BMC Genomics

RESEARCH Open Access

Genomic diversity and structure in arabian horses revealed by whole-genome sequencing: establishment of an allele frequency database of common genetic variation

Tomasz Szmatoła[1,2], Carrie Finno[3], Artur Gurgul[1], Harrison Heath[4], Monika Stefaniuk-Szmukier[2], Saria Almarzook[5], Elaine Norton[6] and Katarzyna Ropka-Molik[2*]

Abstract

Background The Arabian horse is a culturally and historically influential breed that has contributed to the development of many modern horse populations. However, genomic resources for this breed remain limited, particularly population-level allele frequency datasets that support studies of genetic diversity, selection, and disease. The aim of this study was to generate a comprehensive allele frequency database for Arabian horses using representative sampling and high-coverage whole-genome sequencing.

Results We generated the first publicly available allele frequency database for the Arabian horse by sequencing pooled genomic DNA from 120 individuals originating from four distinct populations: Poland, the United States, Egypt, and Syria. Across the dataset, 11.7 million single nucleotide polymorphisms were identified, and both population-specific and global allele frequencies were estimated. Genome-wide analyses revealed substantial variation in polymorphism density, as well as multiple regions exhibiting fixation for alternate alleles. Principal component analysis, hierarchical clustering, and admixture analysis demonstrated clear genetic differentiation among the four populations. We detected 6,210 fixed genomic blocks, many of which overlapped with genes involved in metabolic and signaling pathways, suggesting potential historical selection. One such block corresponded to a region previously associated with insect bite hypersensitivity. To assess clinical relevance, we screened the dataset for known pathogenic variants linked to four inherited disorders described in Arabian horses. None of these variants were detected, and their absence was supported by high sequencing depth and confirmed through Sanger sequencing.

Conclusions This study provides the first comprehensive allele frequency resource for the Arabian horse. The database enables robust variant filtering, facilitates the identification of candidate functional polymorphisms, and supports studies of genetic structure and selection. It represents an important foundation for future work in equine

*Correspondence:
Katarzyna Ropka-Molik
katarzyna.ropka@iz.edu.pl

Full list of author information is available at the end of the article

genomics and has practical applications in breeding, conservation genetics, and health management. Continued expansion of the dataset with additional populations will further enhance its value for research and applied genomics.

Keywords Pooled sequencing, Single nucleotide polymorphism (SNP), Population structure, Equine genetics

Background

The Arabian horse is one of the world's most iconic and influential breeds, renowned for its endurance, versatility, and characteristic phenotype. As a foundation breed for many modern horse populations, the Arabian has been subject to centuries of selection, migration, and population subdivision, resulting in a complex genetic landscape shaped by both historical and contemporary forces [1–3]. Despite its cultural and economic importance, the Arabian horse still lacks the comprehensive genomic resources, such as high-quality reference assemblies, population-scale variant databases, and breed-specific functional annotations, that are now available for many other livestock and companion animal species [4].

One critical gap has been the absence of a robust, population-stratified database of genetic variant frequencies for the Arabian horse. Such databases, which are analogous to resources like gnomAD [5] for humans, are indispensable for both research and clinical genomics. They provide essential context for interpreting genetic variation, distinguishing between common polymorphisms and potentially deleterious or disease-causing mutations, and filtering candidate variants in studies of Mendelian and complex traits [5, 6]. In horses, the lack of allele frequency reference data can hinder the identification of causal variants underlying rare diseases, impede the interpretation of genetic tests, and limit the ability to assess breed-specific risk alleles [7].

The development of an allele frequency database is particularly relevant for Arabian horses. The breed is affected by several inherited diseases and possesses unique traits of interest to both breeders and the scientific community [8]. However, population structure, recent bottlenecks, and the geographic dispersal of subpopulations complicate the interpretation of genetic data and the identification of private or founder mutations [9–11]. Establishing a comprehensive, publicly available resource of population-specific allele frequencies will facilitate more accurate genetic diagnosis, improve disease variant filtering, and enable deeper insight into the breed's evolutionary and selection history [1, 4].

In designing this study, we aimed to establish a comprehensive allele frequency resource that would be broadly applicable to genetic research and clinical genomics in the Arabian horse. To maximize global relevance, we selected four geographically and historically distinct populations from — Poland (PL), United States (US), Egypt (EGY), and Syria (SYR) — as representatives of major contemporary and ancestral genetic lineages within the breed [1, 2, 9, 10]. The Polish and US populations, each represented by multiple sample pools corresponding to established breeding lines, reflect regions with large, genetically diverse Arabian horse populations and structured breeding programs that have shaped global Arabian horse genetics [2, 10]. In contrast, the Egyptian and Syrian Arabians, each represented by a single pool due to logistical constraints, were included to anchor the database with Middle Eastern source lineages. Although the sample size for these groups is limited, their representation is crucial for capturing ancestral diversity and providing a genetic reference point for lineages that have historically contributed to the global Arabian horse gene pool [1, 9]. This population-stratified approach enables the resulting allele frequency database to function as an effective filtering tool for the identification and interpretation of common genetic variants across a broad spectrum of Arabian horses, including those from underrepresented or mixed ancestry backgrounds. By making population-specific frequency data publicly available, our study provides an essential resource for researchers and clinicians seeking to distinguish rare or private variants from common population polymorphisms, and to improve the accuracy of genetic testing and disease variant discovery in the Arabian breed [5, 7].

In this study, we address this need by generating a large-scale, genome-wide catalog of common genetic variation and allele frequencies across 120 Arabian horses originating from four major geographic regions. Through pooled whole-genome sequencing and downstream population genetic analyses, we provide the first, as far as we know, publicly accessible variant frequency database for the breed, alongside a detailed characterization of genetic structure, regions of fixation, and candidate loci relevant to health and performance.

Methods

Sample collection and pooling strategy

A total of 120 purebred Arabian horses were sampled, comprising 12 pools each representing a unique population or subpopulation of 10 horses from four distinct geographic origins: Egypt, Syria, Poland, and the United States. Polish Arabian horses were selected from major state studs, including Janów Podlaski (4 pools) and Michałów (3 pools), which are the principal breeding centers in Poland. United States Arabian horses were sampled across four US states (Alabama, Arizona, Florida and Wisconsin), representing the diverse genetic backgrounds present in the US population. Egyptian Arabians

were obtained from northern suburb of Cairo, while Syrian Arabians were sourced from Damascus, Hama and Al-Hasakah Governorates. Each pool consisted of 10 individuals selected to maximize genetic representation within each group. Pools included both male and female individuals, although the sex distribution differed between populations. All Polish Arabian horses (7 pools) were females. The Egyptian and Syrian pools each consisted of 6 females and 4 males, while the United States population included 24 mares and 6 geldings across three pools. Due to sample availability, the sex ratio was therefore not balanced across all pools. While autosomal analyses are not expected to be affected, this may influence analyses involving the X chromosome, and therefore results for sex chromosomes should be interpreted with caution. Hair follicle samples were collected using standard veterinary procedures with owner consent, and genomic DNA was extracted using the Sherlock AX kit (A&A Biotechnology, Gdansk, Poland) for Poland, Egyptian and Syrian populations and using whole blood from the Qiagen Puregene Blood Kit (Qiagen, Aarhua, Denmark) for US population according to the manufacturer's instructions. DNA quality and concentration were assessed using NanoDrop 2000, and equal amounts of DNA from each individual within a pool were combined to create equimolar pooled DNA samples. Detailed pedigree information was not available for all sampled individuals. However, efforts were made to select horses representing diverse breeding lines within each population. It should be noted that shared ancestry, particularly within structured breeding populations such as Polish Arabians, may influence observed genetic patterns.

Library preparation and whole-genome sequencing

Library preparation was commercially performed, and pooled DNA samples were sequenced on the NovaSeq S4 platform generating paired-end reads of 150 bp in length. Each pool was sequenced to an average depth of 62.6×, yielding an average of 635.18 million (M) raw reads per pool.

Read processing and mapping

Sequencing quality was evaluated using FastQC v0.12.1 [12]. Adapter sequences and low-quality bases were trimmed using Flexbar v3.5 [13] with default parameters, retaining reads with a minimum length of 36 bp and base quality above 20. Quality-filtered reads were mapped to the soft masked Equus caballus reference genome (EquCab3.0) using BWA-MEM v2.2 [14] with default settings. Alignment files were sorted, indexed and corresponding read group tags were added using SAMtools v1.21 [15]. Reads with mapping quality below 30 were discarded. Duplicate reads were identified and marked using Picard tools v3.4 (http://broadinstitute.g.ithub.io/picard/). Mapping statistics, including coverage and m apping rate, were computed using Samtools flagstat and idxstats options. The EquCab3.0 reference genome was selected due to its comprehensive annotation and widespread use in equine genomics, ensuring compatibility with existing resources and tools.

Variant calling and filtering

Variant calling was performed using GATK v4.6.2.0 [16], employing settings optimized for pooled sequencing data. For each pool, the HaplotypeCaller was executed in GVCF mode with the --sample-ploidy parameter set to 20, corresponding to the pooled DNA of 10 diploid individuals. This approach models the pooled data as a 20-ploid population, allowing approximate inference of allele frequencies from the read data. Joint analysis across all pools was performed using GenotypeGVCFs. To ensure high-confidence variant detection, we retained only biallelic SNPs with a minimum site depth of 20 per pool, base quality ≥ 30, and mapping quality ≥ 30. Loci with more than 5% missing data across pools were excluded from further analysis.

To minimize potential artifacts, variants located within repetitive or low-complexity regions, as annotated by RepeatMasker [17], were removed. The final filtered SNP set was used for downstream allele frequency estimation and population genetic analyses.

No explicit minor allele frequency (MAF) filtering was applied at this stage, as the primary objective of the study was to construct a comprehensive allele frequency database capturing the full spectrum of genetic variation. However, we acknowledge that inclusion of rare variants may introduce noise in downstream population structure analyses.

Estimation of allele frequencies

Allele frequencies at each SNP site were estimated from read counts in each pool using a combination of GATK's allele frequency annotations, bcftools query, and custom Python scripts. For each variant, population-specific allele frequencies were calculated for the PL, US, EGY, and SYR pools, as well as combined allele frequencies across all samples. Variant annotation was performed using the Ensembl GTF file for EquCab3.0 (release 114) with a custom Python script, allowing genomic features such as gene IDs, gene symbols, and functional categories (e.g., intronic, exonic, intergenic) to be associated with each variant. To accommodate user preferences and ensure long-term usability of the dataset, we provide two versions of the allele frequency database: one with annotations, and one without. This separation addresses concerns about large file size and allows future users to re-annotate variants as gene models are updated. Additionally, we provide a standalone Python

script that allows users to query the database for specific variant positions, as well as to retrieve variants located in close proximity to a given coordinate. This functionality enables rapid lookups of known or candidate sites and facilitates integration of the database into downstream genomic workflows. Detailed information for each variant, including genomic position, genotype, and allele frequencies, is compiled into a BED-format database and deposited on Zenodo (https://doi.org/10.5281/zenodo.17483784).

The presence or absence of known pathogenic variants associated with inherited diseases in Arabian horses was assessed by comparing reported positions (lifted over to EquCab3 coordinates using the UCSC LiftOver tool) to entries in the final allele frequency database while also analyzing the results found in OMIA database. The diseases of interest were as follows: Lavender Foal Syndrome (LFS; *MYO5A*), Severe Combined Immunodeficiency (SCID; *PRKDC*), Cerebellar Abiotrophy (CA; *TOE1*/*MUTYH locus*), and Occipitoatlantoaxial Malformation (OAAM1; *HOXD3*/*HOXD4* intergenic deletion).

Genome-wide distribution and density of SNPs

Variant distribution and density were assessed by calculating the number of SNPs per chromosome and normalizing by chromosome length (variants/Mb). Additional analysis was performed by partitioning the genome into non-overlapping 1 Mb windows and calculating the number of SNPs per window. All computations and visualizations were carried out in R v4.4 [18] using the ggplot2, data.table, dplyr and viridis packages.

Population structure analysis

Population genetic structure was examined using principal component analysis (PCA), hierarchical clustering, and admixture analysis. PCA was performed on allele frequency data using prcomp R package [18], and the resulting principal components were visualized in R with ggplot package. Hierarchical clustering was conducted on pairwise genetic distance matrices calculated using Nei's distance. Neighbor-joining trees were constructed and visualized using the ape package (v5.8; [19]) in R, allowing for graphical representation of the genetic relationships among the horse populations.

For admixture analysis, individual genotypes were not directly available due to the pooled sequencing design. Instead, allele counts for each SNP in each pool were used to approximate genotype information. Each pool was treated as a single "pseudo-individual," representing the aggregated allele frequencies of 10 diploid individuals (ploidy = 20). For each SNP, the number of reference and alternate alleles was derived from read counts and normalized according to sequencing depth.

These allele counts were converted into genotype-like input suitable for ADMIXTURE v1.3.0 [20] by scaling allele frequencies to the expected ploidy and discretizing them into integer allele counts. This approach enables the use of ADMIXTURE in a pooled sequencing context, although it represents an approximation of true individual-level genotype data.

Analyses were performed for K values ranging from 2 to 10, with cross-validation (--cv) applied to determine the optimal number of clusters. Visualization of admixture proportions was carried out in R using ggplot2.While methods such as NGSAdmix are specifically designed to incorporate genotype likelihoods from next-generation sequencing data, the relatively high sequencing depth achieved in this study (~ 60× per pool) allowed for reliable estimation of allele frequencies from read counts. Therefore, we applied an allele count–based approach using ADMIXTURE for exploratory analysis of population structure. Nevertheless, this approach has limitations compared to genotype likelihood–based methods and should be interpreted accordingly. However, given the high sequencing depth and the focus on population-level allele frequency patterns, this approach provides a reasonable approximation for exploratory population structure analysis.

Analysis of genome-wide allele frequency patterns

Mean alternate allele frequency was computed in non-overlapping 10 kb windows for all autosomes and the X chromosome. Population-specific allele frequency heatmaps were generated in 1 Mb windows for each population and chromosome, visualized using R and the ggplot package. This analysis was conducted to identify genomic regions that exhibit unusually high or low allele frequencies, which may indicate areas influenced by selection, demographic history, or structural constraints, and to highlight patterns of population-specific variation across the genome. Analyses including the X chromosome were conducted for completeness; however, interpretation of allele frequency patterns on sex chromosomes should consider potential biases related to sex composition of pooled samples.

Identification of fixed alternate allele blocks

Genomic blocks fixed for the alternate allele (allele frequency = 1) were identified by scanning for regions containing at least four consecutive fixed SNPs and spanning a minimum length of 400 base pairs. Blocks were defined as contiguous stretches where the alternate allele frequency remained at 1 in the combined dataset. Block sizes, variant counts per block, and chromosomal distribution were computed using custom scripts in R. Genes overlapped by fixed blocks were identified by intersecting block coordinates with gene annotation files (Ensembl

EquCab3.0 release 114). This analysis aimed to detect candidate regions potentially shaped by historical or artificial selection, population bottlenecks, or drift, and to explore their possible functional relevance by identifying genes consistently fixed for alternate alleles across populations.

Genes overlapping fixed blocks were subjected to KEGG pathway enrichment analysis using KOBAS-i [21], with statistical significance assessed by the Fisher's exact test and adjusted for multiple testing using the Benjamini-Hochberg correction (0.05 threshold). Coordinates of all equine QTLs were downloaded from the Horse QTL Database (https://www.animalgenome.org/QTLdb/horse). Overlap between fixed alternate allele blocks and QTL regions was assessed using BEDTools v2.31.1 [22].

Data availability

All raw sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA) under accession numbers PRJNA1332950. The allele frequency databases with and without annotation and the script that can be used as a search tool of specific genetic variants are available at Zenodo (https://doi.org/10.5281/zenodo.17483785).

Results

Whole-genome sequencing and variant calling

Pooled whole-genome sequencing (WGS) was conducted on 12 pools of Arabian horses (120 horses in total). Sequencing generated an average of 635 million reads per pool (ranging from 520 to 690 million reads). After quality trimming with Flexbar, between 518 and 675 million high-quality reads per pool were retained. The trimmed reads were aligned to the EquCab3.0 reference genome, with an average of 87.7% of read pairs mapping uniquely across all pools. Duplicate reads were marked using Picard tools, and variant calling was performed with GATK, using settings optimized for pooled sequencing data. Sequencing coverage was consistent across all pools at average coverage depth of 61.6, supporting accurate estimation of allele frequencies. To assess the completeness and reliability of the variant catalog, per-base coverage was computed across the genome for each pool. Regions with insufficient depth (< 20 reads per pool) were excluded from downstream allele frequency estimation, and only variants meeting this minimum coverage threshold were retained. This filtering ensures that variants absent from the database are unlikely to be due to lack of coverage. A genome-wide coverage summary file, indicating mean depth per base, is available alongside the frequency database to enable users to distinguish between true absence of a variant and a lack of data at that position. Additional sequencing metrics are summarized in Table 1.

Variant discovery and allele frequency estimation

After variant calling and filtering, a total of 11,752,974 SNPs were identified across the analyzed samples. The alternate allele frequency (AF) values ranged from 0.0042 to 1, with a mean ± standard deviation of 0.23 ± 0.26 and a median of 0.12. Moreover, 65,269 variants were found to be fixed for the alternate allele (alternate allele frequency = 1).

Example of database output: part of a GHR gene region

To illustrate the structure and utility of the generated allele frequency database, Table 2 presents a sample excerpt of the data corresponding to a segment of the *GHR* gene (*Growth Hormone Receptor*), one of the candidate *loci* identified in our analysis. For each single nucleotide polymorphism (SNP), this table displays the chromosomal position, reference and alternate alleles, genotype, and allele frequencies calculated for each

Table 1 Sequencing and mapping statistics for Arabian horse DNA pools

Sample pool	Raw reads	Filtered reads	Percent of pruned reads	Properly mapped pairs	Percent of properly mapped pairs	Depth of coverage (x)
1PL	613,648,328	600,464,150	3	522,265,768	86.98	62.5
2PL	593,679,254	580,902,680	3	514,444,334	88.56	58.5
3PL	555,471,740	543,384,464	3	481,763,356	88.66	56.0
4PL	559,416,860	547,206,786	3	460,814,126	84.21	60.3
5PL	587,964,986	575,219,738	3	496,173,568	86.26	74.4
6PL	690,149,938	675,304,084	3	611,640,188	90.57	66.1
7PL	626,107,866	625,096,088	1	544,259,902	87.07	60.7
8US	564,514,060	563,122,054	1	499,939,862	88.78	63.8
9US	595,657,384	594,171,318	1	524,763,464	88.32	58.9
10US	576,720,290	575,273,170	2	484,714,030	84.26	55.5
11SYR	519,810,810	518,497,868	2	456,893,904	88.12	69.8
12EGY	639,061,754	637,452,242	1	574,732,194	90.16	62.5
Mean (All)	597,174,421	584,763,360	2.17	508,446,937	87.43	61.58

Table 2 Example output from the Arabian horse allele frequency database, showing variant data for a segment of the *GHR* gene

CHR	POS (EquCab3.0.)	REF	ALT	AF_ALL	AF_PL	AF_US	AF_EGY	AF_SYR
21	24,856,009	A	T	0.45	0.4	0.5	0.5	0.5
21	24,856,024	G	A	0.01	0	0	0.5	0
21	24,856,467	T	C	0.07	0.21	0	0	0.5
21	24,857,028	A	G	0.05	0.21	0	0	0
21	24,857,665	G	A	0.5	0.5	0.5	0.5	0.5
21	24,857,954	G	T	0.01	0.07	0	0	0
21	24,857,982	T	A	0.04	0.21	0	0	0
21	24,858,128	T	C	1	1	1	1	1
21	24,858,250	T	G	1	1	1	1	1
21	24,858,494	T	C	1	1	1	1	1
21	24,858,888	G	A	1	1	1	1	1
21	24,859,551	C	T	0.08	0.28	0	0	0
21	24,860,039	A	G	0.08	0.28	0.5	0	0
21	24,860,054	C	A	0.07	0.21	0.5	0	0
21	24,860,824	G	C	0.62	0.62	0.71	0.5	0.5
21	24,860,872	A	G	0.1	0.21	0.33	0	0.5
21	24,860,958	A	G	0.28	0.42	0.5	0.5	0
21	24,861,015	T	C	0.01	0	0.16	0.5	0
21	24,861,095	C	T	0.07	0.14	0.33	0	0.5
21	24,861,231	G	A	0.02	0	0	0.5	0
21	24,861,415	C	T	0.46	0.44	0.5	0.5	0.5
21	24,861,483	A	G	0.01	0.07	0	0	0

CHR Chromosome number, *POS* Genomic position (base-pair coordinate), *REF* Reference allele, *ALT* Alternate allele, *AF_ALL* Alternate allele frequency across all populations, *AF_PL* Alternate allele frequency in the Polish population, *AF_US* Alternate allele frequency in the United States population, *AF_EGY* Alternate allele frequency in the Egyptian population, *AF_SYR* Alternate allele frequency in the Syrian population, *GHR* Growth Hormone Receptor

Table 3 Major known disease-associated variants (OMIA database) in Arabian horses and their presence in the current allele frequency database

Disease	Gene	OMIA ID	Variant (Description)	Position (EquCab2)	Position (EquCab3)*	Reference	Found in Our Database?
Lavender Foal Syndrome (LFS)	*MYO5A*	OMIA:001501	c.1462delG (frameshift)	1:119,903,709	1: g.139,290,592del (OMIA variant table; EquCab3.0)	Brooks et al., [7]	No
Severe Combined Immunodeficiency (SCID)	*PRKDC*	OMIA:000220	5 bp del (c.4675_4679del)	9:50,511,072–76	9: g.36,395,752_36,395,756del (EquCab3.0)	Finno et al., [24]	No
Cerebellar Abiotrophy (CA)	*TOE1/MUTYH* region	OMIA:000175	c.218Tdel (frameshift)	2:132,222,294	2: g.13,122,415 C > T (EquCab3.0/3.1)	Scott at al., [25]	No
Occipitoatlantoaxial Malformation (OAAM)	*HOXD3/HOXD4* intergenic	OMIA:000081	upstream deletion (regulatory)	7:66,143,664–65	7:68,325,817 – 68,325,818	Finno et al., [26]	No

*Positions acquired with the use of UCSC Liftover tool or from OMIA database

population (from Poland, US, Egypt, Syria), as well as the overall frequency. This format enables researchers to readily filter variants based on population-specific or global allele frequencies, facilitating the identification of common, rare, private, or potentially functional alleles within the breed.

Presence of known disease-associated variants

To evaluate the clinical utility of the database, we searched for several well-characterized pathogenic variants known to cause inherited diseases in Arabian horses (Table 3). None of the reported pathogenic alleles for Lavender Foal Syndrome (LFS), Severe Combined Immunodeficiency (SCID), Occipitoatlantoaxial Malformation (OAAM), or Cerebellar Abiotrophy (CA) were detected in the current cohort. Importantly, we confirmed that the genomic positions corresponding to these variants were adequately covered in our data, with all sites exhibiting a sequencing depth exceeding 20 reads. This ensures that the absence of these pathogenic alleles reflects a true negative result rather than insufficient coverage.

Furthermore, the absence of pathogenic alleles associated with SCID, CA, and LFS was independently validated through Sanger sequencing for PL, EGY and SRY

samples. This validation followed the methodology described by Ayad et al. [23], and was conducted for each individual horse and each specific disease-associated variant. The Sanger sequencing results confirmed the allele frequencies observed in our dataset, thereby reinforcing the reliability of the negative findings derived from whole-genome sequencing.

Genome-wide distribution and density of SNPs

To investigate the distribution and density of genetic variation across the Arabian horse genome, we analyzed the number and density of SNPs detected in our dataset. The total number of SNPs identified per chromosome varied widely (Fig. 1A), largely reflecting differences in chromosome length. Chromosome 1 harbored the highest number of variants, with over 800,000 SNPs, followed by the X chromosome and several other large autosomes. In contrast, smaller chromosomes exhibited substantially fewer variants. To account for differences in chromosome size, we calculated SNP density as the number of variants per megabase (Mb) for each chromosome (Fig. 1B). This normalization revealed that certain chromosomes, such as chromosome 12 and chromosome 20, were particularly enriched for SNPs relative to their size, reaching densities above 8,000 SNPs per Mb. Other chromosomes exhibited more moderate densities, with the lowest densities observed on the smallest chromosomes.

A more detailed view of SNPs distribution was obtained by plotting the number of variants within non-overlapping 1 Mb windows across the entire genome (Fig. 1C). This analysis revealed considerable heterogeneity in variant density along and between chromosomes. While most regions contained between 2,000 and 6,000 SNPs per window, several genomic intervals exhibited pronounced peaks, indicating potential local hotspots of genetic diversity.

Population structure and genetic relationships

The genetic relationships among the 12 Arabian horse pools from four origins (PL, US, SYR, EGY) were assessed using principal component analysis (PCA), hierarchical clustering of genetic distances, and admixture analysis (Fig. 2).

PCA separated the four origins into distinct clusters. The first principal component (PC1) distinguished the SYR pool from the PL, US and EGY horses, while PC2 further separated EGY from SYR. The PL and US

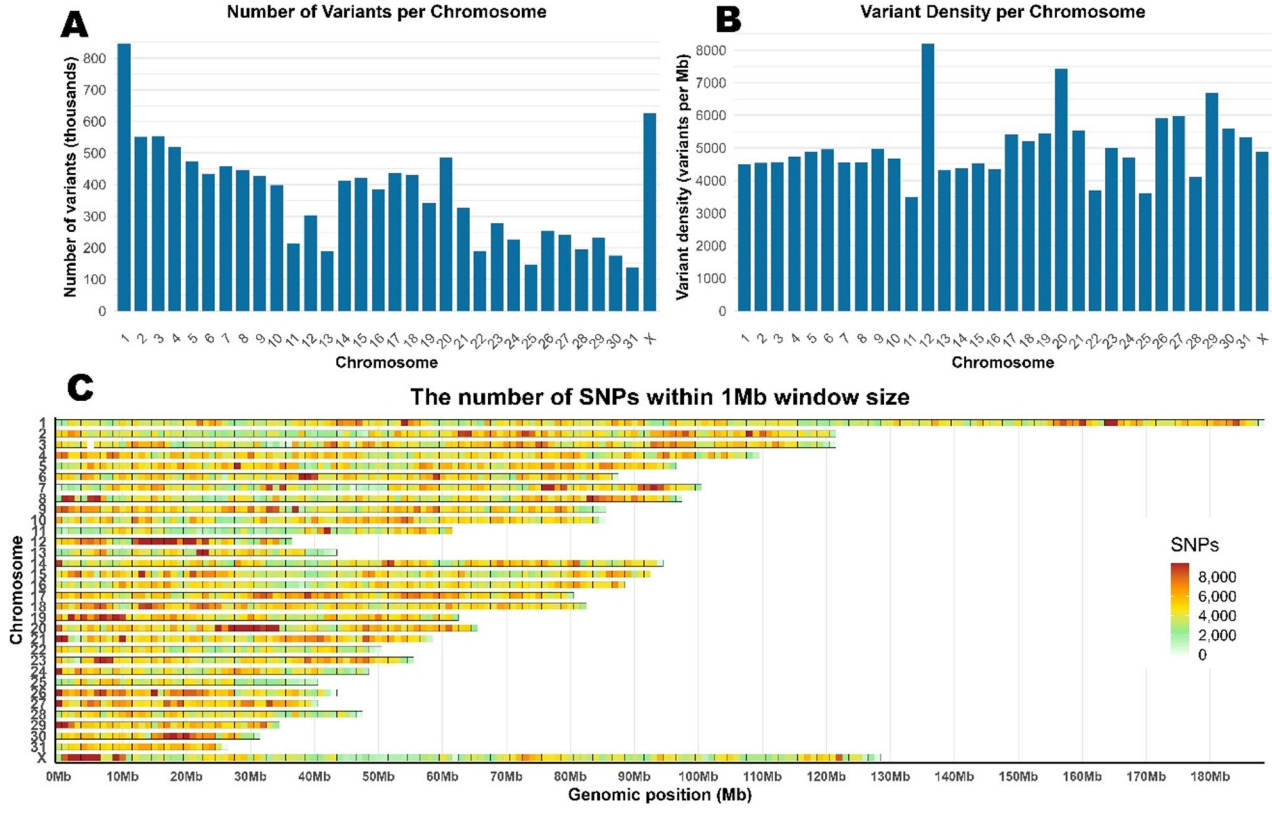

Fig. 1 Genome-wide distribution and density of SNP variants across the Arabian horse genome. **A**Total number of variants identified on each chromosome, shown in thousands. **B** Variant density, calculated as the number of variants per Mb for each chromosome. **C** Genomic distribution of SNPs in 1 Mb windows. Heatmap showing the number of SNPs within non-overlapping 1 Mb windows across all chromosomes. Color scale reflects the number of SNPs per window.

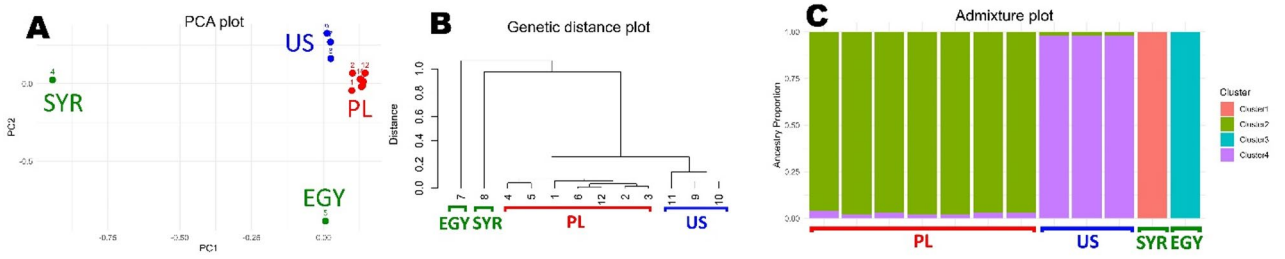

Fig. 2 Genetic structure and ancestry of Arabian horse populations. **A** Principal component analysis (PCA) of allele frequencies showing genetic clustering by origin. **B** Genetic distance dendrogram grouping pools by relatedness. **C** Admixture plot illustrating ancestry proportions for each population

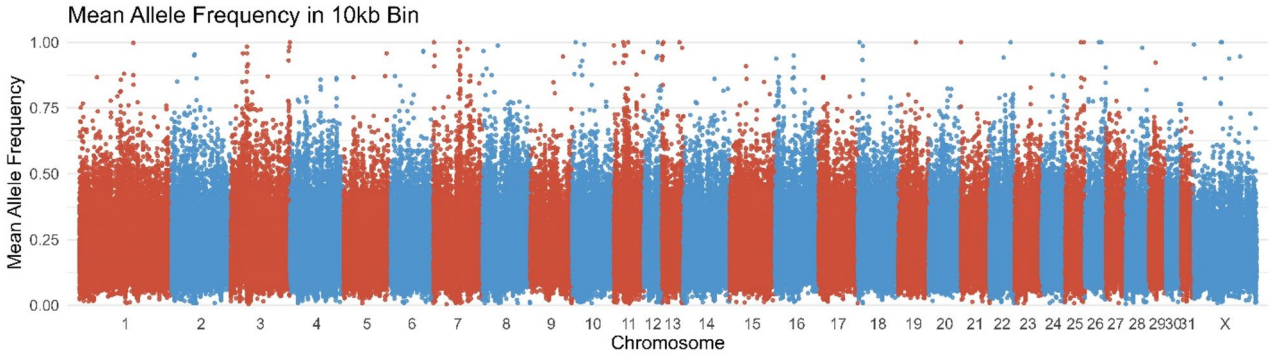

Fig. 3 Genome-wide distribution of mean (all pools) allele frequency in 10 kb bins across all chromosomes

pools formed tight, well-defined clusters, reflecting clear genetic differentiation between these populations (Fig. 2A). The hierarchical clustering based on pairwise genetic distances confirmed the PCA results, grouping pools by origin. EGY and SYR pools clustered closely together, distinct from the PL and US groups, which also formed separate branches. Within the PL and US clusters, pools grouped tightly, suggesting genetic homogeneity within each population (Fig. 2B). For the admixture analysis, cross-validation error indicated that K = 4 was the optimal number of ancestral clusters, and this value was selected for detailed examination. The results revealed a high degree of ancestry homogeneity within each geographic origin. Polish pools were characterized by a single predominant ancestry component (Cluster 2), while the US pools were dominated by Cluster 4. Only minimal admixture was observed in both the Polish and US pools. In contrast, the Syrian and Egyptian pools each formed distinct and separate clusters, with no evidence of admixture between these groups (Fig. 2C).

Genome-wide distribution of allele frequencies

To characterize genome-wide patterns of genetic diversity in the Arabian horse pools, we calculated the mean alternate allele frequency in non-overlapping 10 kb windows across all autosomes and the X chromosome (Fig. 3, Supplementary Material 1). The resulting Manhattan-style plot revealed substantial heterogeneity in allele frequencies along and between chromosomes. Most chromosomes showed a broad range of mean AF values, with several regions displaying notably high frequencies. In particular, chromosomes 3, 7 and 11 exhibited clusters of windows with elevated mean allele frequency.

Population-specific allele frequency patterns across the genome

To explore differences in genetic variation among the four Arabian horse populations, we generated a heatmap of mean alternate allele frequency in non-overlapping 1 Mb bins across all chromosomes (Fig. 4). Each facet corresponds to a chromosome and, within each chromosome, the distribution of allele frequencies is visualized for each population. We chose 1 Mb binning to improve visual clarity and interpretability of genome-wide trends by reducing noise from individual variants and enabling easier detection of population-specific patterns.

The heatmap reveals considerable heterogeneity in allele frequency patterns both along and between chromosomes. For most chromosomes, the overall mean AF values appear broadly similar among populations, suggesting a shared genetic background and limited genome-wide differentiation. However, certain chromosomes (e.g., 12, 20, X) and specific genomic regions display population-specific shifts in mean AF, indicating localized differences in genetic diversity or possible signatures of selection. In particular, several bins on chromosomes 12 and X show lower mean AF values in the PL and US populations compared to EGY and SYR, while

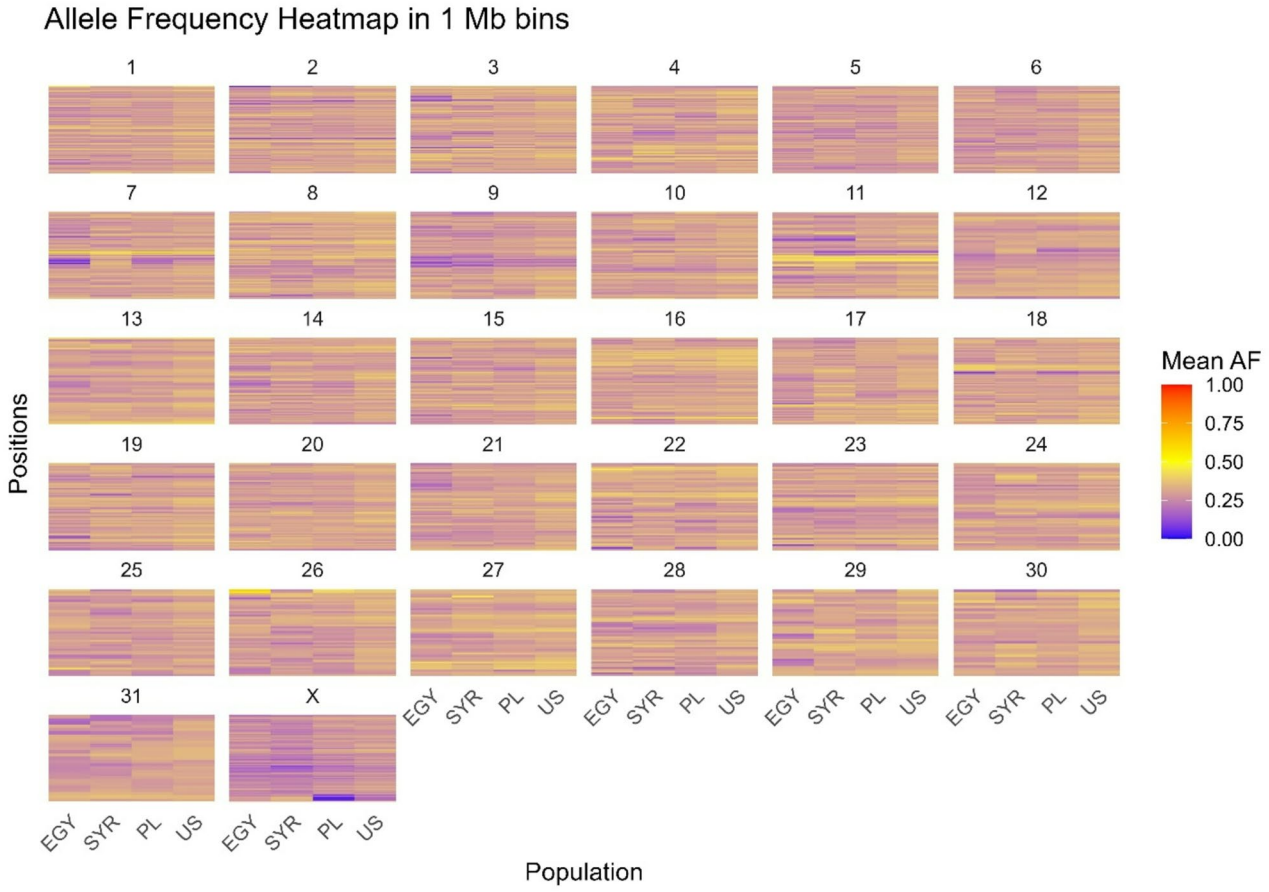

Fig. 4 Population allele frequency heatmap in 1 Mb bins across Arabian horse chromosomes

chromosomes 20 and 26 contain regions where the EGY population exhibits relatively higher allele frequencies.

Blocks of fixed alternate alleles across the genome

To identify regions of putative fixation, we searched for genomic blocks where the alternate allele frequency reached 1. Only blocks spanning at least four consecutive variants were considered, ensuring that isolated singleton SNPs were excluded from this analysis.

In total, 6,210 genomic blocks fixed for the alternate allele (allele frequency = 1.0) were identified across the genome based on the combined allele frequency data from all 12 pools. These regions represent contiguous stretches where the alternate allele was consistently present in all individuals, suggesting fixation at the population level rather than in a single sample. The mean block size was 700 bp, with the largest extending nearly 8.1 kb. On average, each block consisted of 5.8 variants, with the longest containing 70 variants. Detailed information regarding the positions, sizes, and variant counts of these blocks is provided in Supplementary Material 2.

The chromosomal distribution of fixed blocks was highly uneven (Fig. 5A). Chromosome 3 harbored the largest number of blocks, followed by chromosomes 4, 8, and 1, whereas chromosomes 30, 31, 28, and 21 had the fewest. The X chromosome also contained a moderate number of fixed regions. To investigate the potential functional relevance of these blocks, we identified genes that were most frequently overlapped. Among these, *SLITRK3*, *BCHE*, *SI*, and *ZKSCAN2* stood out, each intersected by over 60 fixed blocks (Fig. 5B; details in Supplementary Material 2).

Functional analysis of genes overlapping with fixed blocks

To investigate the biological processes associated with regions of fixation, we performed KEGG pathway enrichment analysis with the use of Kobas-i software on genes overlapping blocks with allele frequency equal to 1. Several pathways were significantly enriched among these genes (Fig. 6).

The most prominent enrichment was observed for the Metabolic pathways category, with over 60 genes in fixed regions mapping to this pathway. Additionally, both the Ras signaling pathway and the Rap1 signaling pathway were significantly enriched, each involving nearly 20 genes. The Focal adhesion pathway was also overrepresented among genes intersected by fixed blocks. All

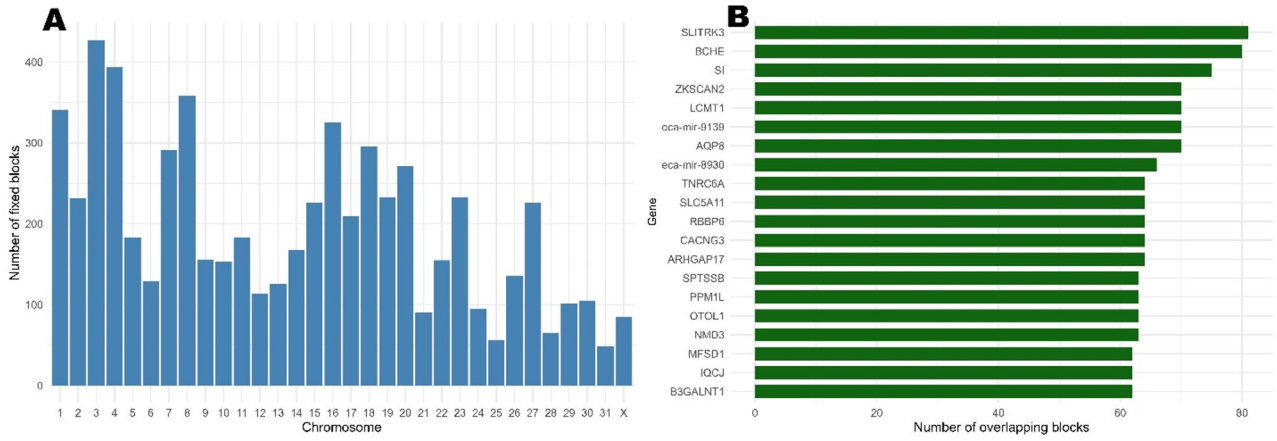

Fig. 5 Genomic distribution and gene overlap of blocks fixed for the alternate allele. **A** Number of genomic blocks with allele frequency of 1 and spanning at least four consecutive SNPs, shown for each chromosome. **B** Top genes most frequently overlapped by fixed blocks

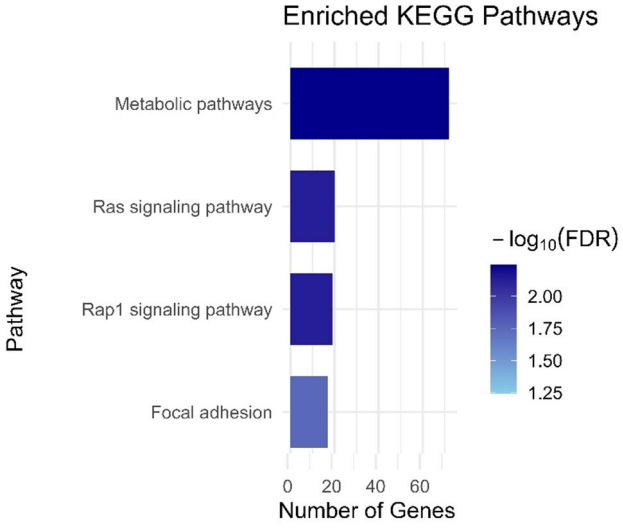

Fig. 6 Top enriched KEGG pathways among genes overlapping blocks of fixed alternate alleles

detailed information about identified pathways is presented in Supplementary Material 3.

To assess the potential functional significance of genomic regions fixed for the alternate allele, we intersected these regions with the current catalogue of equine quantitative trait loci (QTLs) available in the horse QTL database. This analysis identified several instances where fixed blocks overlapped with one known QTL – a QTL associated with insect bite hypersensitivity (QTL ID: 29289; Supplementary Material 3). No other overlaps were identified.

Discussion

Genome-wide genetic variation and the importance of an common variants allele frequency database in arabian horses

A central achievement of this study is the creation, to the best of our knowledge, of the first comprehensive, population-stratified allele frequency database for Arabian horses, encompassing 12 distinct pools derived from 120 individuals across four major geographic regions of Arabian horse breeding: Egypt, Syria, Poland, and the United States. By leveraging pooled whole-genome sequencing and robust variant calling pipelines, we catalogued over 11.7 million SNPs and estimated their allele frequencies, both globally and within each population group.

This publicly available allele frequency database for the Arabian horse represents a valuable new resource of information on common polymorphisms for the equine genomics community. In particular, it will enable more effective classification of genetic variants, supporting both research and clinical applications. In addition, it is important to note that the use of pooled sequencing, while enabling efficient characterization of population-level variation, may limit the detection of low-frequency variants and introduce some uncertainty in allele frequency estimation. These limitations are inherent to pooled designs and should be considered when interpreting rare variant patterns.

The database can be used to rapidly exclude common polymorphisms when searching for putative causal mutations underlying rare Mendelian diseases or complex traits, analogous to how human reference databases such as gnomAD are used in rare disease genomics [5, 6]. While population-specific data are somewhat robust for the Polish and United States groups—represented by seven and three pools, respectively—data for the Egyptian and Syrian origins are more limited, each being represented by a single pool of ten individuals. Nevertheless,

this approach enables the identification of founder or private alleles across all populations and therefore supports the refinement of variant interpretation pipelines for Arabian horses, addressing previous gaps in genomic resources for the breed [1].

Beyond its utility for disease variant prioritization, our study also highlights fundamental features of Arabian horse genomic diversity. The allele frequency spectrum observed here is consistent with those reported for other domestic horse populations, confirming a high level of genetic diversity within the Arabian breed [4]. At the chromosomal scale, we observed marked heterogeneity in SNP density and allele frequency, and relative enrichment of SNPs on chromosomes 12 and 20 was detected. Such patterns are well documented in equine and other mammalian genomes, reflecting influences from local mutation rates, recombination hotspots, and natural or artificial selection [2, 27].

Population structure and differentiation

Analyses of population structure using PCA, hierarchical clustering, and admixture approaches revealed clear genetic differentiation among the four geographic origins. Horses from Egypt and Syria formed distinctly separate clusters, each showing unique genetic signatures. Likewise, the Polish and United States pools formed their own distinct clusters, with these two groups being more closely related to each other than to the Egyptian or Syrian populations. This pattern is consistent with documented breeding histories and reflects the influence of both founder effects and population bottlenecks, particularly among Polish Arabian horses [9, 10].

In the case of the Polish Arabians, a genetic bottleneck occurred due to intensive selection and the use of a limited number of influential stallions after World War I and II, leading to reduced within-population diversity and stronger clustering in multivariate analyses [9]. This historical context helps explain the relatively high genetic homogeneity observed within the Polish pools. The US population, shaped by breeding stock imported from Poland and other European countries during the 20th century, exhibits similar patterns, further supporting this shared ancestry [9].

The optimal number of ancestral clusters identified in the admixture analysis (K = 4) supports the presence of distinct ancestry components corresponding to geographic origins, with minimal admixture among the main groups. Similar structure has been observed in previous studies, such as Petersen et al. [2], who reported discrete clusters in SNP-based analyses of Arabian and other horse breeds, highlighting the persistence of genetic differentiation despite modern breeding. The limited admixture and strong ancestry signals observed here underscore the impact of historical breeding practices, geographic separation, and conservation programs. In contrast to the Polish and US horses, the Syrian and Egyptian pools formed clearly distinct clusters, suggesting limited recent gene flow and retention of unique ancestral lineages [10].

It is also important to consider that the observed population structure may be influenced by the study design, particularly the unequal representation of populations. While the Polish and United States populations were represented by multiple pools, the Egyptian and Syrian groups were each represented by a single pool. As each pool was treated as a "pseudo-individual" in the admixture analysis, this imbalance may exaggerate apparent differentiation and reduce the ability to detect within-population variation in underrepresented groups. Therefore, the results of population structure analyses should be interpreted with caution, especially for comparisons involving Egyptian and Syrian populations.

Regions of fixation and functional enrichment

To identify genomic regions that may have been shaped by selection or other evolutionary forces, we scanned the genome for blocks where the alternate allele was fixed (AF = 1) across all studied populations. The presence of such fixed regions suggests past events—such as selection, bottlenecks, or genetic drift—that led to the elimination of allele diversity at these loci. An important consideration in the interpretation of these results is the use of a Thoroughbred-derived reference genome (EquCab3.0). In this context, genomic regions identified as "fixed for the alternate allele" may, in some cases, reflect differences between Arabian horses and the reference genome rather than true fixation within the studied populations. Therefore, some of the detected fixed blocks may represent positions where the reference genome carries derived alleles specific to Thoroughbreds. Future analyses incorporating comparisons with additional equine genomes, such as Przewalski's horse or multiple breed references, would help to better disentangle true selection signals from reference bias.

Our analysis identified 6,210 genomic blocks that were completely fixed for the alternate allele (AF = 1) across the combined set of Arabian horse pools. The distribution of these fixed blocks was highly uneven, with the greatest number observed on chromosome 3. The concentration of fixed genomic blocks within certain chromosomes and genes suggests the action of non-random evolutionary forces in Arabian horses, including potential signals of directional artificial selection, although alternative explanations such as genetic drift and demographic history cannot be excluded. Particularly, the overrepresentation of metabolic, Ras/Rap1 signaling, and focal adhesion pathways among genes overlapping with fixed regions was observed. Such pathway-level enrichment

is commonly interpreted as a signature of past selection targeting physiological or behavioral traits relevant to breed identity and performance [3, 28]. For example, the prominence of metabolic and cell signaling pathways aligns with the selection for endurance and robust health in Arabian horses, traits for which the breed is renowned [1]. Nevertheless, it is important to acknowledge the possible contribution of demographic factors, such as drift or founder effects, especially in populations with restricted gene flow in fixed regions formation [11].

In addition to broad pathway-level signals of selection and fixation, our study highlights several individual genes that overlap fixed blocks and are of particular interest given their known or plausible roles in equine biology and performance. For example, *RYR2* and *RYR3*, which encode ryanodine receptor calcium channels, are integral to muscle contraction and cardiac function. While most equine research has focused on *RYR1*, especially in the context of malignant hyperthermia and muscle disorders, *RYR2* is essential for normal cardiac excitation-contraction coupling. Selection or fixation in this locus could plausibly be linked to the high aerobic capacity and cardiac performance for which Arabian horses are renowned [29]. Another selection candidate is *IGF1R* (insulin-like growth factor 1 receptor), which is known as a regulator of growth, muscle development, and metabolism. Variants in the IGF1/IGF1R pathway have been associated with performance, adaptability, and muscle traits in horses [30]. Fixation of blocks within *IGF1R* could reflect historical selection for growth efficiency and endurance, which are hallmarks of the Arabian breed. Selection signals have also been found at the *GHR* gene (growth hormone receptor) locus, which is a key determinant of body size and energy metabolism. Previous genome-wide association studies have linked *GHR* variants to height and conformation traits in horses [31], which may have been shaped by both natural and artificial selection in Arabians. The *SOX9* gene, a master transcription factor in cartilage and skeletal development, is another notable candidate. Its importance for bone growth and joint health is well-established across mammals, and fixation within *SOX9* may relate to the soundness and distinctive conformation of Arabian horses [32]. Taken together, these candidate genes warrant further investigation, as they may contribute to key Arabian horse traits such as endurance, metabolic efficiency, musculoskeletal soundness, and temperament. Their fixation suggests an important role for historical selection and adaptation in shaping the Arabian horse genome.

A notable aspect of our findings is the overlap between several fixed genomic blocks and a previously reported QTL associated with insect bite hypersensitivity (IBH; QTL ID: 29289; [33]), a recurrent allergic condition in horses. Within this region, we observed complete fixation of alternate alleles relative to other populations, suggesting historical divergence or targeted selection. Although our data do not include phenotypic information to directly assess IBH susceptibility, the haplotype structure in this breed differs from that reported in IBH-affected populations, implying that it may represent a distinct, potentially less susceptible genetic background. The fixation of alleles in this QTL region could therefore reflect historical selection favoring immune modulation or skin barrier traits that incidentally confer resistance to IBH. Alternatively, these patterns may have arisen through breed-specific drift or founder effects given the closed breeding history of the population. Further investigation integrating functional annotation, gene expression, and phenotypic data will be required to determine whether these variants indeed contribute to reduced IBH risk. Collectively, the overlap of fixed regions with trait-associated loci underscores how both adaptive processes and demographic history have shaped the genomic architecture of this breed.

The regions identified as fixed in this study may also be related to previously reported runs of homozygosity (ROH) islands in Arabian and related horse breeds. For example, Grilz-Seger et al. [34] described genomic regions with extended homozygosity that were associated with selection and demographic history in horse populations. Although our study is based on pooled sequencing data and does not directly estimate ROH at the individual level, the overlap between fixed regions and known ROH islands may provide additional support for the role of historical selection and population structure in shaping these genomic patterns.

Screening for disease-causing variants

We screened our dataset for several well-characterized disease-causing variants in Arabian horses—including mutations responsible for Lavender Foal Syndrome (LFS), Severe Combined Immunodeficiency (SCID), Cerebellar Abiotrophy (CA), and Occipitoatlantoaxial Malformation (OAAM). None of these known pathogenic variants were detected in our dataset, even though we had a sufficient coverage for the sites of interest. This absence is consistent with their low population frequencies; however, it should be interpreted with caution given the sample size and pooled sequencing design [7]. In particular, the absence of these variants in a cohort of 120 individuals does not exclude their presence at low frequencies in the broader Arabian horse population. Under a simple binomial model, the probability of detecting at least one carrier depends strongly on the true allele frequency; for example, variants with carrier frequencies below approximately 2–5% may not be detected in a dataset of this size. Therefore, our results should be interpreted as indicating that these variants are not common in the studied

populations rather than being entirely absent. This distinction is important for accurate interpretation in both research and breeding contexts. Nevertheless, our findings emphasize the importance of continuously updating population-specific allele frequency resources to monitor the prevalence of deleterious alleles and to detect emerging pathogenic variants before they become widespread.

Study limitations and perspectives of database practical applications

A major limitation of this study is, in our opinion, the uneven sampling of populations, with Polish and US groups represented by multiple pools and the Egyptian and Syrian populations by only a single pool each. This restricts the ability to detect substructure or rare variants in these underrepresented groups. Additionally, pooled sequencing approaches reduce the ability to detect rare variants and prevent direct inference of individual genotypes. While high sequencing depth improves allele frequency estimation, some uncertainty remains, particularly for low-frequency alleles and in regions with uneven coverage [35]. As the resource evolves, increasing sample size and pool diversity, particularly for Middle Eastern populations, will further enhance the accuracy and utility of the database.

This allele frequency resource represents a valuable tool for equine genomic research, particularly for improving the interpretation of genetic variants in the Arabian horse. By providing population-level allele frequencies, the database facilitates the distinction between rare potentially deleterious mutations and common benign polymorphisms [5, 7]. Such data are especially useful in genome-wide association studies (GWAS), where background allele frequencies can help refine association signals, reduce false positives, and prioritize true candidate loci. In this context, the resource serves as a critical reference for the identification of genetic variants associated with health-related traits or breed-specific phenotypes.

As equine GWAS results accumulate and causal variants are validated, they can ultimately inform breeding decisions and health management strategies, thereby supporting the transition from research findings to practical genomic selection tools. In this way, our dataset lays the groundwork for future integration of genomic data into precision breeding approaches aimed at improving genetic health and performance traits in Arabian horses.

To address these limitations, the database is intended to remain an open and continuously evolving resource. Future expansions will incorporate additional samples from currently represented populations, as well as new cohorts from previously underrepresented or unrepresented Arabian horse populations, where available. This ongoing enrichment will enhance the accuracy of allele frequency estimates, increase the detection power for rare variants, and improve resolution of population substructure across geographic and genetic lineages. Particular emphasis will be placed on achieving a more balanced representation, especially of Middle Eastern populations, to ensure that the resource remains broadly applicable and maximally informative for both genomic research and practical breeding applications.

Furthermore, the use of "pseudo-individuals" derived from pooled allele counts for admixture analysis represents an approximation of true individual-level data. While this approach enables exploratory analysis of population structure, it may introduce biases compared to methods based on genotype likelihoods or individual genotypes. Therefore, the admixture results should be interpreted as indicative rather than definitive, and future studies using individual-level sequencing data would provide higher resolution insights.

Finally, analyses including the X chromosome should be interpreted with caution, as differences in sex composition among pools and the hemizygous nature of the X chromosome in males may influence allele frequency estimates and population structure patterns.

Conclusion

In this study we created a database of common genetic variation and characterized genome-wide genetic diversity and structure in Arabian horses using pooled sequencing across multiple populations. Our analysis also identified regions of fixation, presumably formed by artificial selection, which were enriched for genes related to metabolism, growth, and musculoskeletal function, as well as overlapped with a QTL for insect bite hypersensitivity. Importantly, we generated a novel allele frequency database for Arabian horses, which can be used for genetic variant filtering, disease association studies, and the interpretation of rare variants in this breed. This resource will facilitate future research on Arabian horse health, performance, and population genetics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12862-0.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements

We thank the horse owners and breeding institutions for providing samples and the sequencing service providers for technical support. We also acknowledge the assistance of colleagues who contributed to data processing and manuscript preparation.

Ethical statement

The material used in this study comprised hair follicles and blood samples that originated from the genetic material bank of the National Research Institute

of Animal Production (Balice, Poland). The samples were obtained and stored in accordance with the applicable Polish law, including the Act of 15 January 2015 on the Protection of Animals Used for Scientific or Educational Purposes (Journal of Laws 2015, item 266).
All samples had been previously collected for routine veterinary and diagnostic purposes and subsequently deposited in the Institute's genetic resource bank. Therefore, no experimental procedures were performed on live animals specifically for the purpose of this study. According to the above-mentioned Act, and as confirmed by institutional regulations, this type of research does not require the approval of the Local Ethical Committee for Animal Experimentation.

Authors' contributions

T.S. designed the study, performed analyses, and wrote the manuscript. C.F., A.G., H.H., M.S.-S., S.A., and E.N. contributed to sample acquisition, methodology, and data interpretation. K.R.-M. and C.F. supervised the project, critically revised the manuscript, and K.R.-M. additionally secured funding for the study from the Ministry of Agriculture and Rural Development of the Republic of Poland. All authors read and approved the final manuscript.

Funding

This research was supported by the Ministry of Agriculture and Rural Development of the Republic of Poland (Grant No. 503-182-609).

Data availability

All raw sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA) under accession numbers PRJNA1332950. The allele frequency databases with and without annotation and the script that can be used as a search tool of specific genetic variants are available at Zenodo (https://doi.org/10.5281/zenodo.17483784) - Databases of genetic variants with allele frequency of Arabian horses.

Declarations

Ethics approval and consent to participate

The material used in this study comprised hair follicles and blood samples that originated from the genetic material bank of the National Research Institute of Animal Production (Balice, Poland). The samples were obtained and stored in accordance with the applicable Polish law, including the Act of 15 January 2015 on the Protection of Animals Used for Scientific or Educational Purposes (Journal of Laws 2015, item 266).
The samples used as archival material from the Biological Material Bank at the National Research Institute of Animal Production were collected with the approval of the Institutional Animal Care and Use Committee (no. 1173/2015 and 102/2025). Hair follicle and blood samples were collected with informed owner consent.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Basic Sciences, University of Agriculture in Kraków, Kraków, Poland
[2]Department of Animal Molecular Biology, Laboratory of Genomics, National Research Institute of Animal Production, Krakowska 1, Kraków 32-083 Balice, Poland
[3]Department of Population Health and Reproduction, Davis School of Veterinary Medicine, University of California, Davis, USA
[4]Department of Biomolecular Engineering and Bioinformatics, University of California, Santa Cruz, USA
[5]Department of Health and Care Management, Arden University, Berlin Campus, Berlin, Germany
[6]College of Veterinary Medicine, University of Arizona, Oro Valley, USA

Received: 2 December 2025 / Accepted: 10 April 2026
Published online: 20 April 2026

References

1. Cosgrove EJ, Sadeghi R, Schlamp F, Holl HM, Moradi-Shahrbabak M, Miraei-Ashtiani SR, et al. Genome diversity and the origin of the Arabian horse. Sci Rep. 2020;10:9702. https://doi.org/10.1038/s41598-020-66232-1.
2. Petersen JL, Mickelson JR, Cothran EG, Andersson LS, Axelsson J, Bailey E, et al. Genetic diversity in the modern horse illustrated from genome-wide SNP data. PLoS ONE. 2013;8(1):e54997. https://doi.org/10.1371/journal.pone.0054997.
3. Librado P, Fages A, Gaunitz C, Khan N, Schubert M, Albrechtsen A, et al. Ancient genomic changes associated with domestication of the horse. Science. 2017;356(6336):442–5. https://doi.org/10.1126/science.aam5298.
4. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499(7456):74–8. https://doi.org/10.1038/nature12323.
5. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43. https://doi.org/10.1038/s41586-020-2308-7.
6. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91. https://doi.org/10.1038/nature19057.
7. Brooks SA, Gabreski NA, Miller DC, Brisbin A, Brown HE, Streeter C, et al. Whole-genome SNP association in the horse: identification of a deletion in myosin Va responsible for lavender foal syndrome. PLoS Genet. 2010;6(4):e1000909. https://doi.org/10.1371/journal.pgen.1000909.
8. Aleman M, Finno CJ, Weich K, Penedo MCT. Investigation of known genetic mutations of Arabian horses in Egyptian Arabian foals with juvenile idiopathic epilepsy. J Vet Intern Med. 2017;32(1):465–8. https://doi.org/10.1111/jvim.14873.
9. Khanshour A, Conant E, Juras R, Cothran EG. Microsatellite analysis of genetic diversity and population structure of Arabian horse populations. J Hered. 2013;104(3):386–98. https://doi.org/10.1093/jhered/est003.
10. Klecel W, Kloch M, Wojciechowska M, Gajewska M, Martyniuk E. Population structure and genetic diversity of Polish Arabian horses based on pedigree data. Animal. 2024;18(5):101148. https://doi.org/10.1016/j.animal.2024.101148.
11. Wallner B, Vogl C, Shukla P, Burgstaller JP, Druml T, Brem G. Y chromosome uncovers the recent oriental origin of modern stallions. Curr Biol. 2017;27(13):2029–e355. https://doi.org/10.1016/j.cub.2017.05.086.
12. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
13. Dodt M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology (Basel). 2012;1(3):895–905. https://doi.org/10.3390/biology1030895.
14. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013. Available from: https://arxiv.org/abs/1303.3997.
15. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
16. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. https://doi.org/10.1101/gr.107524.110.
17. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. 2013–2015. Available from: http://www.repeatmasker.org.
18. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. 2023. Available from: https://www.R-project.org/.
19. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526–8. https://doi.org/10.1093/bioinformatics/bty633.
20. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19(9):1655–64. https://doi.org/10.1101/gr.094052.109.
21. Bu D, Luo H, Huo P, Wang Z, Zhang S, He Z, et al. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. Nucleic Acids Res. 2021;49(W1):W317–25. https://doi.org/10.1093/nar/gkab447.

22. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.
23. Ayad A, Almarzook S, Besseboua O, Aissanou S, Piórkowska K, Musiał AD, et al. Investigation of Cerebellar Abiotrophy (CA), Lavender Foal Syndrome (LFS), and Severe Combined Immunodeficiency (SCID) variants in a cohort of three MENA region horse breeds. Genes (Basel). 2021;12(12):1893. https://doi.org/10.3390/genes12121893.
24. Finno CJ, Spier SJ, Valberg SJ. Equine diseases caused by known genetic mutations. Vet J. 2009;179(3):336–47. https://doi.org/10.1016/j.tvjl.2008.03.016.
25. Scott E, Woolard K, Finno CJ, Murray JD. Cerebellar abiotrophy across domestic species. Cerebellum. 2018;17(3):372–9. https://doi.org/10.1007/s12311-017-0914-1.
26. Finno CJ. Occipitoatlantoaxial malformations in equids. Equine Vet Educ. 2025. https://doi.org/10.1111/eve.14160. in press.
27. Doan R, Cohen ND, Sawyer J, Ghaffari N, Johnson CD, Dindot SV. Identification of copy number variants in horses. Genome Res. 2012;22(5):899–907. https://doi.org/10.1101/gr.128991.111.
28. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 2014;46(8):858–65. https://doi.org/10.1038/ng.3034.
29. Bers DM. Cardiac excitation–contraction coupling. Nature. 2002;415(6868):198–205. https://doi.org/10.1038/415198a.
30. Schröder W, Klostermann A, Distl O. Candidate genes for physical performance in the horse. Vet J. 2011;190(1):39–48. https://doi.org/10.1016/j.tvjl.2010.09.029.
31. Signer-Hasler H, Flury C, Haase B, Burger D, Simianer H, Leeb T, et al. A genome-wide association study reveals loci influencing height and other conformation traits in horses. PLoS ONE. 2012;7(5):e37282. https://doi.org/10.1371/journal.pone.0037282.
32. Akiyama H, Chaboissier M, Martin JF, Schedl A, de Crombrugghe B. The transcription factor SOX9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of SOX5 and SOX6. Genes Dev. 2002;16(21):2813–28. https://doi.org/10.1101/gad.1017802.
33. Schurink A, Wolc A, Ducro B, Frankena K, Garrick DJ, Dekkers JCM, et al. Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. Genet Sel Evol. 2012;44(1):31. https://doi.org/10.1186/1297-9686-44-31.
34. Grilz-Seger G, Neuditschko M, Ricard A, Velie B, Lindgren G, Mesarič M, et al. Genome-wide homozygosity patterns and evidence for selection in a set of European and Near Eastern horse breeds. Genes (Basel). 2019;10(7):491. https://doi.org/10.3390/genes10070491.
35. Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. Nat Rev Genet. 2014;15(11):749–63. https://doi.org/10.1038/nrg3803.

Publisher's note