animals

MDPI

Article

Genetic Diversity and Population Structure of the Kustanay Horse Breed Based on 16 STR Markers

Bakhit Baimenov [1], Alma Dossova [1], Stepan Khlyupko [2], Vitaly Raketsky [2], Aigul Tastemirova [3], Vadim Ulyanov [4,*] and Tatyana Ulyanova [4,*]

[1] Laboratory of Molecular-Genetic Research, Research Institute of Innovative Technologies, Akhmet Baitursynuly Kostanay Regional University, Kostanay 110000, Kazakhstan; bahytbajmenov@gmail.com (B.B.); dossova.alma95555@gmail.com (A.D.)
[2] Kostanay Branch, Kazakh Scientific Research Veterinary Institute LLP, 94 Dulatov Street, Kostanay 110000, Kazakhstan; hlupkostepan@gmail.com (S.K.); vitali29071982@gmail.com (V.R.)
[3] Republican Chamber for Pedigree Horse Breeds, 6/137 Korgalzhyn Highway, Astana 010000, Kazakhstan; amandykova3393@gmail.com
[4] Laboratory of Biotechnology and Diagnostics of Infectious Diseases, Testing Center, Zhangir Khan West Kazakhstan Agrarian and Technical University, 51 Zhangir Khan Street, Uralsk 090009, Kazakhstan
* Correspondence: vadimkst19@gmail.com (V.U.); tatyana.poddudinskaya@gmail.com (T.U.)

Simple Summary

The Kustanay horse is a breed developed in Kazakhstan by crossing local Kazakh mares with Don, Thoroughbred, Streletskaya, and Astrakhan (Kalmyk) stallions. However, many horses considered by their owners to belong to the Kustanay breed have no official pedigree records, making their true relationship to the breed uncertain. We studied more than 700 horses from different regions of Kazakhstan, including animals with official pedigree records and horses without documented ancestry. Their genetic material was examined using 16 internationally recognized markers commonly used to compare horses and verify parentage. Group sizes were highly unbalanced: 595 horses lacked pedigree documentation, while the other groups had only 9–36 animals. The studied groups showed a high level of genetic variation and only small differences between them. Horses without pedigree documentation were especially similar to officially registered Kustanay horses. However, the absence of most founder breeds from the study does not allow this similarity to be considered evidence of Kustanay origin. Establishing the breed affiliation of individual animals requires combining genetic data with reliable pedigree information. The findings provide useful genetic information for improving pedigree records, supporting future identification studies, and helping preserve the Kustanay horse as an important genetic resource of Kazakhstan.

Check for updates

Academic Editor: Marta Siemieniuch-Tartanus

Received: 19 August 2026
Revised: 21 October 2026
Accepted: 29 October 2026
Published: 1 November 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) license.

Abstract

The Kustanay horse is a riding breed developed in Kazakhstan whose genetic diversity and differentiation remain insufficiently studied. We assessed genetic diversity and population structure using 16 autosomal STR loci from a 17-locus panel comprising 12 ISAG core markers. Of 765 genotyped horses, 703 remained after excluding duplicate records and close relatives; the X-linked LEX3 locus was excluded from population analyses. Five groups were studied: registered Kustanay (KST, n = 32), Kustanay from Qazaq Tulpary (KT, n = 36), Thoroughbred (TB, n = 31), Kustanay crossbred (CROSS, n = 9), and undocumented origin (UB, n = 595). Expected heterozygosity was high across groups (H_E = 0.661–0.725). Overall F_{ST} was 0.0088, and AMOVA attributed 99.13% of molecular variance to variation within individuals. The Puechmaille estimators favored $K = 2$, whereas Evanno ΔK peaked at $K = 4$; neither solution revealed clear separation among groups. Nei's genetic distances and

PCoA were consistent with weak differentiation and similarity between registered Kustanay and undocumented horses. Unequal group sizes limit interpretation, and these findings do not establish breed identity for individual undocumented horses. The results inform pedigree recording and conservation planning.

Keywords: Kustanay horse; microsatellite markers; genetic diversity; population structure; Kazakhstan

1. Introduction

Horse breeding is a strategically important sector of livestock production in the Republic of Kazakhstan, contributing to the production of horse-derived products, the advancement of pedigree and sport horse breeding, and the conservation of the nation's equine genetic resources [1]. The Kustanay horse breed occupies a unique position among Kazakhstan's domestic horse breeds as the first specialized riding breed developed in the country through systematic selective breeding and officially recognized in 1951 [2,3]. The Kustanay horse breed combines strong working ability, favorable athletic performance, and adaptation to the harsh continental climate of northern Kazakhstan, making it a genetic resource of considerable scientific and practical importance for the country's horse breeding industry.

The breed was developed during the first half of the twentieth century through reproductive crossbreeding of local Kazakh steppe mares with Don, Thoroughbred, Streletskaya, and Astrakhan (Kalmyk) stallions, followed by long-term selective breeding for conformation, performance, and adaptive traits [2,4]. The Kustanay horse is a valuable indigenous genetic resource of Kazakhstan that is currently facing a serious risk of extinction. According to official statements made in November 2024, the remaining population in Kazakhstan was estimated at approximately 250–280 horses [5,6]. This small reported population size underscores the urgent need to assess genetic diversity, population structure and the potential risks associated with the loss of genetic variation.

In recent decades, the conservation of genetic diversity in local horse breeds has become increasingly important due to declining breeding populations, the widespread use of industrial crossbreeding, and the associated risk of losing unique genetic variants. In the Kustanay breed, this challenge is further complicated by the large number of animals identified as Kustanay despite lacking documented pedigree records, which hinders effective breeding management and threatens the preservation of the breed's genetic integrity. Microsatellite genotyping has also been used to characterize the genetic diversity and population structure of other local horse breeds with documented pedigrees, such as the Lipizzan [7,8].

In accordance with the recommendations of the Food and Agriculture Organization of the United Nations (FAO) and the International Society for Animal Genetics (ISAG), monitoring of livestock genetic resources is recommended to be carried out using molecular genetic markers [9,10]. Among these, short tandem repeats (STRs) have gained the widest application; owing to their high polymorphism, codominant inheritance, and international standardization, they are used for animal identification, parentage verification, assessment of genetic diversity, and analysis of breed population structure [9,11].

Despite the widespread adoption of high-throughput SNP genotyping arrays [12], the STR marker panel used in this study—comprising 17 loci, including the 12 core markers recommended by ISAG—remains the international reference for horse identification and parentage verification [10]. In addition to individual identification, STR genotypes can be used to assess genetic diversity and population structure.

In recent years, the study of genetic resources in domestic horse breeds has advanced considerably owing to the widespread application of molecular genetic approaches. Initial

investigations of the Kustanay horse breed using microsatellite markers focused on the analysis of genetic polymorphism within individual stud lines and confirmed the suitability and high discriminatory power of the standard 17-STR marker panel for genetic characterization of the breed [13]. Subsequently, a preliminary assessment of the genetic diversity of the modern Kustanay breed population was carried out on a limited sample of animals [4]. Similar studies using STR markers have been conducted for other domestic breeds, including the Kushum and Mugalzhar breeds, confirming the high effectiveness of this approach for assessing genetic diversity and population structure [14,15].

The development of genomic technologies has substantially expanded opportunities for investigating the genetic diversity and population structure of domestic horse breeds. High-density SNP genotyping arrays have been successfully applied to characterize the population structure and genetic diversity of traditional Kazakh horse populations [2,16]. In parallel, studies have been conducted on mitochondrial DNA and Y-chromosome haplotypes of the Kushum breed [17], as well as maternal diversity in domestic horse breeds [3]. Furthermore, the first high-quality de novo genome assembly of the Kazakh Zhabe horse has been reported, providing an important genomic resource for future investigations of structural variation and the functional genomics of domestic horse breeds [18].

Despite these advances, the current state of the Kustanay horse breed's gene pool remains insufficiently characterized. Previous studies have focused primarily on individual stud lines, limited sample sizes, or other domestic horse breeds. To date, no comprehensive assessment of the contemporary Kustanay horse population has been performed using the internationally standardized ISAG panel of 17 STR markers. Furthermore, the applicability of this marker panel has not been evaluated for the molecular characterization of both registered Kustanay horses and animals classified as Kustanay despite lacking documented pedigree records. Addressing this knowledge gap is essential for improving breeding management, preserving the breed's genetic integrity, strengthening animal identification and parentage verification systems, and providing a scientific basis for long-term conservation programs targeting Kazakhstan's national equine genetic resources.

The aim of the present study was to comprehensively assess the genetic diversity, population structure, and degree of genetic differentiation of the Kustanay horse breed based on the analysis of 16 autosomal STR loci of a 17-locus panel that includes the 12 ISAG core markers. In addition, the study aimed to evaluate the genetic structure of horse groups differing in origin and documented breed status.

2. Materials and Methods

2.1. Sampling

The study included 765 individuals of *Equus caballus*, distributed across five groups (Table 1).

Sampling was carried out in 2021–2025 from adult animals of both sexes kept on stud farms, commercial farms, and private smallholdings in various regions of the Republic of Kazakhstan (Figure 1, Table S1).

Hair follicles were used as the source of genomic DNA. From each animal, 30–40 hairs with intact follicles were collected. Biological material was collected by breeding inspectors of the Republican Chamber for Pedigree Horse Breeds. The Chamber provided written consent for laboratory analysis of the collected material and for the use of the research results in scientific publications. Each sample was assigned a unique identification number and stored in an individually labeled paper envelope.

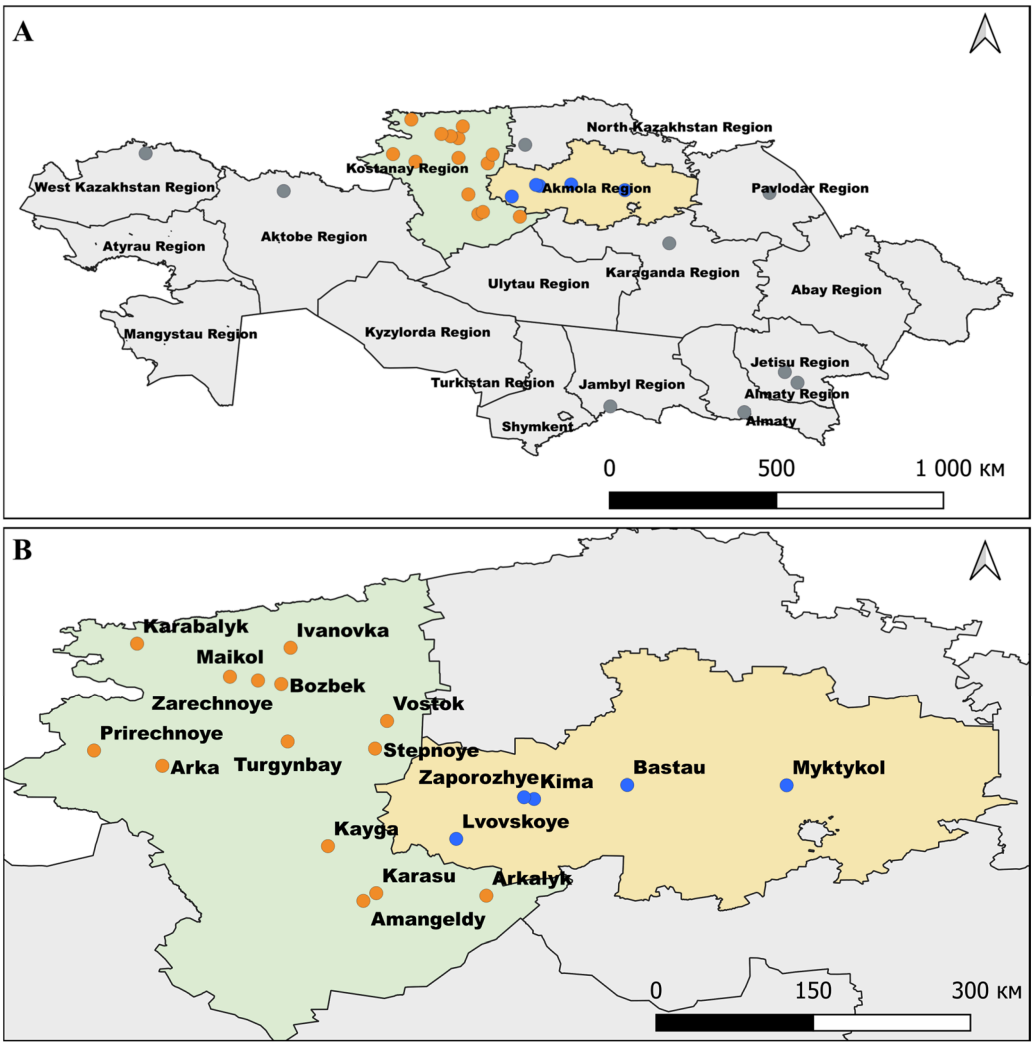

Figure 1. Geographic distribution of horse sampling locations in Kazakhstan. (**A**) Country-level map of Kazakhstan showing all sampling locations; Kostanay and Akmola Regions are highlighted in green and yellow, respectively. Orange and blue points indicate sampling locations in Kostanay and Akmola Regions, respectively, whereas gray points indicate sampling locations in other regions of Kazakhstan. (**B**) Detailed view of the sampling locations in Kostanay and Akmola Regions, with individual sampling localities labeled.

Table 1. Description of the studied horse groups (n, number of animals; before quality control).

Group Code	n	Males	Females	Sex Not Specified	Group/Breed
CROSS	9	3	6	0	Kustanay crossbred horses
KST	32	7	23	2	Registered Kustanay horses
KT	39	15	24	0	Kustanay horses from “Qazaq Tulpary” LLP
TB	31	16	15	0	Registered Thoroughbred horses
UB	654	236	374	44	Horses classified as Kustanay by owners but lacking pedigree documentation
Total	765	277	442	46	

Animals were assigned to groups based on breeding certificates, pedigree records, and information provided by their owners. The UB group comprised horses identified by their owners as Kustanay horses but lacking official pedigree documentation confirming their origin. The Thoroughbred (TB) breed was included because of its historical contribution to the formation of the Kustanay breed and the availability of samples from registered horses of this breed. Other breeds that contributed to the formation of the Kustanay breed were not included in the study owing to the lack of available samples. Moreover, the study was not designed to assess the relative genetic contribution of individual founder breeds to the modern Kustanay horse population. KST comprised registered Kustanay horses from private owners, whereas KT consisted exclusively of horses from a single stud farm, "Qazaq Tulpary" LLP, which specializes in breeding the Kustanay breed. The sample size of each group was determined by the availability of animals meeting the respective inclusion criteria at the time of sampling.

The study was approved by the Local Ethics Committee of the Non-profit Limited Company "Akhmet Baitursynuly Kostanay Regional University" (Protocol No. 9, 30 September 2024). Hair follicle samples collected prior to the approval date had originally been obtained by authorized breeding inspectors of the Republican Chamber for Pedigree Horse Breeds for routine pedigree registration and genetic identification purposes, independent of the present study, and were subsequently used as archived material under the approved protocol. All sampling was performed using a non-invasive procedure and in accordance with national and institutional guidelines for the humane treatment of animals.

2.2. DNA Extraction

DNA extraction and genotyping were performed at the Laboratory of Molecular Genetic Research, Non-profit Limited Company "Akhmet Baitursynuly Kostanay Regional University". Genomic DNA was extracted using the DNA-Extran-2 kit (SYNTOL LLC, Moscow, Russia) according to the manufacturer's instructions. DNA concentration and purity were determined using a NanoDrop™ One spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). DNA samples were stored at −20 °C until further molecular genetic analyses.

2.3. STR Genotyping

Genotyping was performed using the commercial "COrDIS Horse" kit (GORDIZ, Moscow, Russia), comprising a 17-locus STR panel that includes 12 core markers recommended by ISAG for genetic identification and five additional loci [10]. The panel comprised 12 core (AHT4, AHT5, ASB2, ASB17, ASB23, HMS2, HMS3, HMS6, HMS7, HTG4, HTG10, VHL20) and 5 additional (CA425, HMS1, HTG6, HTG7, LEX3) STR loci. The chromosomal location, structure, and repeat types of the 17 STR loci used in the study are presented in Table 2.

DNA amplification was performed by multiplex PCR according to the manufacturer's instructions using a ProFlex™ PCR System thermal cycler (Applied Biosystems, Waltham, MA, USA). Each PCR run included a positive control.

PCR products were separated by capillary electrophoresis using an ABI 3500 Genetic Analyzer (Applied Biosystems, Waltham, MA, USA). Electropherograms were analyzed using GeneMapper™ Software version 6.0 (Applied Biosystems, Waltham, MA, USA).

Alleles were designated according to the international nomenclature proposed by van de Goor et al. (2010) [19]. Samples that failed to amplify or produced poor-quality electropherograms were re-amplified and re-genotyped; only profiles that passed quality control were included in the analysis. During genotype scoring, samples were identified

by unique codes, and the personnel responsible for electropherogram interpretation were blinded to the animals' study-group assignments.

Table 2. Description of the microsatellite markers.

Locus	Chromosomal Location	Repeat Type	Repeat Unit Structure
AHT4	24q14	compound	$(AC)_nAT(AC)_n$
AHT5	8	simple	$(GT)_n$
ASB2	15q21.3–q23	simple	$(GT)_n$
ASB17	2p14–p15	simple	$(AC)_n$
ASB23	3q22.1–q22.3	simple and compound	$(TG)_n$ and $(TG)_nTT(TG)_4$
CA425	28q18	simple	$(GT)_n$
HMS1	15	simple	$(TG)_n$
HMS2	10	compound	$(CA)_n(TC)_2$
HMS3	9	compound	$(TG)_2(CA)_2TC(CA)_n$ and $(TG)_2(CA)_2TC(CA)_nGA(CA)_5$
HMS6	4	simple	$(GT)_n$
HMS7	1q25	compound	$(AC)_2(CA)_n$
HTG4	9	compound	$(TG)_nAT(AG)_5AAG(GA)_5ACAG(AGGG)_3$
HTG6	15q26–q27	simple	$(TG)_n$
HTG7	4	simple	$(GT)_n$
HTG10	21	simple and compound	$(TG)_n$ and $TATC(TG)_n$
LEX3	Xq	simple	$(TG)_n$
VHL20	30	simple	$(TG)_n$

The accuracy of genotyping was further assessed through the laboratory's participation in the 2024–2025 ISAG International Horse STR DNA Typing Comparison Test (100 participating laboratories). The laboratory achieved the highest absolute genotyping accuracy rank (rank 1, 98–100%), based on the 12 core ISAG markers.

2.4. Data Analysis

2.4.1. Data Quality Control and Compilation of the Final Dataset

All 765 samples were genotyped at all loci, including repeated genotyping (100% data completeness across loci and samples). Duplicates were identified by pairwise comparison of multilocus genotypes; pairs of animals differing at no more than two loci were rechecked against the electropherograms. The X-linked locus LEX3 was excluded from all analyses because it is hemizygous in males, which distorts estimates of heterozygosity and F. To assess the effect of its exclusion, H_O, H_E, F, and the STRUCTURE analysis were compared for the original 765 animals using 17 and 16 loci.

Genotyping errors due to stutter peaks and large-allele dropout, as well as null alleles, were assessed in each group using MICRO-CHECKER v2.2.3 [20] (dinucleotide repeat model, 1000 randomizations, 95% confidence interval). Null allele frequencies were estimated using the van Oosterhout, Chakraborty, and Brookfield (1 and 2) methods implemented in the program.

Close relatives were identified by sibship reconstruction in COLONY v2.0.7.2 [21] using the full-likelihood method [22,23]. Each group was analyzed separately, so that allele frequencies were estimated from the sample of that group; five runs were performed per group (seed 1234), with an allelic dropout rate of 0 and another-error rate of 0.01 for all loci. Full-sib pairs with a probability ≥ 0.95 were considered reliable. A full-sib family was defined as a connected chain of such pairs; one animal was retained from each family (the one with the lowest sample number). Half-sibs were not removed.

All subsequent analyses were performed on the final dataset (n = 703, 16 autosomal loci) unless stated otherwise. Deviations from Hardy–Weinberg equilibrium (HWE) and linkage disequilibrium (LD) were tested in each group in R v4.5.2–4.6.1 [24] using the genepop package v1.2.17 [25]. HWE was tested for each locus using an exact test with the *p* value estimated by a Markov chain method [26]; genotypic LD was tested for all 120 pairs of loci in each group. Chain parameters were 10,000 dememorization steps and 100 batches of 5000 iterations (random seed 67144630). *p* values were Bonferroni-corrected within each group (16 tests for HWE, 120 for LD).

2.4.2. Assessment of Diversity and Differentiation

Genetic diversity was assessed using GenAlEx version 6.51 software [27]. For each study group, the following parameters were calculated: allele frequencies (Table S2), number of alleles per locus (Na), effective number of alleles (Ne), Shannon's information index (I), observed (H_O) and expected (H_E) heterozygosity, unbiased expected heterozygosity (uH_E), the fixation index (F), and Nei's genetic distances [28].

Means and SE for each study group were calculated across the 16 STR loci. To account for differences in sample size, allelic richness (Ar) and private allelic richness (PAr) were estimated by rarefaction in ADZE v1.0 [29], standardized to the smallest sample (g = 18 gene copies, corresponding to the CROSS group, n = 9).

Differentiation was assessed using the multilocus F_{ST} statistic of Weir and Cockerham [30]. Overall F_{ST}, F_{IS}, and F_{IT}, as well as their per-locus values, were obtained by analysis of molecular variance (AMOVA) in GenAlEx (allelic distance); pairwise F_{ST} were obtained in the same way. Significance was tested with 999 permutations, with Bonferroni correction for 16 loci and 10 pairs of groups. The same estimates were calculated in R using the hierfstat package v0.5-11 [31]; 95% confidence intervals (CI) for overall and pairwise F_{ST} were obtained by bootstrapping over loci (1000 replicates). The number of migrants per generation was calculated from multilocus F_{ST} under Wright's island model [32]: $Nm = (1/F_{ST} - 1)/4$; the CI for Nm was obtained by applying the same transformation to the CI limits of F_{ST}.

2.4.3. Analysis of Genetic Structure

Cluster analysis was performed using STRUCTURE version 2.3.4, based on a Bayesian algorithm and the Markov chain Monte Carlo (MCMC) method, under the admixture model with independent allele frequencies, without prior group information [33]. The number of genetic clusters (*K*) was evaluated from 1 to 10, with 10 independent runs conducted for each *K* value. Each run included 250,000 burn-in iterations, followed by 1,000,000 MCMC iterations.

Principal coordinates analysis (PCoA) was performed based on a matrix of Nei's genetic distances using GenAlEx version 6.51 software [27].

The number of clusters was estimated in StructureSelector [34] using two approaches: the Δ*K* method [35] and four Puechmaille estimators [36] (MedMedK, MedMeanK, MaxMedK, MaxMeanK). Replicate runs were aligned and modes were identified in

CLUMPAK [37] (CLUMPP, LargeKGreedy algorithm, 2000 repeats; MCL threshold 0.1). For the comparison of 17 and 16 loci, runs on all 765 animals were used.

As a non-Bayesian complement to STRUCTURE, discriminant analysis of principal components (DAPC) [38] was performed in R with the adegenet package v2.1.11 [39]. First, the number of clusters was estimated without group information: k-means clustering on the principal components of allele frequencies for K = 1–10 (100 random starts), compared by the Bayesian information criterion (BIC) using the "diffNgroup" selection criterion. For K = 2–5, DAPC was performed on the inferred clusters and plots of individual membership probabilities were produced. DAPC was then performed using the five predefined groups. In both cases, the number of principal components was selected by cross-validation (xvalDapc; 90% training set, 30 repetitions, lowest mean squared error). Assignment accuracy was evaluated by leave-one-out cross-validation of linear discriminant analysis on the retained components (MASS package v7.3-66 [40], with either equal prior probabilities for all groups or priors proportional to group sizes).

An individual-level assignment test and PCoA were also performed in GenAlEx version 6.51. The assignment test was based on the log-likelihood of the multilocus genotype in each group; the tested individual was excluded from the allele frequency calculation (leave-one-out), and zero frequencies were replaced with 0.01. The individual-level PCoA was based on the genetic distance matrix for codominant markers, using the standardized covariance matrix.

To assess the effect of unequal sample sizes, 10 balanced datasets were created: 36 animals (the size of the largest documented group, KT) were randomly drawn without replacement from the UB group and combined with all animals of the other groups, giving 144 animals per dataset. For each dataset, STRUCTURE (K = 1–6), DAPC, and overall F_{ST} were calculated with the same settings as for the full dataset.

3. Results

3.1. Data Quality Control

Data quality control (MICRO-CHECKER and comparison of multilocus genotypes) identified five HTG4 genotypes carrying an allele inconsistent with a dinucleotide repeat, and four pairs of nearly identical genotypes. After re-examination of the electropherograms at these loci, the HTG4 genotypes were corrected, and the four pairs proved to be repeated samples of the same animals; the duplicate records were removed. The further analysis included 761 animals (CROSS 9, KST 32, KT 39, TB 31, UB 650).

On the corrected data, MICRO-CHECKER detected no evidence of stutter peaks or large-allele dropout at any of the 16 loci in any group. A null-allele signal was detected only at HMS3 in the UB group (0.0230–0.0305; Table S3); HMS3 was therefore retained in the analysis.

COLONY detected no duplicated genotypes in any group. The final log-likelihoods of the five runs were identical in CROSS (−284.77) and differed by less than 3.2 units in KT, TB, and KST. In UB the difference reached 73 units (from −21,863.3 to −21,936.4), but the number of full-sib pairs in the best configuration was similar across runs (173–182). Full-sib pairs with a probability ≥ 0.95 were found only in KT (3 pairs, three families of two animals) and UB (59 pairs, 95 animals in 40 families of 2–5 animals). In the other groups, the highest full-sib pair probability was 0.826 (KST), 0.276 (CROSS), and 0.261 (TB). Half-sib pairs with a probability ≥ 0.95 were found only in UB (8 pairs). After removal of 3 animals from KT and 55 from UB, 703 animals remained in the analysis (UB 595, KT 36, KST 32, TB 31, CROSS 9; Table S4).

After Bonferroni correction, no locus deviated from HWE in any group (Table S5). Before correction, $p < 0.05$ was observed in only three of 80 locus × group combinations

(HMS3 and ASB2 in UB, HTG7 in KST). Significant LD after correction was found only in UB (17 of 120 locus pairs) and in none of the other groups (Table S6). Because LD in UB involved many different locus pairs and was not observed in the other groups, it more likely reflects the heterogeneity of the UB group than physical linkage between loci.

In the original 765 animals, excluding LEX3 changed H_O and H_E little but removed the heterozygote deficit: F in TB decreased from 0.029 to 0.002, and in UB from 0.014 to −0.010 (Table S7). At the LEX3 locus itself, F ranged from 0.16 (CROSS) to 0.46 (TB).

3.2. Allelic Diversity and Allele Frequencies

Analysis of allele frequencies across the 16 STR loci revealed a high level of polymorphism in all studied groups (Table S8). A total of 155 distinct alleles were detected across the analyzed loci, and the complete allele frequency distribution is provided in Table S2. The number of alleles detected per locus ranged from 2 (HTG7 in the CROSS group) to 16 (ASB17 in the UB group). The most polymorphic loci were ASB17, HTG10, ASB2, ASB23, and CA425, whereas HTG7 exhibited the lowest level of polymorphism.

The greatest allelic diversity was observed in the UB group, which displayed the broadest allelic spectrum among all studied groups. In contrast, the groups with documented breed affiliation (KST, KT, and TB) exhibited fewer alleles. Most rare alleles occurred at frequencies below 5%, whereas a limited number of common alleles accounted for the majority of the allelic spectrum at each locus.

Despite differences in the number of alleles, the distribution of the most frequent alleles was similar across all studied groups (Table S2).

At the AHT4 locus, allele 157 was the most frequent allele in four of the five groups, whereas allele 149 predominated only in the CROSS group. Similarly, alleles 222, 238, and 148 were the most frequent alleles at the HMS2, CA425, and HMS3 loci, respectively.

The UB group displayed a greater diversity of rare alleles that were either absent or present at low frequency in the other groups. This feature indicates a broader allelic spectrum within the studied sample and provided the basis for its further analysis using genetic diversity indices and population structure analyses.

3.3. Genetic Diversity Indices

The average values of genetic diversity indices, calculated across the 16 STR loci, are presented in Table 3.

Table 3. Mean (±SE) values of genetic diversity parameters across 16 STR loci in the five studied horse groups.

Group (n)	Na	Ar	PAr	Ne	I	H_O	H_E	uH_E	F
CROSS (9)	4.250 ± 0.296	4.250 ± 0.296	0.158 ± 0.049	3.260 ± 0.263	1.244 ± 0.077	0.729 ± 0.042	0.661 ± 0.028	0.700 ± 0.030	−0.099 ± 0.040
KST (32)	6.063 ± 0.335	4.673 ± 0.239	0.213 ± 0.064	3.558 ± 0.233	1.402 ± 0.064	0.715 ± 0.025	0.700 ± 0.020	0.711 ± 0.020	−0.025 ± 0.029
KT (36)	6.438 ± 0.387	4.948 ± 0.216	0.413 ± 0.091	3.575 ± 0.219	1.452 ± 0.056	0.736 ± 0.024	0.705 ± 0.018	0.715 ± 0.018	−0.045 ± 0.026
TB (31)	5.688 ± 0.405	4.537 ± 0.265	0.124 ± 0.037	3.567 ± 0.259	1.376 ± 0.073	0.696 ± 0.030	0.695 ± 0.024	0.706 ± 0.024	0.002 ± 0.022
UB (595)	9.688 ± 0.669	4.969 ± 0.253	0.346 ± 0.066	3.974 ± 0.314	1.530 ± 0.070	0.732 ± 0.021	0.725 ± 0.021	0.725 ± 0.021	−0.011 ± 0.009

n, number of animals; Na, number of different alleles; Ar, allelic richness; PAr, private allelic richness; Ne, number of effective alleles; I, Shannon's information index; H_O, observed heterozygosity; H_E, expected heterozygosity; uH_E, unbiased expected heterozygosity; F, fixation index. Ar and PAr were estimated by rarefaction to 18 gene copies (ADZE). SE was calculated across loci.

The highest mean number of observed alleles (Na) was recorded in the UB group (9.688 ± 0.669), whereas the remaining groups exhibited values ranging from 4.250 ± 0.296 (CROSS) to 6.438 ± 0.387 (KT).

After rarefaction to 18 gene copies, Ar ranged from 4.25 (CROSS) to 4.97 (UB) and differed little among the groups (KST 4.67, TB 4.54, KT 4.95). Therefore, the high Na in UB mainly reflects its sample size: after rarefaction, Ar in UB was almost equal to that in KT. Private allelic richness was low in all groups (0.12–0.41 alleles per locus): it was highest in KT (0.41) and UB (0.35) and lowest in TB (0.12). The effective number of alleles ranged from 3.260 ± 0.263 (CROSS) to 3.974 ± 0.314 (UB), and Shannon's information index from 1.244 ± 0.077 (CROSS) to 1.530 ± 0.070 (UB).

Heterozygosity was high in all groups: H_O ranged from 0.696 ± 0.030 (TB) to 0.736 ± 0.024 (KT), H_E from 0.661 ± 0.028 (CROSS) to 0.725 ± 0.021 (UB), and uH_E from 0.700 ± 0.030 (CROSS) to 0.725 ± 0.021 (UB). The fixation index was close to zero or negative in all groups, ranging from −0.099 (CROSS) to 0.002 (TB); it was −0.025 in KST, −0.045 in KT, and −0.011 in UB. No pronounced heterozygote deficiency was observed in any group.

All groups had high heterozygosity and similar allelic richness after correction for sample size.

3.4. Genetic Differentiation Among the Studied Groups

F-statistics (F_{IS}, F_{IT} and F_{ST}) were estimated by AMOVA for each locus and across all loci and are presented in Table 4.

Table 4. F-statistics (AMOVA) across 16 STR loci in the five horse groups.

Locus	F_{IS}	F_{IT}	F_{ST}	*P* (F_{ST})
AHT4	−0.070	−0.051	0.0175	0.001 *
AHT5	−0.029	−0.026	0.0033	0.180
ASB17	−0.007	0.000	0.0071	0.038
ASB2	0.001	0.007	0.0057	0.041
ASB23	−0.003	0.007	0.0095	0.011
CA425	−0.016	−0.013	0.0024	0.217
HMS1	−0.026	−0.019	0.0068	0.099
HMS2	0.018	0.025	0.0072	0.058
HMS3	0.042	0.041	−0.0011	0.573
HMS6	−0.051	−0.036	0.0149	0.016
HMS7	0.043	0.052	0.0087	0.016
HTG10	−0.023	−0.018	0.0057	0.062
HTG4	0.041	0.082	0.0430	0.001 *
HTG6	−0.031	−0.016	0.0147	0.007
HTG7	−0.041	−0.040	0.0014	0.337
VHL20	−0.006	−0.006	−0.0001	0.408
Mean	−0.0095	−0.0006	0.0088 (95% CI 0.0051–0.0142)	0.001

P, permutation test (999 permutations); * significant after Bonferroni correction for 16 loci ($p < 0.0031$). The 95% CI for the multilocus F_{ST} was obtained by bootstrapping over loci (1000 replicates).

Locus-specific F_{ST} ranged from −0.001 (HMS3) to 0.043 (HTG4). Differentiation was significant at eight loci ($p < 0.05$) and remained significant after Bonferroni correction at AHT4 and HTG4. Across all loci, differentiation was low but significant ($F_{ST} = 0.0088$, 95% CI 0.0051–0.0142; $p = 0.001$), and there was no heterozygote deficiency within groups ($F_{IS} = -0.0095$; $p = 0.921$) (Table 5).

Table 5. Analysis of molecular variance (AMOVA) and F-statistics, n = 703.

Source of Variation	df	SS	MS	Variance Component	%	F-Statistic	*P*
Among groups	4	42.932	10.733	0.051	0.87	F_{ST} = 0.0088	0.001
Among individuals within groups	698	3997.568	5.727	0.000	0.00	F_{IS} = −0.0095	0.921
Within individuals	703	4103.500	5.837	5.837	99.13	F_{IT} = −0.0006	0.573
Total	1405	8144.000	—	5.889	100	—	—

P, permutation test (999 permutations).

The number of migrants per generation, calculated from the overall F_{ST} (Table 5) and its 95% CI (Table 4), was Nm = 28.1 (95% CI 17.4–48.6).

According to the AMOVA results, 0.87% of the total variation was attributable to differences among groups and 99.13% to variation within individuals; the component among individuals within groups was zero (Table 5). The Weir and Cockerham estimates obtained in hierfstat agreed with the AMOVA estimates.

Pairwise F_{ST} ranged from 0.0017 (KST–UB) to 0.0463 (CROSS–KT) (Table 6). After Bonferroni correction, differentiation was significant for the pairs KT–TB, CROSS–KT, TB–UB, and CROSS–UB; the 95% CIs included zero for the pairs KST–UB and CROSS–TB. Because the CROSS group is small (n = 9), F_{ST} values for pairs involving this group should be interpreted with caution.

Table 6. Pairwise F_{ST} (Weir and Cockerham), n = 703.

Pair	F_{ST}	95% CI	*P*
CROSS–KST	0.0150	0.0020–0.0297	0.060
CROSS–KT	0.0463	0.0234–0.0757	0.002 *
CROSS–TB	0.0051	−0.0082–0.0194	0.225
CROSS–UB	0.0284	0.0121–0.0489	0.003 *
KST–KT	0.0101	0.0038–0.0181	0.022
KST–TB	0.0076	0.0013–0.0138	0.038
KST–UB	0.0017	−0.0005–0.0040	0.167
KT–TB	0.0289	0.0134–0.0508	0.001 *
KT–UB	0.0083	0.0035–0.0129	0.005
TB–UB	0.0090	0.0023–0.0180	0.002 *

95% CI, confidence interval obtained by bootstrapping over loci (1000 replicates); *P*, permutation test (999 permutations); * significant after Bonferroni correction for 10 pairs ($p < 0.005$).

In the 10 balanced subsamples, the overall F_{ST} ranged from 0.0120 to 0.0157 (mean 0.0136), and none of the 95% CIs included zero (lower limits 0.0064–0.0095) (Table S9; composition of the subsamples in Table S10). These values were slightly higher than in the full dataset, in which the estimate is dominated by the large UB group, but remained low. Thus, differentiation among the studied groups was low but statistically significant, and this conclusion does not depend on unequal sample sizes.

3.5. Genetic Structure of the Studied Groups

The mean log-likelihood, Ln$P(K)$, increased monotonically from $K = 1$ (−34,427.2 ± 0.05) to $K = 10$ (−33,557.4 ± 6.7) without reaching a plateau (Table S11). All four Puechmaille estimators indicated $K = 2$, whereas ΔK was highest at $K = 4$ ($\Delta K = 47.9$), with smaller peaks at $K = 7$ (19.4) and $K = 6$ (16.2) (Figure 2). The ΔK peak at $K = 4$ coincided with a sharp increase in variation among runs at $K = 5$ (SD of Ln$P(K)$ = 35.2 versus 2.1–5.3 at $K = 2$–4). CLUMPAK identified a single mode across all 10 runs at $K = 1$–4, 6, and 9, and two to three modes at $K = 5$, 7, 8, and 10.

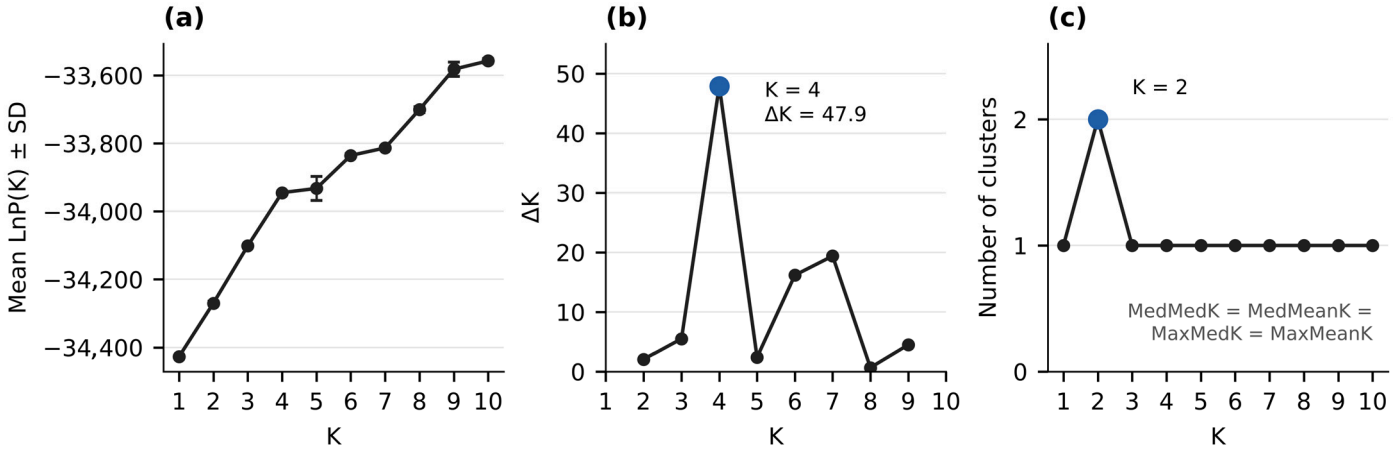

Figure 2. Selection of the number of genetic clusters (K) for the final dataset (n = 703): (**a**) mean Ln$P(K)$ ± SD over 10 runs; (**b**) ΔK [35]; (**c**) Puechmaille estimators [36]. MedMeanK and MaxMeanK are the median and the maximum across runs of the number of clusters in which at least one predefined group has a mean membership coefficient > 0.5; MedMedK and MaxMedK are the same quantities calculated using the median membership coefficient of each group.

Individual membership coefficients were highly admixed at all values of K (Figure 3). At $K = 2$, the mean proportion of the first cluster was 0.72 in CROSS, 0.67 in TB, 0.59 in KST, 0.51 in UB, and 0.43 in KT; the second cluster exceeded the 0.5 threshold only in KT (0.57), and only 14.9% of the animals had a membership coefficient ≥ 0.8 to one of the clusters. At $K = 3$–5, the mean membership of UB was close to 1/K for each cluster, and no group formed a separate cluster.

The Puechmaille estimators indicated $K = 2$ in the original 17-locus dataset and in both 16-locus datasets, and the mean membership of the groups remained practically unchanged. With LEX3, ΔK was also highest at $K = 2$ ($\Delta K = 94.0$); after exclusion of this locus, this peak disappeared ($\Delta K = 0.9$–2.0), and the highest ΔK shifted to $K = 4$. In none of the datasets did the identified clusters correspond to the predefined groups (Table S7).

In the 10 balanced subsamples (n = 144), the Puechmaille estimators indicated $K = 2$ in all cases, and the mean membership of the groups at $K = 2$ was close to that in the full dataset (Table S12). In contrast, the maximum ΔK differed among subsamples ($K = 2$ in two, $K = 3$ in five, $K = 4$ in three), and in five subsamples the mean LnP at $K = 2$ was lower than at $K = 1$. Thus, equalizing group sizes did not reveal additional structure.

BIC decreased gradually with increasing K, from 1235.9 ($K = 1$) to a minimum of 1194.9 ($K = 9$), with a difference of less than 6 units after $K = 5$ and without a clear inflection point (Figure S1). The diffNgroup criterion indicated $K = 2$, but the two clusters found did not correspond to the predefined groups: KST animals were split equally (16 and 16), and UB animals in similar proportions (266 and 329) (Table S13). At $K = 3$–5, individuals of all groups showed mixed membership (Figure S2).

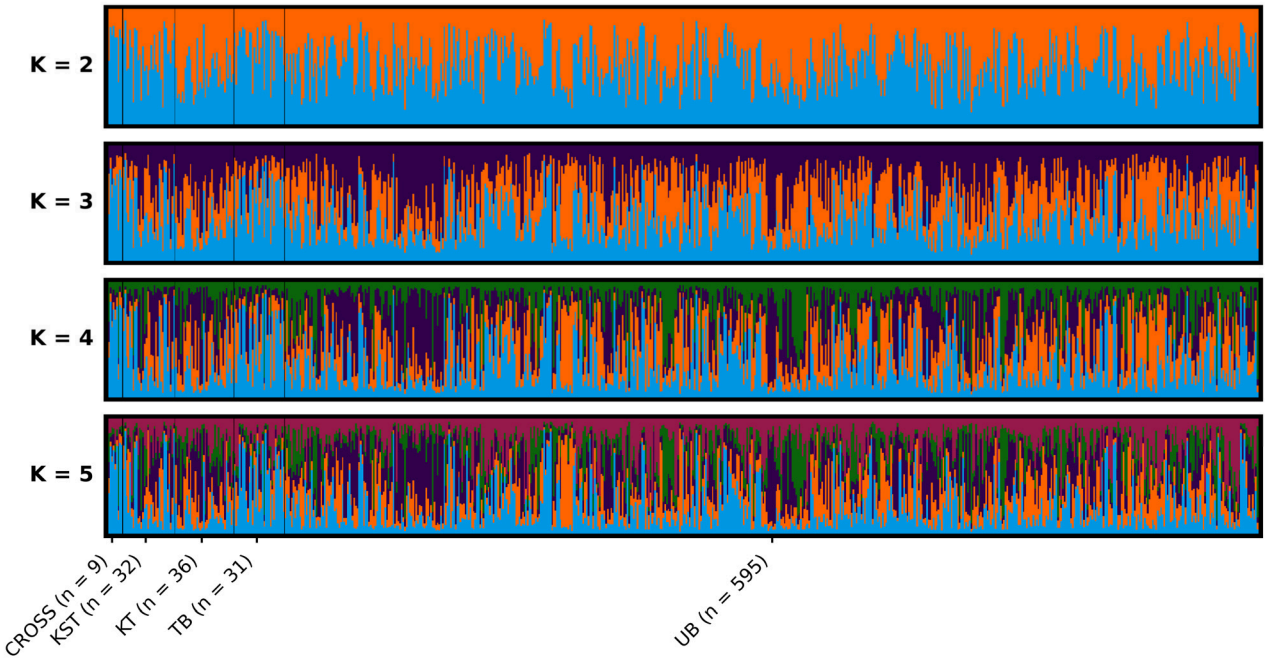

Figure 3. STRUCTURE results at K = 2–5 (n = 703). Each bar represents one animal; colors indicate membership coefficients for the clusters.

In the DAPC with predefined groups (80 principal components, 98% of variance), leave-one-out assignment with equal prior probabilities correctly assigned 33.3% of CROSS, 40.6% of KST, 50.0% of KT, 45.2% of TB, and 52.8% of UB (Table S13). With prior probabilities proportional to group sizes, 92.8% of UB animals were correctly assigned, but most animals of the small groups were assigned to UB (correct assignment 3.1–33.3%). In the subsamples, DAPC also did not yield a stable number of clusters: diffNgroup indicated K = 2 in two subsamples, K = 3 in three, K = 4 in three, and K = 5 in two, and the minimum BIC was at K = 3 in eight subsamples and at K = 2 in two. The mean proportion of correct assignment with equal prior probabilities was 45.5% for CROSS (range 33.3–55.6%), 30.0% for KST (21.9–40.6%), 57.5% for KT (50.0–66.7%), 37.4% for TB (22.6–48.4%), and 32.2% for UB (19.4–44.4%) (Table S9).

The assignment test with exclusion of the tested individual correctly assigned 240 of 703 animals (34.1%) to their own group: 55.6% of CROSS, 40.6% of KST, 50.0% of KT, 38.7% of TB, and 32.3% of UB (Table 7). UB animals were distributed among all groups: 136 to KST, 109 to KT, and 143 to TB. The proportions of correct assignment for KST, KT, and TB were close to the DAPC results with equal prior probabilities.

Table 7. Results of the assignment test, n = 703.

Original Group (n)	→ CROSS	→ KST	→ KT	→ TB	→ UB	Correct, %
CROSS (9)	5	1	0	3	0	55.6
KST (32)	4	13	2	5	8	40.6
KT (36)	0	8	18	3	7	50.0
TB (31)	2	9	0	12	8	38.7
UB (595)	15	136	109	143	192	32.3

Note: The arrow (→) indicates the group to which an individual was assigned.

The first two axes of the individual-level PCoA explained 4.8% and 4.2% of the variation (three axes, 12.9%). The 95% ellipses of the groups almost completely overlapped; only TB

was slightly shifted toward negative values of the first axis, remaining within the overall cloud of points (Figure 4).

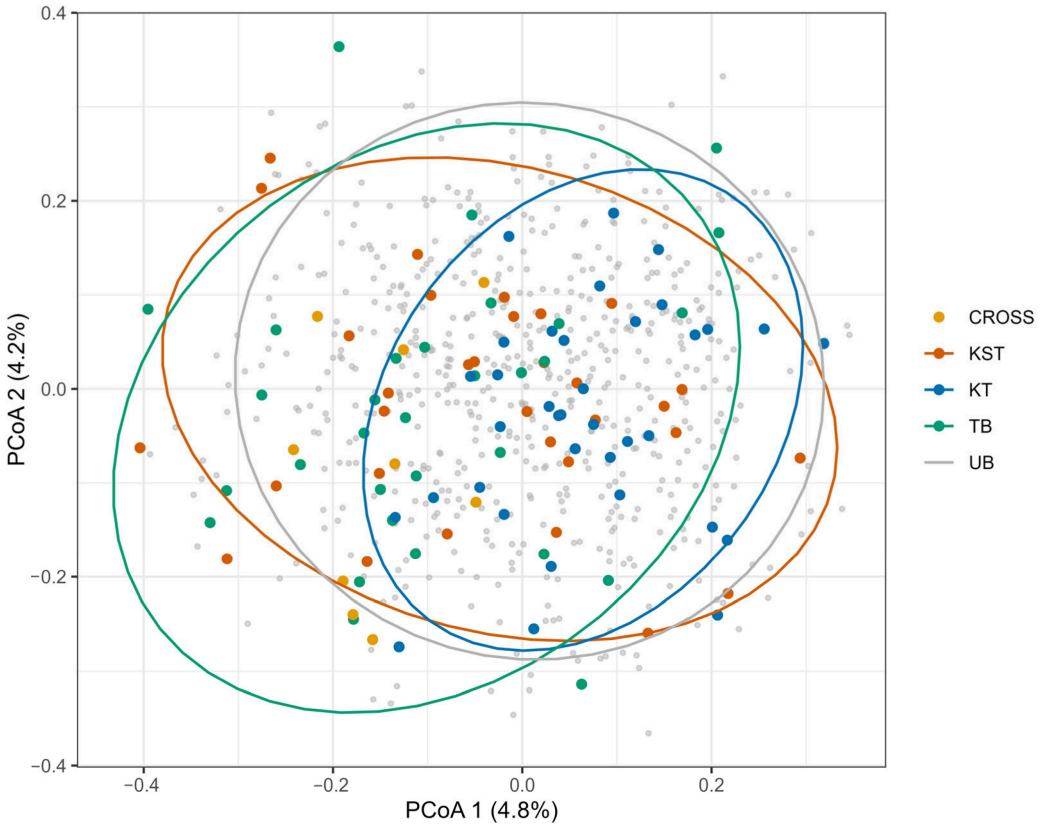

Figure 4. Principal coordinates analysis of individuals (n = 703). UB animals are shown in gray; 95% ellipses are shown for all groups except CROSS (n = 9).

Pairwise Nei's genetic distance analysis likewise demonstrated a high degree of genetic similarity among the studied groups (Table 8).

Table 8. Pairwise Nei's genetic distances among the five studied horse groups.

	CROSS	KST	KT	TB	UB
CROSS	0.000				
KST	0.116	0.000			
KT	0.201	0.061	0.000		
TB	0.093	0.057	0.111	0.000	
UB	0.139	0.024	0.040	0.043	0.000

All pairwise genetic distances were low, ranging from 0.024 to 0.201, indicating only weak genetic differentiation among the studied groups.

The smallest genetic distance was observed between the KST and UB groups (0.024), followed by the KT and UB groups (0.040) and the TB and UB groups (0.043), indicating a high degree of genetic similarity among these groups. Relatively small genetic distances were also recorded between the KST and KT groups (0.061) and between the KST and TB groups (0.057). The largest genetic distance was observed between the CROSS and KT groups (0.201). Likewise, the genetic distances between the CROSS group and the other groups (0.093–0.201) were higher than those observed for most other group pairs. When interpreting the results, the small sample size of the CROSS group (n = 9), which may be insufficient for estimating genetic distances, must be taken into account; therefore, the

observed genetic remoteness of the CROSS group should be interpreted with caution and not as evidence of unambiguous genetic differentiation. The genetic similarity between KST and UB indicates some shared genetic characteristics between the two groups; however, the genetic heterogeneity and potentially different origin of the UB horses must be taken into account.

The results of the principal coordinates analysis (PCoA) were consistent with those obtained from Nei's genetic distance and STRUCTURE analyses (Figure 5).

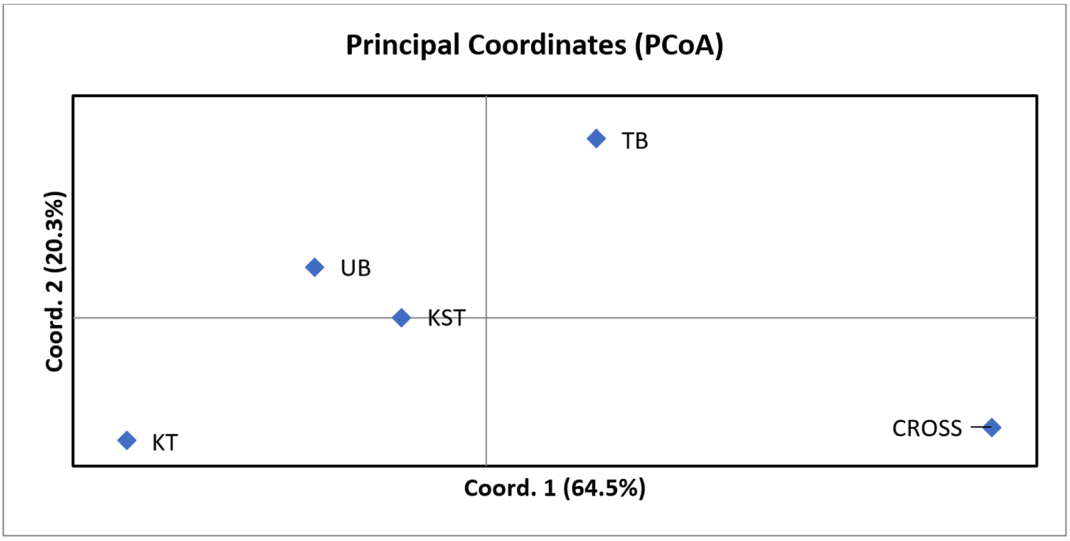

Figure 5. Principal coordinates analysis (PCoA) based on the matrix of Nei's genetic distances among the five studied horse groups.

The studied groups did not form clearly distinct clusters, and their distribution in the principal coordinate space was characterized by substantial overlap. The KST and UB groups were located in close proximity, consistent with the minimal Nei's genetic distance observed between them. Along the first coordinate (64.5% of the variation), CROSS and KT occupied the extreme positions, whereas KST and UB were located centrally and close to each other; TB occupied an intermediate position on the first coordinate and was displaced along the second coordinate (20.3%). However, the spatial separation among groups remained limited and was not accompanied by the formation of distinct clusters. It should be taken into account that the CROSS group is represented by a small number of individuals (n = 9), which limits the reliability of the interpretation of the results. Therefore, the remoteness of the CROSS group revealed by the PCoA cannot unambiguously indicate its genetic distinctness.

Overall, STRUCTURE, DAPC, the assignment test, individual-level PCoA, and Nei's genetic distances consistently show low differentiation: none of the methods identified the studied groups as separate genetic clusters, which is consistent with the AMOVA results.

4. Discussion

4.1. Genetic Diversity of the Kustanay Horse Breed

The high level of allelic polymorphism observed across all studied horse groups confirms the suitability of 17-locus STR panel including the 12 ISAG core markers for assessing genetic diversity in the Kustanay horse breed. Comparable applications of microsatellite markers have been reported for other horse breeds, where they have proved informative for evaluating genetic diversity and population structure [41–43]. The present results are also consistent with previous studies of the Kustanay breed [4,13], while substantially

expanding the available information by including 765 genotyped animals (703 after quality control and removal of close relatives), comprising both registered horses and animals of undocumented origin.

The UB group had the highest number of alleles, including all alleles found in the study. However, after rarefaction to 18 gene copies, allelic richness in UB was practically equal to that in KT. Therefore, the larger number of alleles in UB is explained mainly by sample size rather than by a broader allelic spectrum. Similar patterns of substantial genetic diversity have also been reported in Kazakh and other domestic horse populations using high-density SNP genotyping [2,16] and mitochondrial DNA analyses [3]. Although these marker systems characterize different components of genetic variation, together they indicate that considerable genetic diversity remains within domestic horse populations.

The obtained heterozygosity values were comparable to the results of previous studies of the Kustanay breed using the same 17-marker panel [4,13] and fall within the range typical of other horse breeds, including the Kushum, Mugalzhar, Akhal-Teke, and Thoroughbred breeds [14,15,43,44].

Fixation indices were close to zero or negative in all groups, and the overall F_{IS} from AMOVA did not differ from zero. The slight heterozygote deficit in TB and UB in the original analysis was associated with the X-linked locus LEX3: after its exclusion, F decreased in the same 765 animals. Male horses are hemizygous at this locus, and recording their genotypes as apparent homozygotes can artificially inflate estimates of heterozygote deficiency. Overall, the fixation indices did not indicate a pronounced heterozygote deficit. Similar effects have been reported in Arabian horses genotyped with the same microsatellite panel [45]. The absence of deviations from HWE after correction for multiple comparisons and the low frequency of null alleles (only HMS3 in UB) indicate no heterozygote deficit attributable to inbreeding in the final dataset (from which close relatives had been removed) and the reliability of genotyping.

4.2. Genetic Differentiation and Population Structure

The multilocus F_{ST} indicates weak differentiation among the groups according to Wright's classification [46]; in the balanced subsamples, it was slightly higher but remained low. The Nm estimate, calculated from F_{ST} under the island model, is consistent with the observed low genetic differentiation but should not be interpreted as independent evidence of contemporary gene flow or panmixia. This estimate depends on the assumptions of the underlying model. The observed differentiation may reflect both historical gene exchange and the recent common origin of the groups. In addition, differences in sample sizes and the possible heterogeneity of the UB group must be taken into account. Therefore, the available data do not allow definitive conclusions about the intensity of current gene flow among the studied groups.

AMOVA supported the same general pattern. In the complete dataset, only 0.87% of the total variation was attributable to differences among groups, whereas 99.13% occurred within individuals.

These findings are consistent with previous studies of domestic horses. Orazymbetova et al. [14], using STR markers, and Cosgrove et al. [47], using SNP genotyping and AMOVA, reported that 86% of the genetic variation was distributed within individuals in a comparative analysis involving Arabian horses and other breeds. Studies of Kazakh horses based on SNP data [2] and mitochondrial DNA [3] have similarly demonstrated substantial genetic diversity combined with relatively weak differentiation among populations. Such patterns are compatible with the complex breeding history of Kazakh horses and the historical exchange of genetic material among breeding populations. When interpreting the results of STR genotyping, the possibility of homoplasy must be taken into account. Therefore, the

low F_{ST} values obtained at microsatellite loci cannot serve as definitive evidence of genetic homogeneity of the studied groups. More accurate assessment of genetic structure requires further studies using high-density SNP arrays.

The low level of differentiation observed in the present study may also reflect the formation history of the Kustanay breed. The breed originated through prolonged crossbreeding of local Kazakh mares with Don, Thoroughbred, Streletskaya, and Astrakhan (Kalmyk) stallions, followed by long-term selection within a relatively interconnected breeding population [2,4]. The exchange of breeding stallions among farms and the repeated use of influential sire lines may also have contributed to the maintenance of a shared genetic background and limited the development of strongly differentiated subpopulations.

The two approaches to choosing K gave different results ($K = 2$ by the Puechmaille estimators and $K = 4$ by ΔK), but neither solution corresponded to the predefined groups. With weak differentiation and unequal sample sizes, ΔK tends to produce spurious peaks and cannot evaluate $K = 1$, whereas the Puechmaille estimators are more robust to unequal sample sizes [36,48]. In our data, the solution obtained with the Puechmaille estimators was characterized by only moderate membership of animals in the identified clusters, and $\mathrm{Ln}P(K)$ increased without reaching a plateau. In half of the balanced subsamples, LnP at $K = 2$ was lower than at $K = 1$. Together with the absence of a clear BIC minimum in DAPC and the low assignment accuracy in GenAlEx, this shows that the studied groups do not form separate genetic clusters. A similar combination of STRUCTURE, PCoA, and phylogenetic approaches also revealed weak subdivision among closely related native horse breeds from Xinjiang [49].

Nei's genetic distances were consistent with the close genetic relationships among the studied groups. The smallest distance was observed between KST and UB; KT and TB were similarly close to UB. In the PCoA, KST and UB were located close together in the central part of the plot, whereas CROSS and KT occupied the extreme positions along the first coordinate (Figure 5). The position of CROSS should be interpreted with caution because of the small sample (n = 9). These differences were small and did not lead to the formation of clearly separated clusters.

The small genetic distance between the KST and UB groups indicates their genetic similarity at the group level. This result does not confirm that individual animals from the UB group belong to the Kustanay breed. The overall accuracy of individual assignment was 34.1%, which indicates the limited ability of the markers used to distinguish the studied groups. Despite leave-one-out cross-validation, the results obtained are insufficient to reliably confirm the breed status of individual animals without documented origin. Therefore, molecular genetic data should be interpreted together with reliable pedigree records. Assessing the possibility of individual breed assignment requires further studies using registered Kustanay horses as a reference group.

Overall, the complementary results of F_{ST}, AMOVA, STRUCTURE, DAPC, the assignment test, Nei's genetic distances, and PCoA consistently indicate substantial genetic diversity together with weak differentiation among the studied groups. The agreement among these independent analytical approaches strengthens the conclusion that the studied groups share a largely common genetic background while retaining some differences related to their breeding history. The consistency in the distribution of the predominant alleles suggests the presence of a shared genetic background among the studied groups.

4.3. Study Limitations and Future Directions

Several limitations should be considered when interpreting the present results. First, the study groups differed substantially in sample size, particularly because of the large UB group (n = 595 after removal of close relatives).

To account for this, diversity was compared using rarefied allelic richness, and STRUCTURE, DAPC, and F_{ST} were repeated on 10 balanced subsamples; the conclusions did not change. Second, the CROSS group is represented by only nine animals. This reduces the power of tests for null alleles and HWE, and estimates of relatedness and allele frequencies in this group are based on a very small number of individuals; therefore, the results for CROSS should be interpreted with caution.

Despite these limitations, the large overall dataset and the consistency of results obtained using several complementary approaches provide a robust population-level characterization of the contemporary Kustanay horse gene pool. Future studies should increase the number of documented Kustanay animals, and complement STR analysis with high-density SNP genotyping. Integration of molecular data with detailed pedigree information and mitochondrial DNA may further improve estimates of effective population size, ancestry, and inbreeding. Such combined approaches have been successfully applied to the Sardinian Anglo-Arab horse [50] and the Pura Raza Español breed [51] and could provide a more detailed framework for the long-term genetic management and conservation of the Kustanay horse.

5. Conclusions

The present study represents the most comprehensive analysis to date of the genetic diversity and population structure of the Kustanay horse breed, based on the genotyping of 765 animals using a 17-locus STR panel including the 12 core markers recommended by ISAG (703 animals and 16 autosomal loci after quality control). A high level of genetic diversity, comparable to values reported for other horse breeds, was found in all studied groups. However, heterozygosity values alone do not allow conclusions about the preservation of the entire genetic potential of the Kustanay breed.

A comprehensive analysis incorporating genetic diversity indices, F_{ST}, AMOVA, Nei's genetic distances, Bayesian clustering (STRUCTURE), and principal coordinates analysis (PCoA) revealed only weak genetic differentiation and indicated a largely shared genetic background among the studied groups. The high Nm value is consistent with low genetic differentiation but does not allow unambiguous conclusions about the intensity of historical or contemporary gene flow among the studied groups. The greatest genetic similarity was observed between the registered Kustanay horse (KST) and UB groups. However, because most of the founder breeds were not included in the study, it cannot be excluded that this similarity reflects a broader genetic background of Kazakh horses. Thus, the results characterize genetic relationships at the group level but do not confirm that individual UB animals belong to the Kustanay breed. The limited accuracy of individual assignment underlines the need to combine molecular genetic data with reliable pedigree information.

The results support the applicability of the ISAG STR marker panel for assessing genetic diversity in the Kustanay horse breed, in addition to its established use for individual identification and parentage verification.

The data obtained substantially expand the existing knowledge of the genetic structure of the Kustanay horse breed and provide a basis for further studies. Their practical application for clarifying the origin and pedigree registration of individual animals requires further studies that include founder breeds and other genetically close groups, together with a combined analysis of molecular genetic data and reliable pedigree records.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ani16213453/s1, Table S1: Sampling locations of the studied horses with geographic coordinates; Table S2: Allele frequencies by studied groups; Table S3: MICRO-CHECKER results; Table S4: Close relatives removed from the dataset; Table S5: Hardy–Weinberg equilibrium tests for each locus in each group; Table S6: Linkage disequilibrium tests for all pairs of

loci in each group; Table S7: Comparison of the results obtained with 17 and 16 loci (n = 765); Table S8: Per-locus diversity parameters (Na, H_O, H_E, F) in the studied groups; Table S9: Overall F_{ST} and DAPC results in the 10 balanced subsamples; Table S10: Identifiers of the UB animals included in the 10 balanced subsamples; Table S11: STRUCTURE results for the final dataset (n = 703); Table S12: STRUCTURE results in the 10 balanced subsamples; Table S13: DAPC results for the final dataset (n = 703); Table S14: Anonymized individual multilocus genotypes of the 765 animals with group and sex; Figure S1: BIC values in the DAPC cluster selection; Figure S2: DAPC membership probabilities at K = 2–5; ARRIVE 2.0 checklist.

Author Contributions: Conceptualization, B.B.; methodology, A.D.; validation, B.B. and T.U.; formal analysis, B.B. and V.U.; investigation, A.D., S.K., V.R. and A.T.; resources, S.K., V.R. and A.T.; data curation, B.B.; writing—original draft preparation, B.B.; writing—review and editing, B.B., T.U., S.K., V.R. and A.T.; visualization, B.B. and V.U.; supervision, B.B.; funding acquisition, B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP25794029, "Genetic Diversity and the Influence of Genetic Polymorphisms of MSTN and SINE Inserts on the Athletic Performance of Horses of the Kostanay Breed". The APC was funded by Grant No. AP25794029.

Institutional Review Board Statement: The animal study protocol was approved by the Local Ethics Committee of Akhmet Baitursynuly Kostanay Regional University (Protocol No. 9, 30 September 2024; approval code IRB00014274). Hair follicle samples collected prior to the approval date had originally been obtained for routine pedigree registration and genetic identification purposes by authorized breeding inspectors and were subsequently analyzed as archived material under the approved protocol.

Informed Consent Statement: Biological material was collected by authorized breeding inspectors of the Republican Chamber for Pedigree Horse Breeds during routine zootechnical evaluation and pedigree registration procedures. Horse owners voluntarily presented their animals for evaluation and biological sample collection. The Chamber provided written authorization for laboratory analysis of the collected samples, use of the resulting data in scientific publications, and retrospective analysis of archived samples.

Data Availability Statement: The data supporting the findings of this study are provided in the article and its Supplementary Materials, including anonymized individual multilocus genotypes with group and sex (Table S14). Additional data are available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Kabylbekova, D.; Assanbayev, T.S.; Kassymbekova, S.; Kantanen, J. Genetic Studies and Breed Diversity of Kazakh Native Horses: A Comprehensive Review. *Adv. Life Sci.* **2024**, *11*, 18–27. [CrossRef]
2. 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]
3. Ualiyeva, D.; Dossybayev, K.; Kapassuly, T.; Kozhakhmet, A.; Kozhanov, Z.; Kryukov, K.; Arita, M.; Torekhanov, M.; Torekhanov, A. Mitochondrial DNA Reveals High Maternal Diversity within a Weak Breed Structure in Native Kazakhstani Horses. *Front. Genet.* **2026**, *17*, 1874969. [CrossRef]
4. Ulyanova, T.V.; Baymenov, B.M.; Mukanov, T.M.; Beishova, I.S.; Bugubaeva, A.U. Genetic Diversity of the Kostanay Horse Breed Using Microsatellite Markers. *Sci. Educ.* **2023**, *3*, 19–26. [CrossRef]
5. Tokayev Instructed to Establish a Horse Breeding Institute in Kazakhstan. (In Russian). Available online: https://vlast.kz/novosti/62667-tokaev-porucil-sozdat-v-kazahstane-institut-konevodstva.html (accessed on 20 October 2026).
6. How to Preserve the Endangered Kostanay Horse Breed. (In Russian). Available online: https://24.kz/ru/news/ekologiya/682947-kak-sokhranit-ischezayushchuyu-kostanajskuyu-porodu-loshadej (accessed on 20 October 2026).

7. Kovács, M.; Hegedűs, B.; Mihók, S.; Knop, R.; Szabó, C.; Posta, J. Population Structure and Genetic Diversity Among Lipizzan Mare Families in Hungary Based on Microsatellite Genotyping. *Animals* **2026**, *16*, 1062. [CrossRef]
8. Rogić, B.; Moravčíková, N.; Margeta, P.; Štrbac, L.; Zorc, M.; Kovács, M.; Vostry, L.; Kasarda, R.; Posta, J. Genetic Diversity and Population Structure in Seven Lipizzan Populations Based on Microsatellite Genotyping. *Animals* **2026**, *16*, 1516. [CrossRef]
9. Food and Agriculture Organization of the United Nations. Molecular Genetic Characterization of Animal Genetic Resources. In *FAO Animal Production and Health Guidelines*; FAO: Rome, Italy, 2011; ISBN 978-92-5-107032-1.
10. International Society for Animal Genetics. ISAG 2023 Workshop Report: Equine Genetics and Thoroughbred Parentage Testing Standardization Committee. Available online: https://www.isag.us/Docs/EquineGenParentage2023.pdf (accessed on 24 July 2026).
11. Vignal, A.; Milan, D.; SanCristobal, M.; Eggen, A. A Review on SNP and Other Types of Molecular Markers and Their Use in Animal Genetics. *Genet. Sel. Evol.* **2002**, *34*, 275. [CrossRef]
12. Schaefer, R.J.; McCue, M.E. Equine Genotyping Arrays. *Vet. Clin. N. Am. Equine Pract.* **2020**, *36*, 183–193. [CrossRef]
13. Beyshova, I.; Kikebayev, N.; Kokanov, S.; Chuzhebayeva, G.; Nametov, A. Genetic Polymorphism of 17 Microsatellite DNA Loci in Main Lines of Kostanay Breed Horses. *Life Sci. J.* **2013**, *10*, 85–91. [CrossRef]
14. Orazymbetova, Z.; Ualiyeva, D.; Dossybayev, K.; Torekhanov, A.; Sydykov, D.; Mussayeva, A.; Baktybayev, G. Genetic Diversity of Kazakhstani *Equus caballus* (Linnaeus, 1758) Horse Breeds Inferred from Microsatellite Markers. *Vet. Sci.* **2023**, *10*, 598. [CrossRef]
15. Kassymbekova, S.N.; Iskhan, K.Z.; Rzabaev, S.S.; Bimenova, Z.Z.; Kabylbekova, D.I.; Tursunkulov, S.A. Assessment of Genetic Diversity Using Microsatellite Markers and Milk Productivity of Mugalzhar Horses. *Her. Sci. S. Seifullin Kazakh Agrotech. Res. Univ. Vet. Sci.* **2024**, *3*, 29–36. [CrossRef]
16. 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]
17. Nguyen, T.B.; Paul, R.C.; Okuda, Y.; Le, T.N.A.; Pham, P.T.K.; Kaissar, K.J.; Kazhmurat, A.; Bibigul, S.; Bakhtin, M.; Kazymbet, P.; et al. Genetic Characterization of Kushum Horses in Kazakhstan Based on Haplotypes of mtDNA and Y Chromosome, and Genes Associated with Important Traits of the Horses. *J. Equine Sci.* **2020**, *31*, 35–43. [CrossRef]
18. Assanbayev, T.; Akilzhanov, R.; Sharapatov, T.; Bektayev, R.; Samatkyzy, D.; Karabayev, D.; Gabdulkayum, A.; Daniyarov, A.; Rakhimova, S.; Kozhamkulov, U.; et al. Whole Genome Sequencing and de Novo Genome Assembly of the Kazakh Native Horse Zhabe. *Front. Genet.* **2024**, *15*, 1466382. [CrossRef]
19. Van De Goor, L.H.P.; Panneman, H.; Van Haeringen, W.A. A Proposal for Standardization in Forensic Equine DNA Typing: Allele Nomenclature for 17 Equine-Specific STR Loci. *Anim. Genet.* **2010**, *41*, 122–127. [CrossRef]
20. Van Oosterhout, C.; Hutchinson, W.F.; Wills, D.P.M.; Shipley, P. Micro-Checker: Software for Identifying and Correcting Genotyping Errors in Microsatellite Data. *Mol. Ecol. Notes* **2004**, *4*, 535–538. [CrossRef]
21. Jones, O.R.; Wang, J. COLONY: A Program for Parentage and Sibship Inference from Multilocus Genotype Data. *Mol. Ecol. Resour.* **2010**, *10*, 551–555. [CrossRef]
22. Wang, J. Sibship Reconstruction From Genetic Data With Typing Errors. *Genetics* **2004**, *166*, 1963–1979. [CrossRef]
23. Wang, J.; Santure, A.W. Parentage and Sibship Inference from Multilocus Genotype Data under Polygamy. *Genetics* **2009**, *181*, 1579–1594. [CrossRef]
24. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2026. Available online: https://www.r-project.org/ (accessed on 19 October 2026).
25. Rousset, F. Genepop'007: A Complete Re-Implementation of the Genepop Software for Windows and Linux. *Mol. Ecol. Resour.* **2008**, *8*, 103–106. [CrossRef]
26. Guo, S.-W.; Thompson, E. Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. *Biometrics* **1992**, *48*, 361–372. [CrossRef]
27. Peakall, R.; Smouse, P.E. Genalex 6: Genetic Analysis in Excel. Population Genetic Software for Teaching and Research. *Mol. Ecol. Notes* **2006**, *6*, 288–295. [CrossRef]
28. Nei, M. Genetic Distance between Populations. *Am. Nat.* **1972**, *106*, 283–292. [CrossRef]
29. Szpiech, Z.A.; Jakobsson, M.; Rosenberg, N.A. ADZE: A Rarefaction Approach for Counting Alleles Private to Combinations of Populations. *Bioinformatics* **2008**, *24*, 2498–2504. [CrossRef]
30. Weir, B.S.; Cockerham, C.C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **1984**, *38*, 1358–1370. [CrossRef]
31. Goudet, J.; Jombart, T. Hierfstat: Estimation and Tests of Hierarchical F-Statistics. Available online: https://cran.r-project.org/web/packages/hierfstat/index.html (accessed on 19 October 2026).
32. Wright, S. Evolution in Mendelian Populations. *Genetics* **1931**, *16*, 97–159. [CrossRef]
33. Pritchard, J.K.; Stephens, M.; Donnelly, P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **2000**, *155*, 945–959. [CrossRef]

34. Li, Y.-L.; Liu, J.-X. StructureSelector: A Web-Based Software to Select and Visualize the Optimal Number of Clusters Using Multiple Methods. *Mol. Ecol. Resour.* **2018**, *18*, 176–177. [CrossRef]
35. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the Number of Clusters of Individuals Using the Software Structure: A Simulation Study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [CrossRef]
36. Puechmaille, S.J. The Program STRUCTURE Does Not Reliably Recover the Correct Population Structure When Sampling Is Uneven: Subsampling and New Estimators Alleviate the Problem. *Mol. Ecol. Resour.* **2016**, *16*, 608–627. [CrossRef]
37. Kopelman, N.M.; Mayzel, J.; Jakobsson, M.; Rosenberg, N.A.; Mayrose, I. Clumpak: A Program for Identifying Clustering Modes and Packaging Population Structure Inferences across K. *Mol. Ecol. Resour.* **2015**, *15*, 1179–1191. [CrossRef]
38. Jombart, T.; Devillard, S.; Balloux, F. Discriminant Analysis of Principal Components: A New Method for the Analysis of Genetically Structured Populations. *BMC Genet.* **2010**, *11*, 94. [CrossRef]
39. Jombart, T. Adegenet: A R Package for the Multivariate Analysis of Genetic Markers. *Bioinformatics* **2008**, *24*, 1403–1405. [CrossRef]
40. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*; Statistics and Computing; Springer: New York, NY, USA, 2002; ISBN 978-1-4419-3008-8.
41. Nikolaeva, E.A.; Voronkova, V.N.; Politova, M.A.; Ryabova, E.V.; Demin, V.A.; Stolpovsky, Y.A. Genetic Structure of Russian Riding Horse Breed. *Genetika* **2023**, *59*, 1048–1058. [CrossRef]
42. Fornal, A.; Radko, A.; Piestrzyńska-Kajtoch, A. Genetic Polymorphism of Hucul Horse Population Based on 17 Microsatellite Loci. *Acta Biochim. Pol.* **2013**, *60*, 761–765.
43. Ustyantseva, A.V.; Khrabrova, L.A.; Abramova, N.V.; Ryabova, T.N. Genetic Characterization of Akhal-Teke Horse Subpopulations Using 17 Microsatellite Loci. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *341*, 012070. [CrossRef]
44. Kalashnikov, V.; Khrabrova, L.; Blohina, N.; Zaitcev, A.; Kalashnikova, T. Dynamics of the Inbreeding Coefficient and Homozygosity in Thoroughbred Horses in Russia. *Animals* **2020**, *10*, 1217. [CrossRef]
45. Jlassi, M.; Dhifalli, I.; Ouled Ahmed, H.; Lasfar, F.; El Gtari, M.; Jemmali, B. Genetic Characterization of the Arabian Horse Population in Tunisia Using Microsatellites. *Life* **2025**, *15*, 1925. [CrossRef]
46. Wright, S. *Evolution and the Genetics of Populations, Volume 4: Variability Within and Among Natural Populations*; University of Chicago Press: Chicago, IL, USA, 1978; ISBN 978-0-226-91041-3.
47. Cosgrove, E.J.; Sadeghi, R.; Schlamp, F.; Holl, H.M.; Moradi-Shahrbabak, M.; Miraei-Ashtiani, S.R.; Abdalla, S.; Shykind, B.; Troedsson, M.; Stefaniuk-Szmukier, M.; et al. Genome Diversity and the Origin of the Arabian Horse. *Sci. Rep.* **2020**, *10*, 9702. [CrossRef]
48. Janes, J.; Miller, J.; Dupuis, J.; Malenfant, R.; Gorrell, J.; Cullingham, C.; Andrew, R. The K=2 Conundrum. *Mol. Ecol.* **2017**, *26*, 3594–3602. [CrossRef]
49. Hou, L.; Sulayman, A.; Zeng, Y.; Zhou, L.; Aimaier, A.; Kader, A.; Shi, L. Analysis of Genetic Diversity and Race Genetic Structure of Major Horse Breeds in Xinjiang, China. *Animals* **2025**, *15*, 2690. [CrossRef]
50. Giontella, A.; Sarti, F.M.; Cardinali, I.; Giovannini, S.; Cherchi, R.; Lancioni, H.; Silvestrelli, M.; Pieramati, C. Genetic Variability and Population Structure in the Sardinian Anglo-Arab Horse. *Animals* **2020**, *10*, 1018. [CrossRef]
51. Perdomo-González, D.I.; Sánchez-Guerrero, M.J.; Molina, A.; Valera, M. Genetic Structure Analysis of the Pura Raza Español Horse Population through Partial Inbreeding Coefficient Estimation. *Animals* **2020**, *10*, 1360. [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.