

A Comprehensive SNP Genotype and Phenotype Dataset of Native Kazakh Horses

Rakhim Kanat ^{1,2}, Alexandr Kovalchuk ³ , Indira Beishova ³, Vadim Ulyanov ³ , Malika Shamekova ^{1,2} and Maxat Toishimanov ^{1,*} 

¹ Laboratory of Breeding and Biotechnology, Institute of Plant Biology and Biotechnology, Timiryazev 45, Almaty 050040, Kazakhstan

² Tanir Research Laboratory, Al-Farabi Avenue 75B, Almaty 050060, Kazakhstan

³ Testing Center, Zhangir Khan West-Kazakhstan Agrarian Technical University, Zhangir Khan 51, Oral 090009, Kazakhstan

* Correspondence: m.toishimanov@ipbb.kz

Abstract

The present article provides genotype and phenotype data of 3279 Kazakh local horses from various regions of Kazakhstan. Genotypic data are 81,974 single nucleotide polymorphisms (SNPs) from genome-wide genotyping of all individuals sampled. All horses were phenotyped and their height at withers, oblique body length, chest circumference, cannon bone circumference and body mass were recorded. Field sampling of data took place at various study sites and biological samples and morphometric measurements were taken from each animal at the same time. The dataset comprises both raw and quality-controlled genotypic data, as well as phenotypic data in structured format. All animals were identified and cataloged, with metadata such as sampling location, sex and age if applicable. Following QC filtering, 2704 individuals and 73,377 SNPs were left. The dataset serves as a resource for population genetic studies, genome-wide association studies and genomic selection research, and for studies on genetic architecture of morphological traits. It can also be used for comparative genomics among breeds of horses, genetic diversity studies of central Asian horses and the creation and/or testing of genomic prediction models for body measurement traits. The dataset is freely available and formatted for easy integration with other equine genomic databases.

Dataset: <https://zenodo.org/records/23120773> (accessed on 3 October 2026) (DOI: 10.5281/zenodo.23120773)

Dataset License: Creative Commons Attribution 4.0 International

Keywords: equine genomics; SNP genotyping; phenotypic traits; horse breed; animal genetics; Kazakh horses



Academic Editor: Pufeng Du

Received: 11 September 2026

Revised: 7 October 2026

Accepted: 9 October 2026

Published: 14 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Summary

The horse (*Equus caballus*) has been domesticated on the steppes of the Eurasian continent ~5500 years ago [1,2] and has become an important model organism in scientific research, especially in the field of genomics. Beyond scientific value, the refined horse genome has impacted veterinary medicine and breeding programs. Kazakh local horses are indigenous horses of Kazakhstan and Central Asian countries which are used mainly for meat, milk and endurance in harsh climatic conditions. Although culturally and

economically significant, genomic resources are still limited for this population when compared to European and American horse breeds [3]. We created a large-scale genotype–phenotype resource, covering the genetic diversity and morphological variation of eight Kazakh horse breeds: Adai, Zhabe, Mugalzhar, Naiman, Kushum, Novoaltai, Kyrgyz and Kostanai, to address this gap. About 80,000 SNPs were measured in 3279 individuals and five morphometric traits. This data descriptor presents these data in a structured, open-access format, to facilitate population genetic characterization and genotype–phenotype association studies.

There are three benefits to these data. First, 3279 individuals from a population of about 80,000 SNPs were genotyped, which is one of the largest equine genotype–phenotype datasets for a population of horses from Central Asia, providing a strong basis for population genetic analysis and genome-wide association studies in this underrepresented breed population. Second, the availability of dense SNP data and five conformation measurements per animal provides the opportunity to study the genetic architecture of economically important conformation traits and also to create or test genomic prediction models. Third, the genetic diversity, population structure, and admixture of Central Asian horse breeds can be used in future comparative studies with other world breeds, in the determination of within-population variation, and in the identification of signatures of selection specific to Kazakh local horses. The data can also be used as a reference population for genomic selection breeding programs for body conformation and meat production traits in Kazakh horses.

2. Data Description

The dataset comprises genotypic and phenotypic information from 3279 Kazakh local horse individuals representing multiple breeds. The repository is hosted on Zenodo with four main directories: *raw/*, *raw_qc/*, *metadata/*, and *phenotype/*, containing PLINK (BED, BIM, FAM) raw and quality-controlled genotype data, sample metadata (location, sex, breed), and morphometric measurements in .tsv format, respectively.

2.1. Raw Genotype Data (*raw/*)

The *raw/* directory is the original genotype data scanned from the Illumina Equine80k SNP array on an Illumina iScan platform. Genotypes are available in PLINK binary format, which consists of three files: *dataset.bed* (binary genotype data), *dataset.bim* (SNP variant information including chromosome, SNP identifier, genetic distance, physical position, and alleles), and *dataset.fam* (individual information including family ID, individual ID, paternal and maternal IDs, sex, and phenotype). This directory includes genotype calls for 81,974 SNPs in 3279 individuals without any quality control filtering. The equine SNP distribution around the genome is depicted in Figure 1 (total SNP number per chromosome and number of SNPs per length of chromosome). Figure 2 shows the distribution of SNP positions in each chromosome.

2.2. Quality-Controlled Genotype Data (*raw_qc/*)

The genotype data in the *raw_qc/* directory are obtained after sequential quality control filtering using PLINK 1.9 [4]. The files are in the same PLINK binary format as the raw data (*dataset_qc.bed*, *dataset_qc.bim*, *dataset_qc.fam*). The following filtering criteria were applied sequentially: individual missing genotype rate threshold of 0.1 (*--mind 0.1*), SNP missing genotype rate threshold of 0.1 (*--geno 0.1*), minor allele frequency threshold of 0.01 (*--maf 0.01*).

These filters resulted in 2692 individuals and 78,690 SNPs remaining from the 3207 individuals and 81,974 SNPs in the raw dataset, which represents 83.9% of individuals and

96.0% of SNPs. Table 1 shows the number of individuals and SNPs that were retained after each sequential filter.

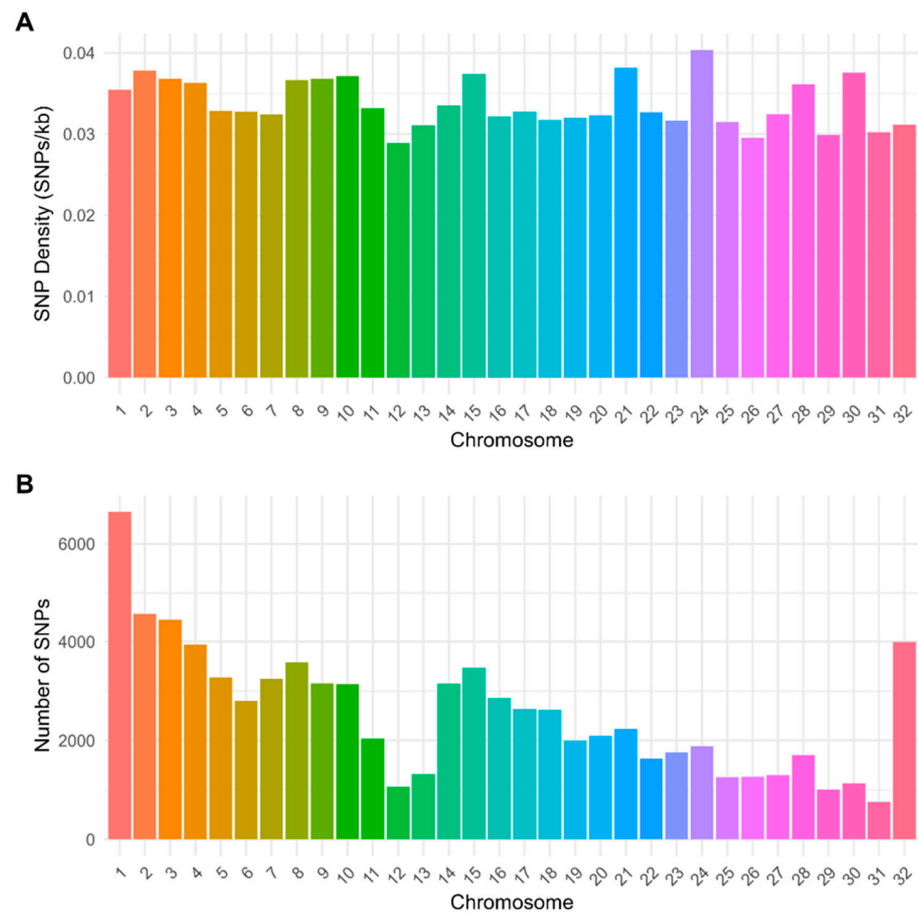


Figure 1. SNP distribution by chromosome: (A) SNP density by horse chromosome (SNPs per kb), accounting for varying chromosome lengths; (B) SNP count by horse chromosome.

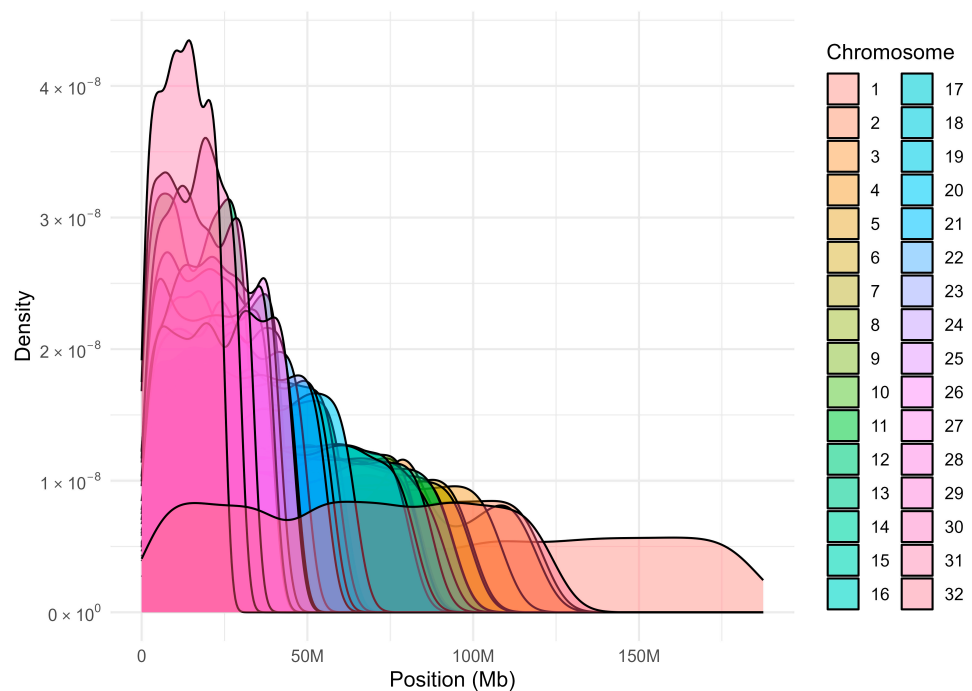


Figure 2. SNP position distribution across all chromosomes.

Table 1. Results of dataset filtering.

Filtering Step	Criterion	Individuals Retained	% Individuals Kept	SNPs Retained	% SNPs Kept
Initial raw data	—	3207	100.0	81,974	100.0
Missing genotype rate (individuals)	--mind 0.1	2692	83.9	81,974	100.0
Missing genotype rate (SNPs)	--geno 0.1	2692	83.9	81,113	99.0
Minor allele frequency	--maf 0.01	2692	83.9	78,690	96.0

Figure 3 presents a principal component analysis (PCA) plot of the entire quality-controlled dataset, generated using a distance matrix. The first three principal components capture 22.75%, 12.18%, and 7.19% of the total genetic variance, respectively. This visualization allows assessment of population structure and potential clustering patterns among the 2692 individuals included in the dataset.

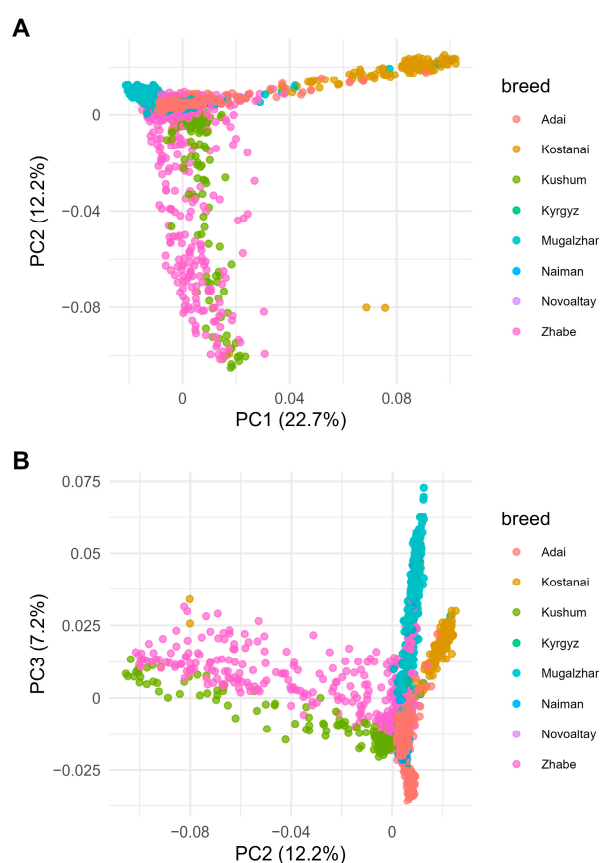
**Figure 3.** Principal component analysis of the whole dataset using a distance matrix. (A) Principal components 1 and 2. (B) Principal components 2 and 3.

Figure 4 displays observed heterozygosity values calculated for each breed represented in the dataset. Heterozygosity was computed as the proportion of heterozygous genotypes

per individual averaged across all 78,690 SNPs. Values are grouped by eight distinct horse breeds, facilitating comparison of genetic diversity levels across populations.

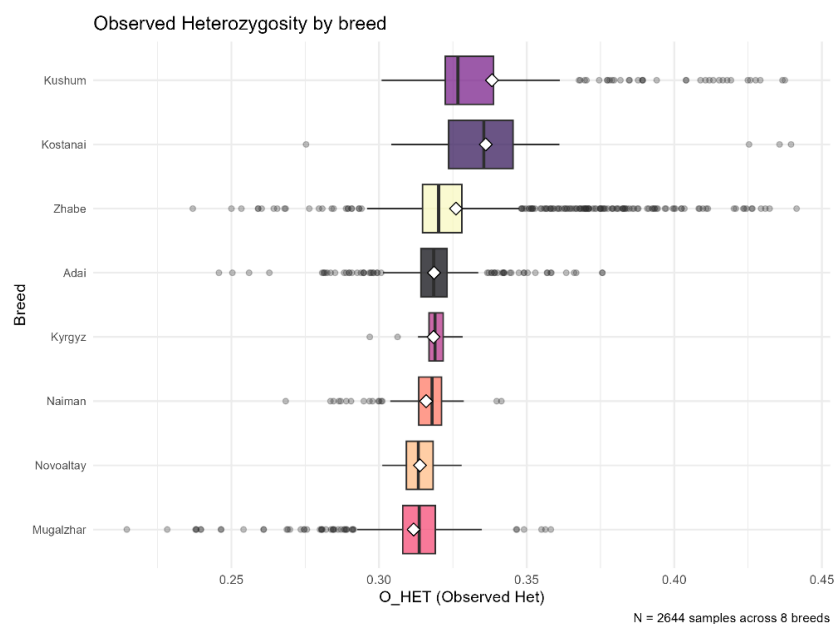


Figure 4. Observed heterozygosity by breed. Boxplots show the distribution of observed heterozygosity (O_HET) across individuals within each breed. The box spans the interquartile range (IQR; Q1–Q3), the horizontal line indicates the median, and the white diamond marks the mean. Outliers (individuals beyond $1.5 \times \text{IQR}$ from the quartiles) are plotted as semi-transparent dots.

Figure 5 illustrates the total length of runs of homozygosity (ROH) calculated per breed. ROH segments were identified using PLINK with a minimum length threshold of 500 kb, a maximum gap between consecutive homozygous SNPs of 100 kb, a minimum number of SNPs per segment of 30, and maximum heterozygous calls allowed per segment of 1. Total ROH length per individual was summed and averaged within each breed group, providing a summary measure of recent inbreeding or population history captured by the dataset.

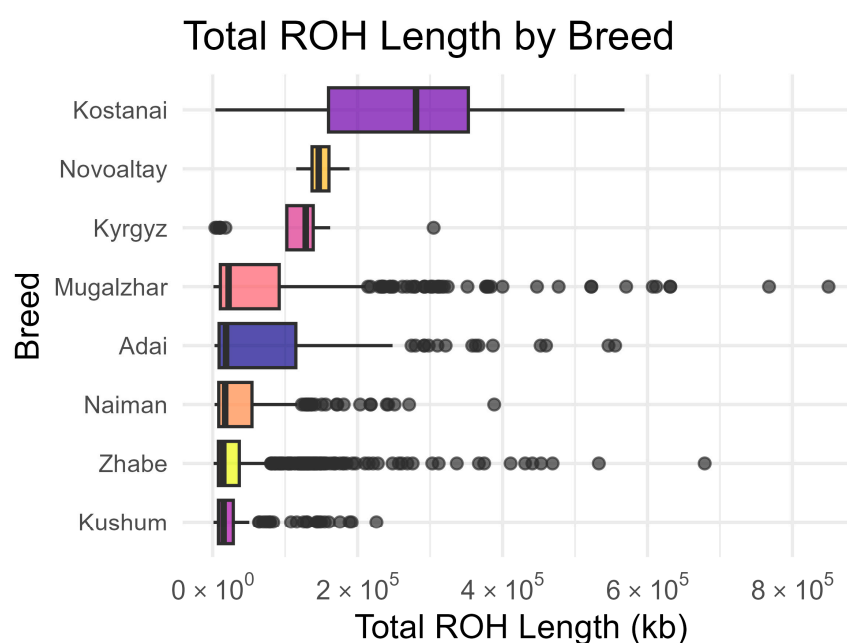


Figure 5. Total ROH length by breed.

Identity-by-descent analysis using PLINK v1.90 identified 28,125 pairs of closely related individuals ($\pi\text{-hat} > 0.05$) among the quality-controlled samples. Of these, 152,182 pairs exhibited kinship coefficients consistent with first cousin relationships or closer, while 10,278 pairs showed half-sibling/avuncular-level relatedness, and 1604 pairs indicated parent–offspring or full sibling relationships. Also, 43 duplicate- or twin-level matches were found ($\pi\text{-hat} \geq 0.9$). The high percentage of apparent familial relationships in this dataset may represent true close relationships within the populations sampled or background genetic similarity typical of indigenous horse breeds with similar histories and traditional breeding systems (Figure 6).

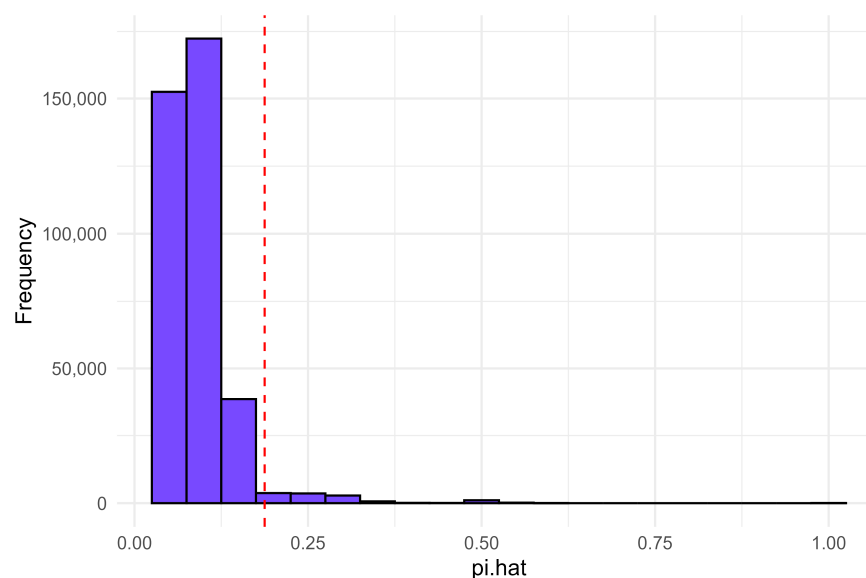


Figure 6. IBD kinship coefficient distribution. Red dotted line represents $\pi\text{-hat} > 0.1875$ threshold.

2.3. Metadata (metadata/)

The metadata/ directory contains a single tab-separated values file, metadata.tsv, providing sample-level information for all 2981 individuals in the dataset. The file includes columns for individual identifiers (matching those in the dataset.fam file), breed classification, sex, age at the time of measurement, measurement date, farm identifier, and geographic region of origin. This metadata enables cross-referencing between genetic data and phenotypic or environmental variables for downstream analyses.

The percentages of missing data for each metadata field are shown in Table 2. The largest number of missing data are found in age (753 samples, 24.12%) and gender (429 samples, 13.74%). Only 194 samples (6.21%) do not have location data and only 36 samples (1.15%) do not have breed information. The field with the lowest missing rate is farm identifiers with 10 samples (0.32%), showing a high completeness rate for this field. These patterns of missing data are pertinent to the choice of which samples to use in particular analyses.

Table 2. Missing data summary for metadata fields.

Field	Missing	Missing (%)
Age	673	22.58
Gender	340	11.41
Location	194	6.51
Farm	10	0.34
Breed	36	1.21

Figure 7 presents the distribution of metadata categories across all samples in four panels: (A) regional distribution of individuals, illustrating geographic sampling coverage; (B) distribution of samples across farm identifiers; (C) breed composition; and (D) gender distribution among all sampled individuals. Together, these visualizations provide an overview of the demographic and geographic structure of the cohort.

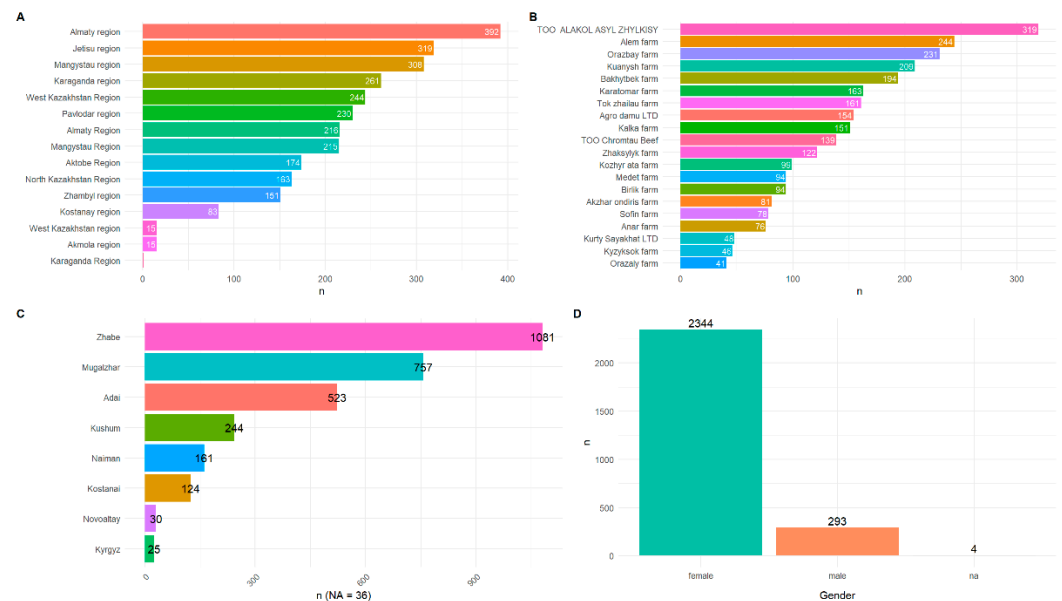


Figure 7. Metadata distributions across all samples: (A) distribution by regions; (B) distribution by farms; (C) distribution by breed name; (D) distribution by gender.

2.4. Phenotypic Data (phenotype/)

The phenotype/folder includes morphometric data in a tab-separated values file, phenotype.tsv, that has the same individual identifiers as the genotype and metadata files. For each animal, the following morphometric measurements were taken: height at withers (cm), oblique body length (cm), chest circumference (cm), circumference of cannon bone (cm) and body mass (kg). The measurements performed are standard morphometric parameters routinely used in the phenotype characterization of horses and were taken with standardized measurement procedures at all sampling sites.

Missing data rates for phenotypic traits are summarized in Table 3, with 2442 samples that have all phenotypes. Body mass has the highest proportion of missing records (481 samples, 16.4%), followed by cannon bone circumference (462 samples, 15.8%), chest circumference (455 samples, 15.5%), height at withers (455 samples, 15.5%), and oblique body length (455 samples, 15.5%). Researchers should account for these missing data patterns when planning analyses involving multiple traits.

Table 3. Missing data summary for phenotypic traits.

Phenotype	Missing	Missing (%)
Body mass	481	16.4
Cannon bone circumference	462	15.8
Chest circumference	455	15.5
Height at withers	455	15.5
Oblique body length	455	15.5

Figure 8 presents the distribution of and relationships among phenotypic traits in four panels: (A) distribution of all five phenotypic traits; (B) distribution of body mass

specifically; (C) a scatter plot of height at withers against body mass; and (D) a scatter plot of chest circumference against body mass.

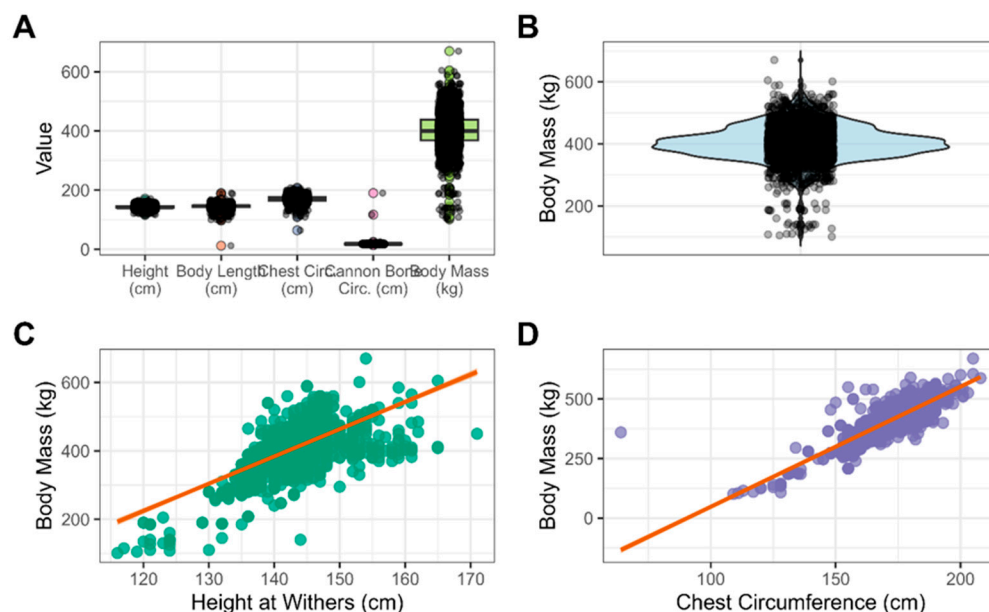


Figure 8. (A) Distribution of all phenotypic traits; (B) body mass distribution; (C) height at withers vs. body mass; (D) chest circumference vs. body mass. The orange line in (C,D) represents best fit line using ordinary least squares regression method.

3. Methods

3.1. Study Population and Sampling

In total, 3279 local horses from several regions of Kazakhstan were sampled. Sampling sites were chosen to cover the geographic distribution and management systems of the Kazakh local horses. At sampling, phenotypic data (height at withers, oblique body length, chest circumference, cannon bone circumference and body mass) were recorded from each animal, and hair was taken from the tail. All animals were identified and data on sex, age, and sampling site were collected. Trait definitions: height at withers: vertical distance from the ground to the highest point of the withers; oblique body length: distance from the shoulder joint to the ischial tuberosity; chest circumference: circumference immediately behind the shoulder blades; cannon bone circumference: circumference at the narrowest point of the cannon bone; body mass: weight measured using calibrated livestock scales.

3.2. DNA Extraction

Genomic DNA was extracted from hair follicle samples using the “DNK-Extran2” reagent (Syntol, Moscow, Russia) following the manufacturer’s protocol. DNA quality and quantity were assessed using a Qubit 4 Fluorometer with Qubit dsDNA Broad Range reagent (Thermo Fisher Scientific, Waltham, MA, USA). Samples meeting sufficient quality and concentration thresholds were selected for genotyping.

3.3. Genotyping

Samples were genotyped using the Equine80k SNP array (custom array) [5] on the iScan system (Illumina, San Diego, CA, USA), targeting 81,974 single nucleotide polymorphisms distributed across the equine genome. Genotype calling was performed using GenomeStudio (v2.0.5.0) with default clustering settings and then exported to a PLINK-compatible format.

3.4. Quality Control

Raw genotype data were subjected to quality control filtering using PLINK v1.90 [4]. The following filters were applied: sample call rate threshold >90% (--mind 0.1), SNP call rate threshold >90% (--geno 0.1), and minor allele frequency threshold >0.01 (--maf 0.01). MAF filters were applied only to downstream analyses performed in this manuscript. The filtered dataset deposited in the repository only passed through sample and SNP call rate filters as variants that are rare in the present dataset may be polymorphic in other populations. A log file documenting the number of SNPs and individuals retained after each filtering step is included in the dataset repository.

3.5. Visualization Pipeline

The quality-controlled genotype data was analyzed using principal component analysis (PCA) with PLINK v1.90 --pca command. A genetic relationship/distance matrix was calculated using the genotype data and eigenvectors/eigenvalues were extracted. The first three principal components were plotted in R (v4.6.1) using ggplot2 and the variance explained was computed based on the eigenvalues.

Runs of homozygosity (ROH) were detected using PLINK v1.90 with the following parameters: minimum ROH segment length of 500 kb (--homozyg-kb 500), maximum gap between consecutive homozygous SNPs of 100 kb (--homozyg-gap 100), minimum of 30 SNPs per segment (--homozyg-snp 30), and a maximum of one heterozygous genotype allowed per ROH segment (--homozyg-window-het 1). Analyses were performed on the quality-controlled genotype data. The total length of ROH segments was summed for each individual and then averaged within each breed group, providing a summary measure of genome-wide autozygosity, which reflects recent inbreeding and demographic history. The results were visualized in R using ggplot2.

IBD analysis was performed using PLINK v1.90 (--genome command) on the quality-controlled dataset. Summarization and visualization were performed using tidyverse package.

All downstream statistical analyses and figure generation were performed in R [6] using tidyverse [7] package.

4. User Notes

The sampling of horses was restricted to those available in certain study areas in Kazakhstan, and this may not represent the geographic and genetic diversity of all sub-populations of the local horse population in Kazakhstan. Phenotypic measurements were carried out by several field operators at various sites, but there is a possibility that there may be some inter-operator variation in morphometric measurements. The SNP panel used (~80,000 SNPs) might not be dense enough for fine-mapping of causal variants or the detection of rare alleles since the array was not designed for variant-level resolution. Tables 2 and 3 list the data patterns that were not obtained and should be taken into account when choosing sample subsets for particular analyses.

Author Contributions: Conceptualization, R.K. and M.S.; methodology, R.K., A.K. and M.T.; software, R.K.; Validation, R.K. and M.T.; formal analysis, R.K. and V.U.; investigation, I.B., A.K. and V.U.; resources, I.B. and A.K.; data curation, R.K. and M.S.; writing—original draft, R.K.; writing—review and editing, M.S., M.T. and A.K.; supervision, M.S. and A.K.; Visualization, R.K. and V.U.; project administration, M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number AP22782840 and the Ministry of Agriculture of the Republic of Kazakhstan, program number BR22887106. The APC was funded by Institute of Plant Biology and Biotechnology.

Institutional Review Board Statement: The animal study protocol was approved by the Institutional Ethics Committee of NJSC “Zhangir Khan West-Kazakhstan Agrarian Technical University” (protocol code 21_2024).

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset presented in this study is openly available in Zenodo at <https://zenodo.org/records/23120773> (accessed on 3 October 2026) (DOI: 10.5281/zenodo.23120773).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. 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.; et al. Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PLoS ONE* **2013**, *8*, e54997. [[CrossRef](#)] [[PubMed](#)]
2. Wade, C.M.; Giulotto, E.; Sigurdsson, S.; Zoli, M.; Gnerre, S.; Imsland, F.; Lear, T.L.; Adelson, D.L.; Bailey, E.; Bellone, R.R.; et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **2009**, *326*, 865–867. [[CrossRef](#)] [[PubMed](#)]
3. Pozharskiy, A.; Abdrakhmanova, A.; Beishova, I.; Shamshidin, A.; Nametov, A.; Ulyanova, T.; Bekova, G.; Kikebayev, N.; Kovalchuk, A.; Ulyanov, V.; et al. Genetic structure and genome-wide association study of the traditional Kazakh horses. *Animal* **2023**, *17*, 100926. [[CrossRef](#)] [[PubMed](#)]
4. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **2015**, *4*, s13742-015-0047-8. [[CrossRef](#)] [[PubMed](#)]
5. Pozharskiy, A.; Beishova, I.; Nametov, A.; Shamshidin, A.; Ulyanova, T.; Kovalchuk, A.; Ulyanov, V.; Shamekova, M.; Bekova, G.; Gritsenko, D. Genetic composition of Kazakh horses of Zhabe type evaluated by SNP genotyping. *Heliyon* **2025**, *11*, e41173. [[CrossRef](#)] [[PubMed](#)]
6. Rizzo, M.L. *Statistical Computing with R*, 2nd ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2019. [[CrossRef](#)]
7. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D.A.; François, R.; Grolemond, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the tidyverse. *J. Open Source Softw.* **2019**, *4*, 1686. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.