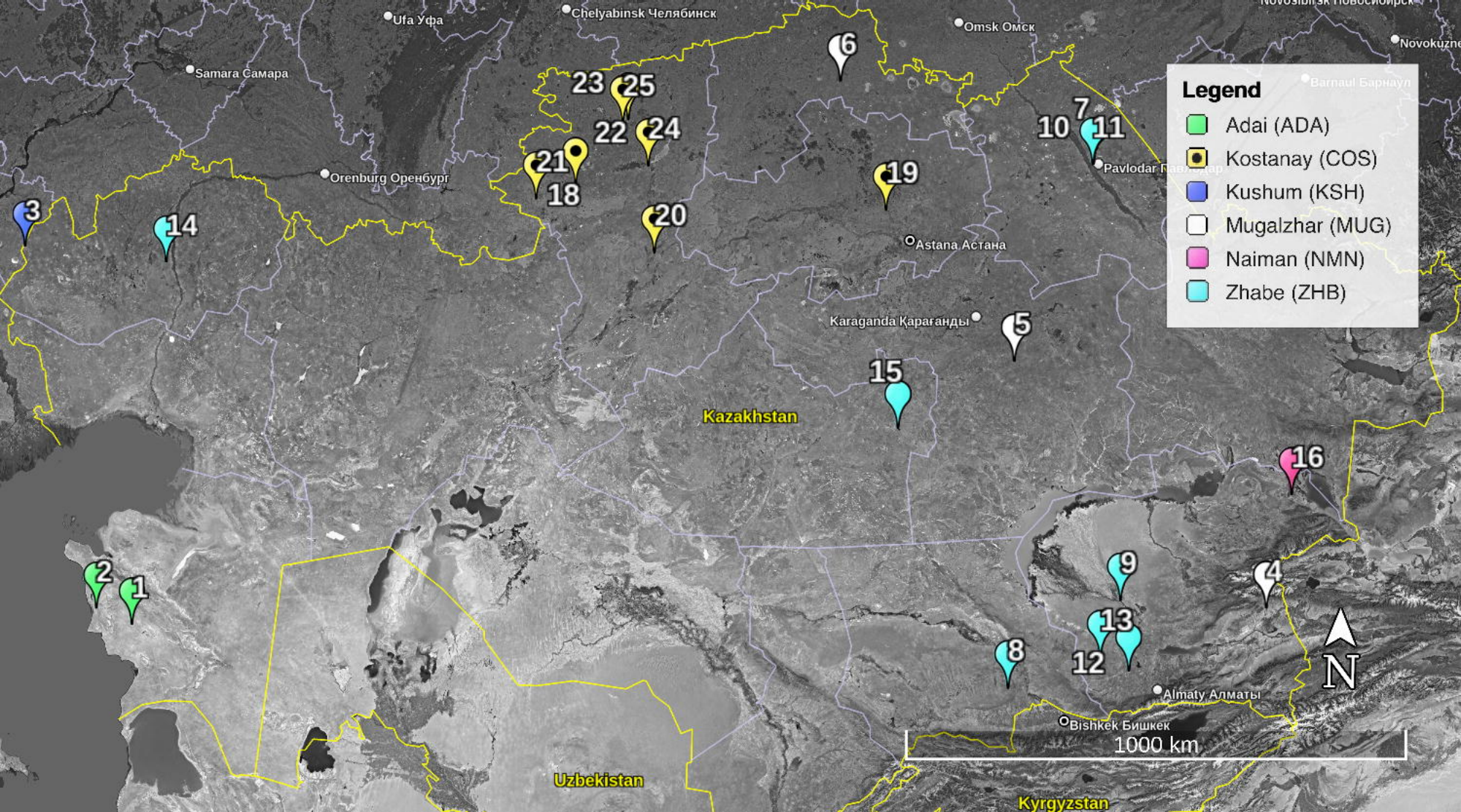
Figure 2. Results of ADMIXTURE analysis. A) complete sample of Kazakh horses of six breeds; B) reduced sample of Kazakh horses in comparison with foreign breeds (data (Petersen et al., 2013a))

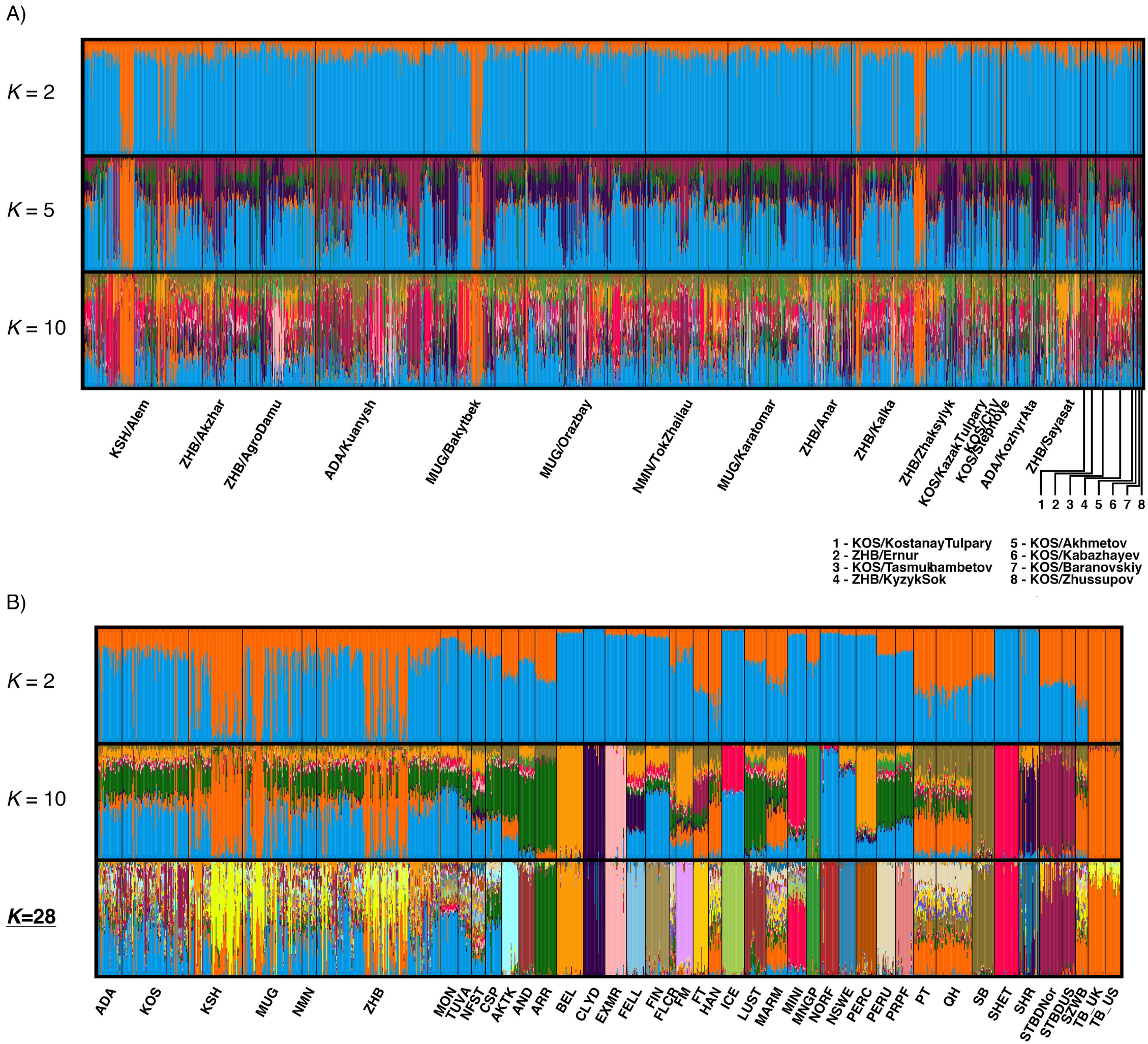
Figure 3. Principal component analysis of Kazakh horses in comparison with foreign breeds. A) outlier genotypes excluded; B) outlier genotypes included. The ellipses indicate the boundaries for points corresponding to breeds, with the confidence level 0.95. The positions of the breed labels correspond to the central points.

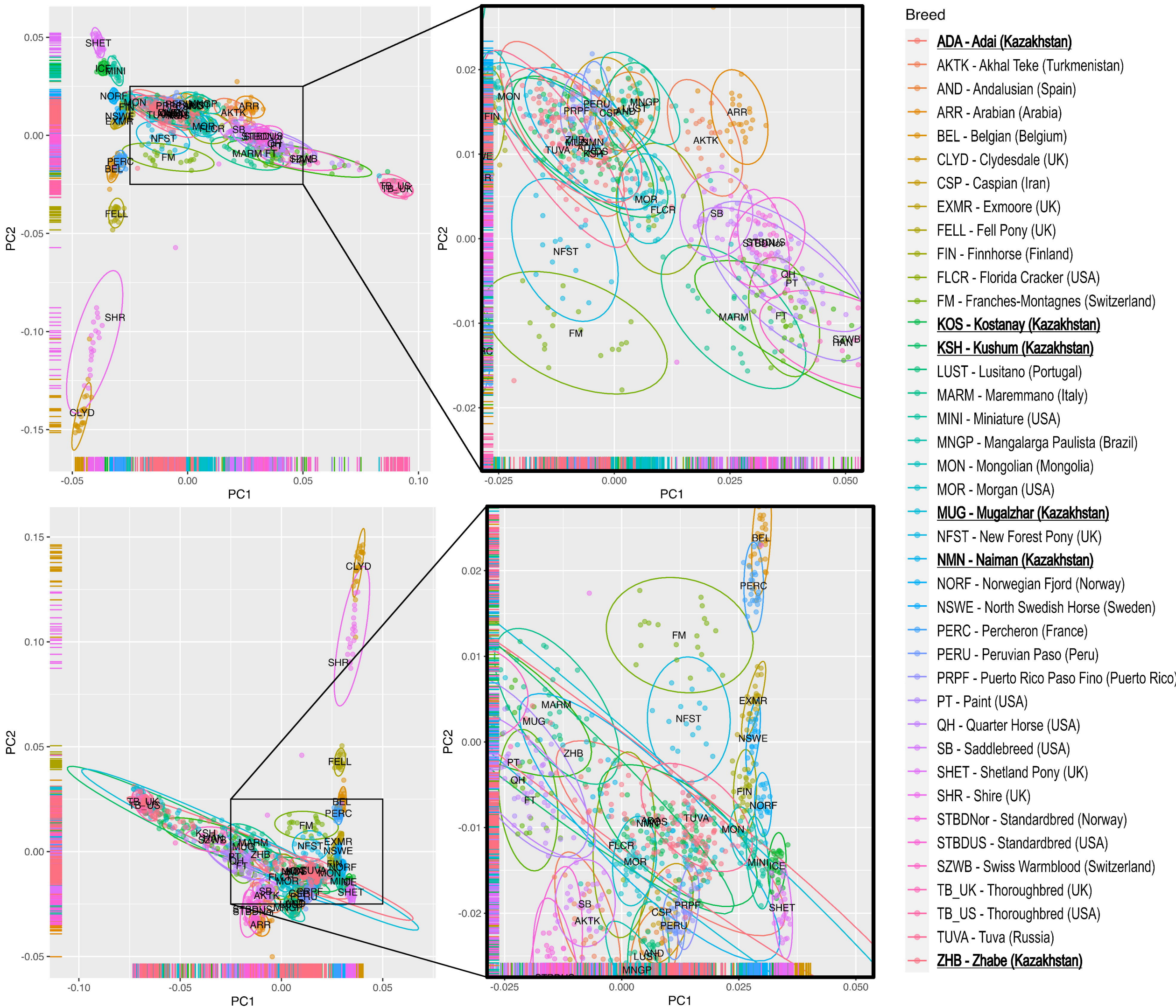
Figure 4. Neighbor joining tree based on Manhattan distances between Kazakh horses and the foreign breeds. A) full view of the tree, with the expanded view of the sub-trees containing outlying Kazakh horse genotypes; B) expanded view of the sub-tree consisting of Kazakh horse breeds.\

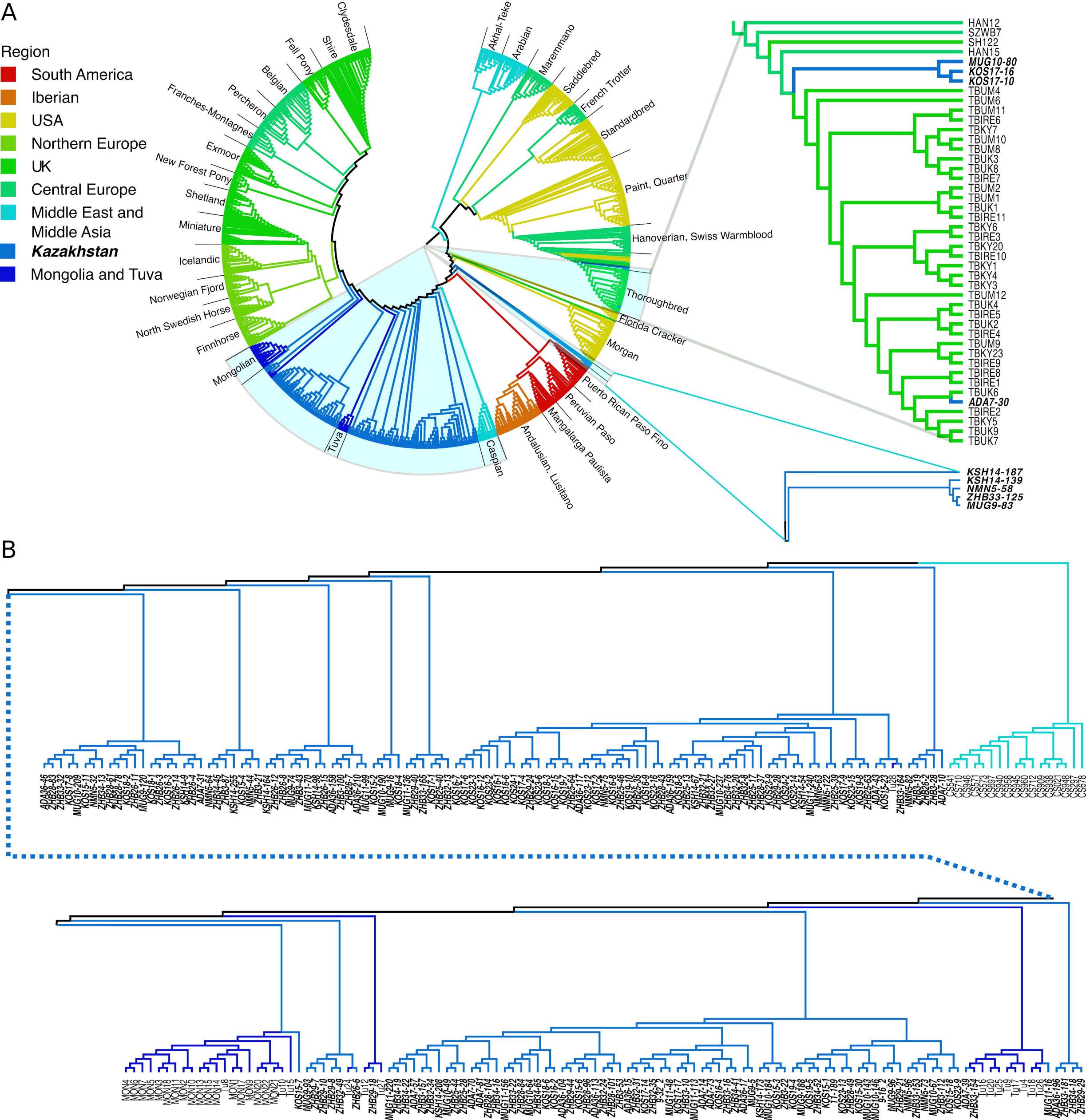
Figure 5. Manhattan plots of *p*-values corrected by the Monte-Carlo permutation test from genome-wide association analysis for horse size (top) and weight (bottom)

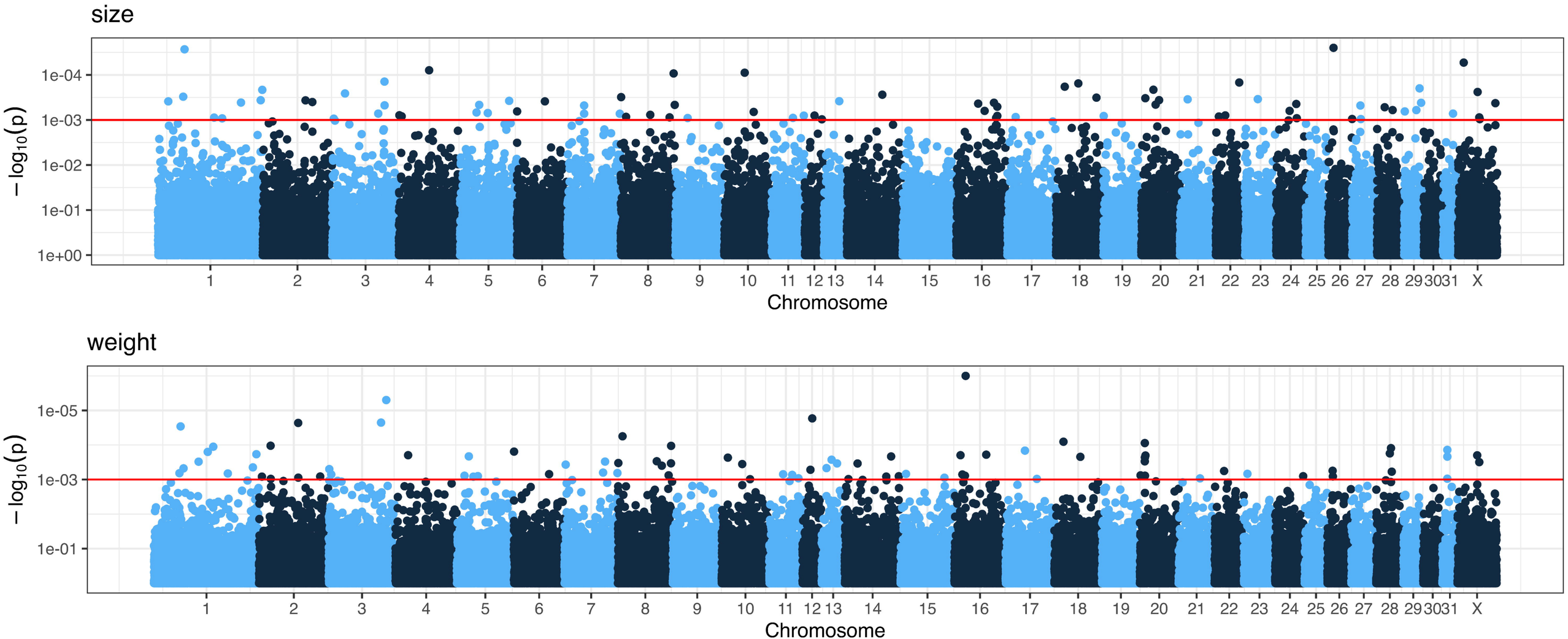
Supplementary table 1. Summary of horse sample collection.
Supplementary table 2. List of Kazakh horses identified as hybrids (outlying genotypes)
Supplementary table 3. Results of annotation of SNP associated with size and weight in the Kazakh horses.
Supplementary figure 1. Neighbor joining tree of the Kazakh horses in comparison with foreign breeds.


Legend
Adai (ADA)
Kostanay (COS)
Kushum (KSH)
Mugalzhar (MUG)
Naiman (NMN)
Zhabe (ZHB)
Ufa Уфа
Chelyabinsk Челябинск
Omsk Омск
Novosibirsk Новосибирск
Novokuzne
Barnaul Барнаул
Samara Самара
Orenburg Оренбург
Pavlodar Павлодар
Astana Астана
Karaganda Қарағанды
Almaty Алматы
Bishkek Бишкек
Kazakhstan
Uzbekistan
Kyrgyzstan
1000 km
N
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
19
20
21
22
23
24
25


A)
K = 2
K = 5
K = 10
KSH/Alem
ZHB/Akzhar
ZHB/AgroDamu
ADA/Kuanysh
MUG/Bakytbek
MUG/Orazbay
NMN/TokZhailau
MUG/Karatomar
ZHB/Anar
ZHB/Kalka
ZHB/Zhaksylyk
KOS/KazakTulpary
KOS/Chy
KOS/Stepnoye
ADA/KozhyrAta
ZHB/Sayasat
1 2 3 4 5 6 7 8
1 - KOS/KostanayTulpary
2 - ZHB/Ernur
3 - KOS/Tasmukhambetov
4 - ZHB/KyzykSok
5 - KOS/Akhmetov
6 - KOS/Kabazhayev
7 - KOS/Baranovskiy
8 - KOS/Zhussupov
B)
K = 2
K = 10
K=28
ADA
KOS
KSH
MUG
NMN
ZHB
MON
TUVA
NFST
CSP
AKTK
AND
ARR
BEL
CLYD
EXMR
FELL
FIN
FLCR
FM
FT
HAN
ICE
LUST
MARM
MINI
MNGP
NORF
NSWE
PERC
PERU
PRPF
PT
QH
SB
SHET
SHR
STBDNor
STBDUS
SZWB
TB_UK
TB_US


PC1
PC2
Breed
ADA - Adai (Kazakhstan)
AKTK - Akhal Teke (Turkmenistan)
AND - Andalusian (Spain)
ARR - Arabian (Arabia)
BEL - Belgian (Belgium)
CLYD - Clydesdale (UK)
CSP - Caspian (Iran)
EXMR - Exmoore (UK)
FELL - Fell Pony (UK)
FIN - Finnhorse (Finland)
FLCR - Florida Cracker (USA)
FM - Franches-Montagnes (Switzerland)
KOS - Kostanay (Kazakhstan)
KSH - Kushum (Kazakhstan)
LUST - Lusitano (Portugal)
MARM - Maremmano (Italy)
MINI - Miniature (USA)
MNGP - Mangalarga Paulista (Brazil)
MON - Mongolian (Mongolia)
MOR - Morgan (USA)
MUG - Mugalzhar (Kazakhstan)
NFST - New Forest Pony (UK)
NMN - Naiman (Kazakhstan)
NORF - Norwegian Fjord (Norway)
NSWE - North Swedish Horse (Sweden)
PERC - Percheron (France)
PERU - Peruvian Paso (Peru)
PRPF - Puerto Rico Paso Fino (Puerto Rico)
PT - Paint (USA)
QH - Quarter Horse (USA)
SB - Saddlebreed (USA)
SHET - Shetland Pony (UK)
SHR - Shire (UK)
STBDNor - Standardbred (Norway)
STBDUS - Standardbred (USA)
SZWB - Swiss Warmblood (Switzerland)
TB_UK - Thoroughbred (UK)
TB_US - Thoroughbred (USA)
TUVA - Tuva (Russia)
ZHB - Zhabe (Kazakhstan)


A
Region
South America
Iberian
USA
Northern Europe
UK
Central Europe
Middle East and Middle Asia
Kazakhstan
Mongolia and Tuva
Clydesdale
Shire
Fell Pony
Belgian
Percheron
Franches-Montagnes
Exmoor
New Forest Pony
Shetland
Miniature
Icelandic
Norwegian Fjord
North Swedish Horse
Finnhorse
Mongolian
Tuva
Caspian
Akhal-Teke
Arabian
Maremmano
Saddlebred
French Trotter
Standardbred
Paint, Quarter
Hanoverian, Swiss Warmblood
Thoroughbred
Florida Cracker
Morgan
Puerto Rican Paso Fino
Peruvian Paso
Mangalarga Paulista
Andalusian, Lusitano
HAN12
SZWB7
SH122
HAN15
MUG10-80
KOS17-16
KOS17-10
TBUM4
TBUM6
TBUM11
TBIRE6
TBKY7
TBUM10
TBUM8
TBUK3
TBUK8
TBIRE7
TBUM2
TBUM1
TBUK1
TBIRE11
TBKY6
TBIRE3
TBKY20
TBIRE10
TBKY1
TBKY4
TBKY3
TBUM12
TBUK4
TBIRE5
TBUK2
TBIRE4
TBUM9
TBKY23
TBIRE9
TBIRE8
TBIRE1
TBUK6
ADA7-30
TBIRE2
TBKY5
TBUK9
TBUK7
KSH14-187
KSH14-139
NMN5-58
ZHB33-125
MUG9-83
B


size
1e-04
1e-03
1e-02
1e-01
1e+00
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 X
Chromosome
weight
1e-05
1e-03
1e-01
$-\log_{10}(p)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 X
Chromosome