

Population Genomics Reveals Genetic Structure, Genomic Inbreeding, and Shared Homozygosity Hotspots in the Peruvian Paso Horse

Carla Saldaña

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Lenin Quiñones Huatangari

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Luis Murga

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Jorge Maicelo

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Carlos Arbizu

`carlos.arbizu@untrm.edu.pe`

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

William Bardales

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM)

Research Article

Keywords: Zoogenetic resources, High-density SNP markers, Sustainable conservation, Equine

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11053775/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The Peruvian Paso horse (PPH) is an iconic gaited breed recognized for its distinctive smooth gait, historical importance, and cultural value. However, genome-wide information on its genetic diversity, population structure, demographic history, and genomic inbreeding remains limited. This study characterized genomic diversity, population structure, linkage disequilibrium (LD), effective population size, and runs of homozygosity (ROH) in southern PPH using high-density SNP markers. A total of 141 individuals were genotyped with the Illumina Equine SNP70 BeadChip, and 49,610 autosomal SNPs passed quality control. Bayesian clustering identified three genetic clusters, with substantial admixture and weak differentiation indicating considerable connectivity among breeding lineages. Principal coordinates analysis supported these patterns and showed partial overlap among clusters. Genetic diversity was moderate, with observed heterozygosity ranging from 0.322 to 0.363 and negative inbreeding coefficients. Demographic analyses revealed contrasting patterns, with Cluster 2 showing the highest LD and lowest effective population size, whereas Cluster 1 exhibited the lowest LD and largest effective population size. ROH analyses revealed differences in genomic inbreeding, with Cluster 3 showing the highest FROH and greatest abundance of long homozygous segments. Several ROH hotspots were identified, including a chromosome 7 hotspot shared across clusters. Functional annotation highlighted genes associated with epigenetic regulation, tissue remodeling, skeletal development, and neuronal processes. Overall, southern PPH retains moderate genomic diversity despite population structure and localized autozygosity. These findings provide novel insights into the breed's genomic architecture and breeding history and support future conservation and sustainable breeding strategies.

Introduction

The horse (*Equus caballus*) has a complex evolutionary and breeding history shaped by domestication, rapid geographic expansion and intensive human-mediated selection. Ancient genome analyses have shown that the ancestry of modern domestic horses emerged in the western Eurasian steppes and spread rapidly across Eurasia around 4,200 years ago, profoundly reshaping equine population history (Librado et al. 2021). In parallel, genome-wide studies in contemporary breeds have revealed marked differences in genetic diversity, population structure, admixture, genomic inbreeding and runs of homozygosity, reflecting founder effects, regional breeding practices and recent selection (Petersen et al. 2013; Criscione et al. 2022). Together, these features make SNP-based genomic analyses especially valuable for investigating how breeding history and geographic subdivision shape equine populations and for informing conservation and breeding strategies.

High-density SNP genotyping has become a powerful and cost-effective tool for equine population genomics. Equine SNP arrays were specifically developed to enable rapid genome-wide characterization of horses and have been widely applied to investigate breed relationships, genetic diversity, linkage disequilibrium, genomic inbreeding and signatures of selection (Petersen et al. 2013; McCue et al. 2012). In recent years, SNP-based analyses have also been successfully used to characterize local and native

horse populations, including the estimation of effective population size, the detection of regional genetic structure and the assessment of runs of homozygosity, all of which are highly relevant for the design of conservation and breeding strategies in domestic animals (Criscione et al. 2022; Bazvand et al. 2024).

Against this background, the Peruvian Paso Horse (PPH) emerges as a particularly relevant target for genome-wide population analysis. As one of Peru's most emblematic equine breeds, it is internationally recognized for its smooth and distinctive gait, the *paso llano*, a symmetrical four-beat gait with a lateral footfall sequence characteristic of gaited horses (Nicodemus and Clayton 2003). In the PPH, functional traits related to gait expression, including overreach, term and acuteness, have also been recognized as relevant breeding traits (Vilela Velarde et al. 2025), and kinematic approaches have been applied to evaluate variation in limb joint angles during locomotion in the breed (Gonzales et al. 2018).

Beyond its zootechnical relevance, the breed is also recognized as a flagship product of Peru (Saldaña et al. 2025). Nevertheless, despite its biological, breeding and cultural importance, genomic research in the PPH remains limited. Recent assembly of its complete mitochondrial genome provided an important first step by establishing a maternal genomic resource for phylogenetic and conservation studies (MINCETUR 2012). However, because mitochondrial DNA reflects only maternal inheritance, it cannot resolve the genome-wide diversity, fine-scale population structure and homozygosity patterns required to support breeding and conservation management (Ceballos et al. 2018).

This limitation is especially relevant in the PPH, which is bred across geographically distinct regions of Peru and has historically been managed through interconnected breeding networks. Under such conditions, differences in mating patterns, founder representation, selection practices, and gene flow may generate fine-scale genomic structure that is not necessarily reflected by geographic origin alone (Bazvand et al. 2024; Grilz-Seger et al. 2019). This has important practical implications because hidden population structure can influence the distribution of genetic diversity, the accumulation of genomic inbreeding, and the interpretation of downstream genomic analyses (Criscione et al. 2022; Bazvand et al. 2024). Evidence from other horse breeds indicates that integrating analyses of population structure with measures of heterozygosity, linkage disequilibrium, effective population size, and runs of homozygosity provides a robust framework for reconstructing recent demographic history, identifying genetically differentiated groups, and recognizing populations at greater risk of diversity loss (McCue et al. 2012; Nicodemus and Clayton 2003; Grilz-Seger et al. 2019).

In the present study, we used genome-wide SNP genotyping data to characterize the genetic diversity, population structure, linkage disequilibrium, effective population size, genomic inbreeding, and runs of homozygosity of the southern Peruvian Paso horse population sampled from Lima, Ica, and Arequipa. Particular emphasis was placed on identifying genetically differentiated groups within the breed and evaluating whether the contemporary southern population is composed of distinct genetic clusters that may represent meaningful units for conservation and breeding management. Understanding the distribution of genetic variation among such groups is essential for maintaining long-term genetic diversity, minimizing the accumulation of inbreeding, and preserving unique genetic lineages within the

breed (Ceballos et al. 2018; Charlesworth 2009). Furthermore, we investigated genomic regions exhibiting elevated homozygosity and potential signatures of historical selection. By providing the first comprehensive genome-wide assessment of these parameters in the southern Peruvian Paso horse, this study establishes a valuable foundation for conservation planning, evidence-based breeding strategies, and future investigations into the genomic basis of breed-specific traits.

Materials and Methods

Animal sampling and DNA extraction

A total of 143 Peruvian Paso horses were randomly selected from three geographical departments of southern Peru: Lima (n = 96), Ica (n = 20) and Arequipa (n = 27). The sampled population included male and female animals of different ages, aiming to represent the genetic diversity of the population under investigation. Breed identity was established based on owner-provided information, breed records when available, and phenotypic characteristics consistent with the Peruvian Paso horse standard.

For each individual, tail hair samples were collected as the source of genomic DNA, preferentially using hairs with intact follicles. Sampling was conducted under standard handling practices to minimize animal stress and prevent injury. All samples were individually labeled and recorded according to the corresponding animal identification and associated metadata.

All animal procedures were performed in accordance with relevant institutional ethical guidelines and current animal welfare regulations. The study protocol was approved by the appropriate institutional ethics committee, and written informed consent was obtained from the owners before sample collection. Following collection, samples were mounted on hair cards provided by Neogen and shipped to the Neogen Geneseek laboratory (Lincoln, NE, USA), where DNA extraction was performed prior to SNP genotyping.

Genotyping and SNP quality control

A total of 143 Peruvian Paso horse samples were genotyped using the Illumina Equine SNP70 BeadChip at the Neogen GeneSeek laboratory (Lincoln, NE, USA), following the manufacturer's standard procedures. Raw genotyping reports were converted into PLINK-compatible formats for downstream population genomic analyses.

Quality control filtering was conducted using PLINK v1.9 (Purcell et al. 2007). Individuals with more than 10% missing genotypes were excluded, as were SNPs with more than 10% missing data, SNPs with a minor allele frequency (MAF) lower than 0.05, and SNPs deviating from Hardy–Weinberg equilibrium (HWE; $P < 1 \times 10^{-6}$). To reduce the effect of linkage disequilibrium in multivariate and clustering analyses, linkage disequilibrium (LD)-based pruning was performed in PLINK using the `--indep-pairwise`

50 5 0.2 option, corresponding to a sliding window of 50 SNPs, a step size of 5 SNPs, and an r^2 threshold of 0.2.

After quality filtering, the final dataset consisted of 141 individuals and 49,444 high-quality autosomal SNP markers. Following LD-based pruning, 7,658 SNPs were retained for population structure analyses, including PCoA, STRUCTURE, and dendrogram reconstruction,

Population Structure and genetic clustering

Population structure and genetic differentiation were evaluated using complementary multivariate, distance-based, and Bayesian clustering approaches. Pairwise F_{ST} values were estimated using the filtered autosomal SNP dataset (Weir and Cockerham 1984). Principal coordinates analysis (PCoA) was performed using the LD-pruned SNP dataset. Pairwise genetic distances among individuals were calculated using the poppr package (Kamvar et al. 2014), and the resulting distance matrix was subjected to eigenanalysis using the pcoa function implemented in the ape package (Paradis and Schliep 2019).

The distribution of genetic variation within and among inferred genetic clusters was further assessed through analysis of molecular variance (AMOVA) using the poppr package in R (Kamvar et al. 2014). Genetic relationships among individuals were additionally evaluated through an unweighted pair group method with arithmetic mean (UPGMA) dendrogram constructed from a genetic distance matrix with 1,000 bootstrap replicates to assess node support.

To determine the population structure, we converted VCF data sets to STRUCTURE format using PGSpider v.2.1.1.5 software (Lischer and Excoffier 2012). Then, we employed STRUCTURE program v.2.3.4 (Pritchard et al. 2000) with 10 runs for each population, ranging from 1 to 12 with a burn – in length of 20,000 and 50,000 Monte Carlo iterations. We used STRUCTURE HARVESTER software (Earl and vonHoldt 2012) to process the STRUCTURE results, detecting the most likely number of clusters by the Evanno method (Evanno et al. 2005). Finally, we used the pophelper package v.2.3.1 (Francis 2017) in R to visualize the STRUCTURE membership.

Diversity Analysis within inferred clusters

Genetic diversity was assessed using the filtered autosomal SNP dataset. Observed heterozygosity (H_o), expected heterozygosity (H_e), inbreeding coefficient (F_{IS}), and minor allele frequency (MAF) were estimated for each inferred genetic cluster and for the complete dataset. The filtered VCF dataset was imported into R v4.3.1 software (R Core Team 2024), where genotype data were converted into a genind object using the vcfR and adegenet packages. Population assignments were defined according to sample identifiers, and genetic diversity statistics (H_o , H_e , and F_{IS}) were calculated using the hierfstat package in R. Mean values were summarized for each population and for the overall dataset. Minor allele frequency (MAF) was estimated separately for each population using PLINK v1.9 (Purcell et al. 2007),

based on the filtered autosomal SNP dataset. Mean MAF values were subsequently summarized by population.

Linkage disequilibrium decay (LD) and effective population size (Ne)

LD decay was evaluated separately for the three genetic clusters inferred by STRUCTURE using the filtered SNP dataset obtained after quality control. Pairwise LD between SNPs was estimated as the squared correlation coefficient (r^2) using PLINK v1.9. Analyses were performed independently for each cluster using cluster-specific sample lists. The following parameters were applied: `--r2`, `--ld-window 999999`, `--ld-window-kb 10000`, and `--ld-window-r2 0`, allowing the calculation of LD across all SNP pairs within a maximum physical distance of 10 Mb. The extent of LD decay was assessed by grouping SNP pairs into physical distance classes (10, 20, 40, 60, 100, 200, 500, and 1000 kb) and calculating the mean r^2 value for each class. LD decay patterns were visualized in R using the ggplot2 package (Wickham 2016).

Historical effective population size (N_e) was inferred independently for each genetic cluster from LD estimates. Physical distances between SNPs were converted into recombination fractions assuming an approximate rate of 1 cM/Mb, and adjusted r^2 values were used to estimate N_e across recent generations. Median N_e values were summarized for 10, 20, 30, 40, and 50 generations ago and visualized in R using ggplot2 (Wickham 2016).

Runs of Homozygosity (ROH) and Genomic Inbreeding

ROH analyses were conducted using the R package detectRUNS v0.9.6 (Biscarini et al. 2019) based on the quality-filtered SNP dataset previously generated in PLINK. Prior to ROH detection, PLINK-formatted files were converted to PED/MAP format using PLINK v1.9 (Purcell et al. 2007) with the `--recode` option. Consecutive homozygous segments were identified using the `consecutiveRUNS.run` function implemented in detectRUNS. ROH calling parameters were selected following commonly applied criteria in livestock SNP-array studies to minimize false-positive ROH generated by background linkage disequilibrium while retaining biologically meaningful homozygous segments. The minimum number of SNPs required to define a ROH was set to 50 SNPs, allowing a maximum of one heterozygous genotype within a run (`maxOppRun = 1`) and no missing genotypes (`maxMissRun = 0`). The minimum ROH length was established at 500 kb (`minLengthBps = 500000`), and the maximum allowed distance between consecutive SNPs within a ROH was set to 500 kb (`maxGap = 500000`). ROH hotspot regions were identified by estimating the frequency of SNP occurrence within ROH segments across individuals within each population. SNPs present within ROH in at least 50% of individuals were considered candidate ROH hotspots.

Consecutive hotspot SNPs were merged into genomic regions, and candidate genes located within these intervals were annotated using the Ensembl BioMart (Kinsella et al., 2011). database based on the EquCab3.0 reference genome assembly. Candidate genes located within shared ROH hotspots were

functionally evaluated according to their reported biological roles in connective tissue organization, cellular signaling, metabolic regulation, and neuromuscular processes. Genome-wide genomic inbreeding coefficients (F_{ROH}) were estimated using the *Froh_inbreeding* function implemented in detectRUNS. The genomic inbreeding coefficient was calculated as the proportion of the autosomal genome covered by ROH segments. To further distinguish recent and ancient autozygosity, F_{ROH} was additionally estimated using different ROH length classes, including ROH segments larger than 2, 4, 8, and 16 Mb.

Results

SNP dataset after quality control

The initial dataset comprised 143 Peruvian Paso Horse individuals genotyped at 65,090 SNP markers. Following quality control filtering, two individuals were excluded due to high levels of missing genotype data. At the marker level, SNPs were removed because of missing genotype rates, deviation from Hardy–Weinberg equilibrium, low minor allele frequency ($MAF < 0.05$), and the presence of multiallelic loci.

After filtering, the final dataset retained for downstream analyses consisted of 141 individuals and 49,444 autosomal SNP markers. Linkage disequilibrium pruning further reduced the dataset to 7,658 SNPs, which were subsequently used for population structure analyses, including PCoA, STRUCTURE, and dendrogram reconstruction.

Population Structure

Structure analysis identified three genetic clusters ($K = 3$) within the southern PPH population (Fig. 1). The mean likelihood values obtained from STRUCTURE analyses across $K = 1–10$ supported the presence of genetic subdivision within the population (Figure S1). The optimal number of clusters was subsequently determined using the Evanno method, which showed a clear peak at $K = 3$ (Figure S2), indicating that three genetic clusters best explained the observed genetic structure. Cluster 1 represented the predominant genetic background and included individuals from all three sampled regions (Arequipa, Ica, and Lima). In contrast, Cluster 2 was composed exclusively of horses from Lima, whereas Cluster 3 was dominated by Lima individuals but also included a small number of horses from Arequipa and Ica. Several individuals displayed admixed ancestry, particularly between Clusters 1 and 3, indicating ongoing genetic connectivity among breeding lineages. Detailed information for all samples, including geographic origin and cluster assignment, is provided in Supplementary Table S1.

Principal coordinate analysis (PCoA) based on the LD-pruned SNP dataset revealed a clear genetic structure among southern PPH, consistent with the clustering pattern inferred (Fig. 2). Three well-defined genetic clusters were identified along the first two principal components. Cluster 1 comprised the majority of individuals and occupied a central position in the ordination space, whereas Cluster 2 and Cluster 3 formed distinct groups located at opposite extremes of PC 1 and PC 2. Cluster 2 was clearly separated from the remaining individuals along positive values of both principal components, while

Cluster 3 was distributed along negative values of PC1 and positive values of PC2. Only limited overlap was observed among clusters, indicating substantial genetic differentiation among the inferred groups. The first two principal components explained 5.79% and 4.12% of the total genetic variation, respectively, reflecting the polygenic nature of population structure while still capturing the major patterns of genetic subdivision within the dataset.

Pairwise F_{ST} estimates revealed moderate levels of genetic differentiation among the inferred genetic clusters (Table 1). The highest differentiation was observed between Cluster 2 and Cluster 3 ($F_{ST} = 0.128$), indicating substantial genetic divergence between these groups. Intermediate differentiation was detected between Cluster 1 and Cluster 2 ($F_{ST} = 0.069$), whereas the lowest differentiation was found between Cluster 1 and Cluster 3 ($F_{ST} = 0.042$). Overall, these results support the presence of a clear genetic structure within the southern Peruvian Paso horse population, with Cluster 1 and Cluster 3 being genetically more similar to each other than either is to Cluster 2.

Table 1
Pairwise F_{ST} values among the three genetic clusters inferred ($K = 3$) in southern PPH

Cluster 1	Cluster 1	Cluster 2	Cluster 3
	NA	0.069	0.042
Cluster 2	0.069	NA	0.128
Cluster 3	0.041	0.128	NA

Analysis of molecular variance (AMOVA) based on the three genetic clusters inferred revealed that most of the genetic variation was distributed within clusters, accounting for 88.65% of the total variation (Table 2). In contrast, variation among clusters represented 11.35% of the total genetic variance. These results are consistent with the pairwise F_{ST} estimates and the clear separation observed in the PCoA, supporting the presence of a well-defined genetic structure within the southern PPH population.

Table 2
Analysis of molecular variance among southern PPH clusters

Source of variation	df	SS	MS	Est.Var.	%
Among clusters	2	12064.16	19226.97	166.59	11.35
Within clusters	138	1301.48	8756.05	1301.48	88.65
Total	140	1369.06	8905.63	1468.07	100.00

The UPGMA dendrogram based on Nei's genetic distance recovered three major genetic groups that closely corresponded to the clusters identified ($K = 3$) (Fig. 3). Cluster 2 formed a distinct and highly differentiated lineage composed predominantly of individuals from Lima, whereas Cluster 3 included individuals from all three regions but showed a relatively higher representation of horses from Ica and

Arequipa. Cluster 1 represented the largest genetic group and comprised individuals from Lima, Ica, and Arequipa, reflecting a substantial degree of genetic admixture among geographic regions. Although some intermixing among clusters was observed, the overall topology was largely consistent with the PCA, pairwise F_{ST} estimates, and AMOVA results. Bootstrap support values were generally high across the principal nodes, with several branches receiving support values above 90%, indicating strong confidence in the inferred clustering pattern.

Genetic diversity

Overall mean values of observed heterozygosity (H_o) and expected heterozygosity (H_e) across the three genetic clusters were moderate, ranging from 0.322 to 0.363 and from 0.294 to 0.358, respectively. Cluster 1 exhibited the highest levels of genetic diversity, with H_o and H_e values of 0.363 and 0.358, respectively, as well as the highest mean minor allele frequency ($MAF = 0.284$). In contrast, Cluster 3 showed the lowest diversity estimates, with $H_o = 0.322$, $H_e = 0.294$, and $MAF = 0.222$. Cluster 2 displayed intermediate values for all diversity parameters ($H_o = 0.361$, $H_e = 0.330$, $MAF = 0.250$). Negative inbreeding coefficients (FIS) were observed in all clusters, ranging from -0.014 in Cluster1 to -0.095 in Cluster 3, indicating a slight excess of heterozygous genotypes relative to Hardy–Weinberg expectations. The consistently negative FIS values suggest the absence of substantial inbreeding within clusters and are consistent with the moderate levels of genetic admixture observed among individuals. Overall, the results revealed moderate genetic diversity across the inferred genetic clusters, with Cluster 1 retaining the greatest allelic variability and representing the largest genetic group identified in the population (Table 3).

Table 3
Genetic diversity statistics for the inferred genetic clusters of southern PPH ($K = 3$)

Cluster	N	H_o	H_e	FIS	MAF
Cluster 1	106	0.363	0.358	-0.014	0.284 ± 0.001
Cluster 2	14	0.361	0.330	-0.093	0.250 ± 0.001
Cluster 3	21	0.322	0.294	-0.095	0.222 ± 0.001

Linkage disequilibrium decay (LD) and effective population size (N_e)

Linkage disequilibrium decreased with increasing physical distance between SNPs in all inferred genetic clusters, following the expected LD decay pattern (Fig. 4, Supplementary Table S2.). However, marked differences were observed among clusters. Cluster 1 consistently exhibited the lowest LD values across all distance classes, whereas Cluster 2 showed the highest LD levels, with Cluster 3 presenting intermediate values. At short genomic distances, mean r^2 values were high in all clusters, ranging from approximately 0.45 in Cluster 1 to 0.50 in Clusters 2 and 3. LD declined progressively as marker distance

increased, although the rate of decay differed among clusters. Cluster 1 showed a rapid decline in LD, reaching values below 0.10 at distances greater than 1 Mb, whereas Clusters 2 and 3 maintained comparatively higher LD values across the entire range of distances evaluated. At the largest distance class, mean r^2 values remained approximately 0.13 in Cluster 2, 0.11 in Cluster 3, and 0.04 in Cluster1.

Historical effective population size estimates revealed marked differences among the three inferred genetic clusters across the last 50 generations (Fig. 5, Supplementary Table S3.). Cluster 1 consistently exhibited substantially larger effective population sizes than the other clusters throughout the analyzed period. Estimated N_e values for Cluster 1 increased from approximately 150 individuals 10 generations ago to nearly 460 individuals 50 generations ago. In contrast, Cluster 2 and Cluster 3 displayed considerably lower effective population sizes, ranging from approximately 40–45 individuals at 10 generations ago to 155–165 individuals at 50 generations ago.

All clusters showed a progressive increase in effective population size toward older generations, indicating a reduction in N_e over recent generations. Cluster 2 consistently exhibited the lowest N_e estimates, whereas Cluster 3 showed slightly higher values across most time points. The larger effective population size observed in Cluster1 is consistent with its lower levels of linkage disequilibrium, whereas the reduced N_e values detected in Cluster 2 and Cluster 3 agree with their comparatively higher LD levels. Overall, these results suggest contrasting demographic histories among the inferred genetic clusters, with Cluster1 retaining a substantially larger breeding population size than the other two clusters.

Runs of Homozygosity and Genomic Inbreeding

A total of 4,675 runs of homozygosity (ROH) were identified across the three genetic clusters inferred from the southern PPH population (Table 4). Cluster 1 exhibited the highest absolute number of ROH segments (3,344), followed by Cluster 3 (917) and Cluster 2 (414). Mean ROH length ranged from 6.95 Mb in Cluster 1 to 9.32 Mb in Cluster 3. Likewise, Cluster 3 accumulated the greatest total genomic length within ROH (8,548.1 Mb), followed by Cluster 1 (23,243.0 Mb) and Cluster 2 (3,175.0 Mb). A summary of ROH statistics for each genetic cluster is presented in Table 4.

Table 4
Statistics of runs of homozygosity (ROH) across southern PPH clusters

Cluster	N	Total ROH	Mean ROH/ individual	Mean ROH length (Mb)	Total ROH length (Mb)
Cluster 1	106	3344	31.55	6.95	23243
Cluster 2	14	414	29.57	7.67	3175
Cluster 3	21	917	43.67	9.32	8548

The distribution of ROH length classes differed among the three genetic clusters (Fig. 6). ROH segments between 2 and 8 Mb accounted for the largest proportion of homozygous tracts in all clusters. The 2–4 Mb class was the most abundant, representing 38.5%, 36.7%, and 32.1% of all ROH segments in Cluster 1, Cluster 2, and Cluster 3, respectively. Segments ranging from 4 to 8 Mb constituted the second most frequent class across the three clusters. In contrast, long ROH segments (> 16 Mb) were more prevalent in Cluster3 (15.8%) than in Cluster2 (11.4%) and Cluster1 (7.4%). Similarly, ROH segments between 8 and 16 Mb represented comparable proportions in Cluster2 and Cluster3 (21.7%), whereas a slightly lower proportion was observed in Cluster1 (19.6%).

Analysis of SNP frequencies within ROH segments identified several genomic regions exhibiting elevated levels of homozygosity, hereafter referred to as ROH hotspots (Fig. 7). Cluster 1 exhibited a single hotspot on chromosome 7, where SNP frequencies reached 58.5%. Cluster 2 displayed three hotspot regions located on chromosomes 7, 11, and 13, with the strongest signal detected on chromosome 7 (64.3%). In contrast, Cluster 3 showed a broader distribution of hotspots across chromosomes 1, 2, 3, 5, 6, 7, 8, 17, and 29. The most prominent hotspot was detected on chromosome 17, where SNP frequencies reached 85.7%, followed by hotspots on chromosomes 7 (66.7%), 2 (61.9%), and 6 (61.9%). A shared ROH hotspot on chromosome 7 was identified in all three genetic clusters, whereas the remaining hotspot regions were cluster-specific. Detailed genomic coordinates of all hotspot regions are provided in Supplementary Table S4.

Functional annotation of genes located within ROH hotspot regions identified several candidate genes with functions related to cellular signaling, tissue organization, skeletal development, and neuronal processes (Table 5). The complete list of annotated genes identified across all hotspot regions is provided in Supplementary Table S5. A shared ROH hotspot detected on chromosome 7 across all three genetic clusters contained genes including *PDE4A*, *TYK2*, *S1PR2*, *DNMT1*, and *ANGPTL6*, which are involved in cyclic nucleotide signaling, cytokine-mediated signal transduction, tissue organization, epigenetic regulation, and growth-related processes. Additional genes identified within this region included *CNN1*, *ELAVL3*, and *EPOR*, associated with muscle contraction, neuronal development, and erythropoiesis, respectively. Cluster specific ROH hotspots also contained genes with diverse biological functions. The chromosome 11 hotspot in Cluster 2 harbored genes such as *PECAM1* and *ERN1*, involved in cell adhesion and cellular homeostasis, whereas the chromosome 13 hotspot contained *AUTS2* and *GALNT17*, both associated with neurodevelopmental processes. In Cluster 3, hotspot regions on chromosome 3 included *IRX3*, *IRX5*, *MMP2*, and *SLC6A2*, which are related to embryonic development, extracellular matrix remodeling, and neuronal signaling. Additional genes identified in Cluster 3 included *NOTCH2* on chromosome 5, *PIK3C2G* and *PDE3A* on chromosome 6, members of the protocadherin family (*PCDH8*, *PCDH17*, and *PCDH20*) on chromosome 17, and *UCMA*, *OPTN*, and *CAMK1D* on chromosome 29 (Table 5).

Table 5
Candidate genes identified within ROH hotspots

Genetic cluster	Chromosome	Gene	Known biological function
Cluster1 / Cluster3	7	<i>PDE4A</i>	Cyclic AMP signaling pathway and regulation of cellular responses
Cluster1 / Cluster3	7	<i>TYK2</i>	Cytokine mediated signal transduction and immune response pathways
Cluster1 / Cluster3	7	<i>S1PR2</i>	Cell migration, tissue organization, and developmental signaling
Cluster1 / Cluster3	7	<i>DNMT1</i>	Maintenance of DNA methylation patterns and epigenetic regulation
Cluster1 / Cluster3	7	<i>ANGPTL6</i>	Growth regulation, angiogenesis, and energy metabolism
Cluster2	7	<i>CNN1</i>	Actin binding protein involved in muscle contraction and cytoskeletal organization
Cluster2	7	<i>ELAVL3</i>	Neuronal RNA binding protein associated with nervous system development
Cluster2	7	<i>EPOR</i>	Erythropoietin receptor involved in erythropoiesis and oxygen transport
Cluster2	11	<i>PECAM1</i>	Cell adhesion and vascular development
Cluster2	11	<i>ERN1</i>	Endoplasmic reticulum stress signaling and cellular homeostasis
Cluster2	13	<i>AUTS2</i>	Neurodevelopmental regulation and neuronal differentiation
Cluster2	13	<i>GALNT17</i>	Glycosylation related processes and neural development
Cluster3	3	<i>IRX3</i>	Embryonic patterning and developmental regulation
Cluster3	3	<i>IRX5</i>	Skeletal development and morphological patterning
Cluster3	3	<i>MMP2</i>	Extracellular matrix remodeling and connective tissue organization
Cluster3	3	<i>SLC6A2</i>	Neurotransmitter transport and neuronal signaling
Cluster3	5	<i>NOTCH2</i>	Cell differentiation and developmental signaling pathways
Cluster3	6	<i>PIK3C2G</i>	PI3K mediated signaling and metabolic regulation
Cluster3	6	<i>PDE3A</i>	Cyclic nucleotide signaling and muscle physiology
Cluster3	17	<i>PCDH8</i>	Cell adhesion and nervous system development

Genetic cluster	Chromosome	Gene	Known biological function
Cluster3	17	<i>PCDH17</i>	Neural connectivity and neuronal communication
Cluster3	17	<i>PCDH20</i>	Cell adhesion and neural tissue organization
Cluster3	17	<i>OLFM4</i>	Cell differentiation and tissue development
Cluster3	17	<i>DIAPH3</i>	Cytoskeleton organization and cellular morphology
Cluster3	29	<i>UCMA</i>	Cartilage matrix formation and skeletal development
Cluster3	29	<i>OPTN</i>	Cellular signaling and autophagy regulation
Cluster3	29	<i>CAMK1D</i>	Calcium dependent signaling involved in muscle and neuronal function

Genome-wide genomic inbreeding coefficients (FROH) differed among the three genetic clusters (Table 6). Cluster 3 exhibited the highest mean genomic inbreeding coefficient (FROH = 0.180), followed by Cluster 2 (FROH = 0.100) and Cluster 1 (FROH = 0.097). Similarly, Cluster 3 showed the highest maximum individual FROH value, indicating a wider range of genomic inbreeding estimates within this cluster. Detailed summary statistics for genomic inbreeding coefficients are presented in Table 6.

The distribution of FROH values estimated from ROH segments of different minimum length thresholds is shown in Fig. 7. Across all ROH length classes (> 2, > 4, >8, and > 16 Mb), Cluster 3 consistently exhibited higher FROH values than Cluster 1 and Cluster 2. Median FROH values decreased as the minimum ROH length threshold increased in all clusters. Greater variability in FROH estimates was also observed in Cluster 3, particularly for ROH segments longer than 2 Mb and 4 Mb (Fig. 8).

Table 6
Genomic inbreeding coefficients (FROH) in the three genetic clusters of the southern PPH clusters

Cluster	Mean FROH	Median FROH	SD	Min	Max
Cluster 1	0.097	0.095	0.037	0.021	0.195
Cluster 2	0.100	0.107	0.054	0.028	0.179
Cluster 3	0.180	0.183	0.070	0.059	0.313

Discussion

The present study provides the first genome-wide characterization of genetic diversity, population structure, linkage disequilibrium, effective population size, genomic inbreeding, and homozygosity patterns in the southern Peruvian Paso horse (PPH) using high-density SNP markers. Bayesian clustering analysis identified three genetic clusters that did not strictly correspond to geographic origin. Individuals from Arequipa, Ica, and Lima were distributed across multiple clusters, and several horses

exhibited admixed ancestry, indicating substantial genetic connectivity among breeding lineages. Consistent with the STRUCTURE results, the weak genetic differentiation observed among clusters was supported by the extensive overlap detected in the multivariate analyses and the low F_{ST} estimates. Similar patterns have been reported in managed horse breeds, where weak population subdivision coexists with detectable genetic structure generated by founder effects, selective breeding, and unequal reproductive contributions among breeding animals (Petersen et al. 2013; Criscione et al. 2022). In the PPH, this connectivity is likely maintained through the historical exchange of breeding animals and the widespread dissemination of successful breeding lineages across regions, suggesting that the observed genetic structure reflects breeding lineages rather than strict geographic subdivision.

Despite this genetic structure and the presence of genomic autozygosity, the southern PPH retained moderate levels of genetic diversity. Observed heterozygosity was similar to or slightly higher than expected heterozygosity across clusters, resulting in negative F_{IS} values and indicating no substantial deficit of heterozygotes. Similar patterns have been reported in several horse breeds, where moderate diversity is maintained despite founder effects, selective breeding, and localized signatures of selection (Grilz-Seger et al. 2019; Metzger et al. 2015). Likewise, ROH-based studies have shown that managed horse populations can retain substantial genetic variation while simultaneously exhibiting genomic regions affected by autozygosity and selection (Grilz-Seger et al. 2019; Metzger et al. 2015). Although Cluster 3 exhibited higher FROH values and a greater abundance of long ROH segments, all clusters maintained slightly negative F_{IS} estimates. Similar discrepancies between heterozygosity-based and ROH-based measures of inbreeding have been described in livestock populations, reflecting the ability of ROH analyses to detect recent autozygosity that may not be captured by heterozygosity-based metrics alone (Purfield et al. 2012; Ferenčaković et al. 2013). Overall, these results suggest that the southern PPH has maintained substantial genetic variation despite the accumulation of autozygous genomic regions in specific lineages.

Patterns of linkage disequilibrium (LD) and historical effective population size (N_e) revealed contrasting demographic histories among the three inferred genetic clusters. Cluster 2 consistently exhibited the highest LD levels, whereas Cluster 1 showed the lowest values and Cluster 3 displayed an intermediate pattern. Similar LD decay profiles have been reported in other domestic horse populations and are generally interpreted as signatures of differences in demographic history, genetic drift, and breeding structure (Criscione et al. 2022; Corbin et al. 2012; Mousavi et al. 2023). Historical effective population size estimates were highly concordant with the LD patterns. Cluster 1 maintained the largest effective population size throughout the analyzed period, whereas Cluster 2 exhibited the lowest N_e values and Cluster 3 showed intermediate estimates. Because effective population size reflects the balance between genetic drift and the maintenance of genetic diversity (Charlesworth 2009), the combination of elevated LD and reduced N_e in Cluster 2 suggests a more restricted breeding pool and stronger effects of genetic drift. Conversely, the lower LD and larger N_e observed in Cluster 1 are consistent with a broader breeding base and greater historical genetic connectivity. Similar relationships between linkage disequilibrium, effective population size, and demographic history have been reported in other horse

populations analyzed using genome-wide SNP data, where founder effects, genetic drift, and breeding practices contributed to differences in genomic diversity and inbreeding levels (Criscione et al. 2022; Mousavi et al. 2023).

Runs of homozygosity analyses revealed contrasting patterns of genomic inbreeding among the three inferred genetic clusters. Although the southern Peruvian Paso horse retained moderate levels of overall genomic diversity, Cluster 3 exhibited substantially higher levels of autozygosity than the remaining clusters. Similar levels of ROH-derived genomic inbreeding have been reported in managed horse breeds subjected to selective breeding and the repeated use of influential stallions, where localized autozygosity may accumulate despite the maintenance of moderate genome-wide diversity (Grilz-Seger et al. 2019; Metzger et al. 2015). Analysis of ROH length classes indicated the coexistence of both historical and recent autozygosity signals within the breed. Intermediate-length ROH segments predominated across all clusters, suggesting the influence of shared ancestry and historical demographic processes. However, Cluster 3 displayed a greater abundance of long ROH segments, which are generally associated with more recent common ancestry and recent inbreeding events (Ceballos et al. 2018; Curik et al. 2014). This pattern contrasts with the lower effective population size observed in Cluster 2, indicating that different demographic processes have shaped the genetic architecture of the inferred clusters. Whereas Cluster 2 appears to reflect a more restricted breeding pool at the population level, Cluster 3 is characterized by a greater accumulation of recent autozygosity within specific lineages.

One of the most notable findings of this study was the identification of a shared ROH hotspot on chromosome 7 across the inferred genetic clusters. The persistence of this hotspot across genetically differentiated clusters suggests the existence of a conserved autozygous block that has been maintained throughout the breeding history of the PPH, potentially through selection, demographic processes, or the recurrent inheritance of shared haplotypes (Pemberton et al. 2012; Peripolli et al. 2017). Such regions are of particular interest because they may harbor variants associated with traits that have been consistently favored during breed development.

Functional annotation of the chromosome 7 hotspot revealed genes involved in epigenetic regulation (DNMT1), growth and tissue remodeling (ANGPTL6), intracellular signaling (PDE4A), and tissue organization (S1PR2) (Oike et al. 2005; Houslay 2010; Moore et al. 2013; Adada et al. 2013). The co-localization of these genes within a high-frequency ROH island shared across all genetic clusters suggests that chromosome 7 may represent a conserved regulatory region maintained throughout the breeding history of the breed. Interestingly, this hotspot does not overlap with the best-known locomotion associated region on chromosome 23 containing DMRT3, which has been identified as a major genomic region associated with gait expression in several gaited horse breeds (Andersson et al. 2012). Instead, the PPH hotspot contains genes associated with epigenetic regulation, tissue remodeling, and cellular signaling, suggesting that additional genomic regions may have contributed to the development and maintenance of breed specific characteristics in this population. Beyond chromosome 7, cluster specific ROH hotspots contained genes associated with musculoskeletal development, extracellular matrix remodeling, cartilage biology, and neuronal function. Notably, Cluster 3

hotspots harbored genes involved in developmental regulation (IRX3, IRX5, and NOTCH2), connective tissue remodeling (MMP2), cartilage maintenance (UCMA), and neuronal signaling (SLC6A2, PCDH8, PCDH17, and PCDH20). Functional annotation based on Ensembl gene annotations indicated that these genes participate in biological processes related to tissue development, extracellular matrix organization, skeletal maintenance, and neuronal function (Kinsella et al. 2011; Page-McCaw et al. 2007; Canalis 2013). Likewise, Cluster 2 contained AUTS2, a gene implicated in neurodevelopment and neuronal differentiation (Oksenberg and Ahituv 2013). Collectively, these findings suggest that different breeding lineages may have accumulated distinct genomic regions affecting biological pathways relevant to skeletal development, connective tissue biology, cartilage maintenance, and neuronal regulation. Although functional validation is required, these ROH islands may represent genomic regions influenced by historical selection, lineage specific breeding practices, or the accumulation of favorable haplotypes associated with breed defining characteristics.

The occurrence of distinct ROH hotspots among genetic clusters further supports the existence of partially different demographic and breeding histories. While the chromosome 7 hotspot likely reflects a conserved genomic region shared across the breed, Cluster 3 exhibited a larger number of hotspot regions and greater genomic autozygosity, whereas Cluster 2 was characterized by higher linkage disequilibrium and lower effective population size. In contrast, Cluster 1 retained the highest genomic diversity and exhibited only a single major hotspot. Together, these patterns indicate that the contemporary genetic architecture of the southern PPH has been shaped by both shared breed history and the differential propagation of haplotypes among breeding lineages, a process commonly observed in managed livestock populations subjected to founder effects, unequal reproductive contributions, and long term selection (Ceballos et al. 2018; Grilz-Seger et al. 2019; Peripolli et al. 2017). Future studies integrating genomic and phenotypic information will be necessary to determine whether these regions contribute to variation in locomotion, conformation, or other performance related traits in the breed.

Overall, the identification of three genetic clusters with contrasting patterns of linkage disequilibrium, effective population size, genomic inbreeding, and homozygosity provides new insights into the demographic history of the southern PPH. The detection of a conserved ROH hotspot on chromosome 7, together with cluster specific genomic regions enriched for genes involved in musculoskeletal and neuronal functions, highlights genomic regions that may have contributed to the development of breed defining traits. These findings provide a valuable framework for future studies and support the implementation of conservation and breeding strategies aimed at maintaining the genetic sustainability of the breed.

Conclusions

This study represents the first genome-wide characterization of genetic diversity and population structure in the southern Peruvian Paso horse using high-density SNP markers. The results revealed the presence of three genetic clusters with moderate differentiation, contrasting patterns of linkage disequilibrium, effective population size, and genomic inbreeding, but without evidence of severe loss of

genetic diversity at the population level. Runs of homozygosity analyses identified both shared and cluster-specific hotspot regions, including a conserved hotspot on chromosome 7 present across all genetic clusters. The genes located within these regions were associated with developmental, musculoskeletal, and neuronal functions, suggesting that historical breeding practices may have contributed to the differential accumulation of genomic regions potentially relevant to breed-defining characteristics. Overall, these findings improve our understanding of the demographic history and genomic architecture of the Peruvian Paso horse and provide genomic information that may be useful for the design of future conservation and breeding programs.

Declarations

Ethics approval and consent to participate

Not applicable.

Animal Ethics and Consent to Participate declarations

Not applicable

Consent for publication

Not applicable.

Availability of data and materials

The dataset generated and analyzed during the current study has been deposited in Zenodo and is available to reviewers during the peer-review process through this link: <https://tinyurl.com/5y6328cc>. The dataset will be made publicly available upon publication of the manuscript or at the end of the embargo period.

Competing interests

The authors declare no competing interests.

Funding

We thank the project “Integrando la genómica, IoT y Data Mining para la caracterización del andar del Caballo Peruano de Paso” (Grant ID N° 501096382-2025-PROCIENCIA) Programa Nacional de Investigación Científica y Estudios Avanzados, PROCIENCIA, and project “Creación del servicio de investigación y enseñanza en Equinos en la Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas” (Grant ID CUI N° 2314510) of the Peruvian Government

Authors' contributions

C.L.S.: Designed the study, sample collection, sample processing, analyzed the data, writing-original draft, writing-review & editing, L.Q. H.: Writing-original draft, writing-review & editing, L.M.: Sample collection, sample processing, writing-original draft, writing-review & editing, J.L.M: Writing-original draft, writing-review & editing, C.I.A.: Designed the study, sample collection, sample processing, analyzed the data, writing-original draft, writing-review & editing, W.B: Designed the study, sample collection, writing-original draft, writing-review & editing . All authors discussed the methodologies, results, and read and approved the manuscript

Acknowledgements

We thank the National Association of Peruvian Paso Horse Breeders (ANPCPP) for its continued support and collaboration. We are also grateful to Stefany Lobato for her assistance with figure preparation and editing, and to Maili Muñoz for her support with project logistics. Finally, the authors acknowledge the High-Performance Computing Cluster (UNTRM–DATA SCIENCE) of the Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), the Bioinformatics High-Performance Computing Server of the Universidad Nacional Agraria La Molina (UNALM), and the University of Wisconsin–Madison for providing the computational resources used in this study.

References

1. Adada M, Canals D, Hannun YA, Obeid LM (2013) Sphingosine-1-phosphate receptor 2. *FEBS J* 280:6354–6366
2. Andersson LS, Larhammar M, Memic F, Wootz H, Schwochow D, Rubin CJ et al (2012) Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. *Nature* 488:642–646
3. Bazvand B, Rashidi A, Zandi MB, Moradi MH, Rostamzadeh J (2024) Genome-wide analysis of population structure, effective population size and inbreeding in Iranian and exotic horses. *PLoS ONE* 19:e0299109
4. Biscarini F, Cozzi P, Gaspa G, Marras G (2019) detectRUNS: Detect runs of homozygosity and heterozygosity in diploid genomes. R package version 0.9.6.
5. Canalis E (2013) Notch signaling in skeletal development. *Endocr Rev* 34:282–303
6. Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF (2018) Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet* 19:220–234
7. Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:195–205
8. Corbin LJ, Liu AYH, Bishop SC, Woolliams JA (2012) Estimation of historical effective population size. *Genet Sel Evol* 44:9
9. Criscione A, Mastrangelo S, D'Alessandro E, Tumino S, Di Gerlando R, Zumbo A et al (2022) Genome-wide survey on three local horse populations with a focus on runs of homozygosity pattern. *J Anim Breed Genet* 139:540–555

10. Curik I, Ferenčaković M, Sölkner J (2014) Inbreeding and runs of homozygosity. *Livest Sci* 166:26–34
11. Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* 4:359–361
12. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* 14:2611–2620
13. Ferenčaković M, Sölkner J, Curik I (2013) Estimating autozygosity from high-throughput information. *Genet Sel Evol* 45:42
14. Francis RM (2017) pophelper: an R package and web app to analyse and visualize population structure. *Mol Ecol Resour* 17:27–32
15. Gonzales R, Li R, Kemper G, Del Carpio C, Ruiz E (2018) An algorithm for estimating the variation of the joint angles of the limbs of Peruvian Paso horse. In: 2018 IEEE XXV International Conference on Electronics, Electrical Engineering and Computing (INTERCON). Piscataway: IEEE; pp. 1–4
16. Grilz-Seger G, Druml T, Neuditschko M, Horna M, Ricard A, Mesarič M et al (2019) High-resolution runs of homozygosity and inbreeding parameters in Noriker horses. *Arch Anim Breed* 62:279–288
17. Houslay MD (2010) Underpinning compartmentalised cAMP signalling. *Trends Biochem Sci* 35:91–100
18. Kamvar ZN, Tabima JF, Grünwald NJ (2014) poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281
19. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G et al (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011:bar030
20. Librado P, Khan N, Fages A, Kusliy MA, Suchan T, Tonasso-Calvière L et al (2021) The origins and spread of domestic horses from the Western Eurasian steppes. *Nature* 598:634–640
21. Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28:298–299
22. McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM et al (2012) A high density SNP array for the domestic horse and extant Perissodactyla: utility for association mapping, genetic diversity, and phylogeny studies. *PLoS Genet* 8:e1002451
23. Metzger J, Schrimpf R, Philipp U, Distl O (2015) Expression levels of LCORL are associated with body size in horses. *PLoS ONE* 10:e0122758
24. MINCETUR (2012) Caballo Peruano de Paso: producto bandera del Perú. Ministerio de Comercio Exterior y Turismo del Perú, Lima
25. Moore LD, Le T, Fan G (2013) DNA methylation and its basic function. *Neuropsychopharmacology* 38:23–38
26. Mousavi SF, Razmkabir M, Rostamzadeh J, Seyedabadi HR, Naboulsi R, Petersen JL et al (2023) Genetic diversity and signatures of selection in four indigenous horse breeds of Iran. *Heredity* 131:96–108

27. Nicodemus MC, Clayton HM (2003) Temporal variables of four-beat, stepping and non-stepping gaits of gaited horses. *Appl Anim Behav Sci* 80:133–142
28. Oike Y, Akao M, Kubota Y, Suda T (2005) Angiopoietin-like proteins. *Trends Mol Med* 11:473–479
29. Oksenberg N, Ahituv N (2013) The role of AUTS2 in neurodevelopment and human evolution. *Trends Genet* 29:600–608
30. Page-McCaw A, Ewald AJ, Werb Z (2007) Matrix metalloproteinases and the regulation of tissue remodelling. *Nat Rev Mol Cell Biol* 8:221–233
31. Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528
32. Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ (2012) Genomic patterns of homozygosity. *Am J Hum Genet* 91:275–292
33. Peripolli E, Stafuzza NB, Munari DP, Lima ALF, Irgang R, Machado MA et al (2017) Assessment of runs of homozygosity islands. *BMC Genomics* 18:34
34. Petersen JL, Mickelson JR, Cothran EG, Andersson LS, Axelsson J, Bailey E et al (2013) Genetic diversity in the modern horse illustrated from genome-wide SNP data. *PLoS ONE* 8:e54997
35. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
37. Purfield DC, Berry DP, McParland S, Bradley DG (2012) Runs of homozygosity and population history in cattle. *BMC Genet* 13:70
38. R Core Team (2024) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
39. Saldaña CL, Justo S, Murga L, Vásquez HV, Maicelo JL, Arbizu CI, Bardales W (2025) Mitochondrial genome assembly of the Peruvian Paso horse through PacBio long-read sequencing. *Sci Rep* 16:85
40. Vilela Velarde JL, Quintana Dolores PGM, Velarde Marcos ML, Dextre Chacon JC (2025) Correlations among morphometric traits, functional performance, and gluteal temperature in the Peruvian Paso horse. *Austral J Vet Sci* 57:e570116
41. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358–1370
42. Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-, New York

Figures

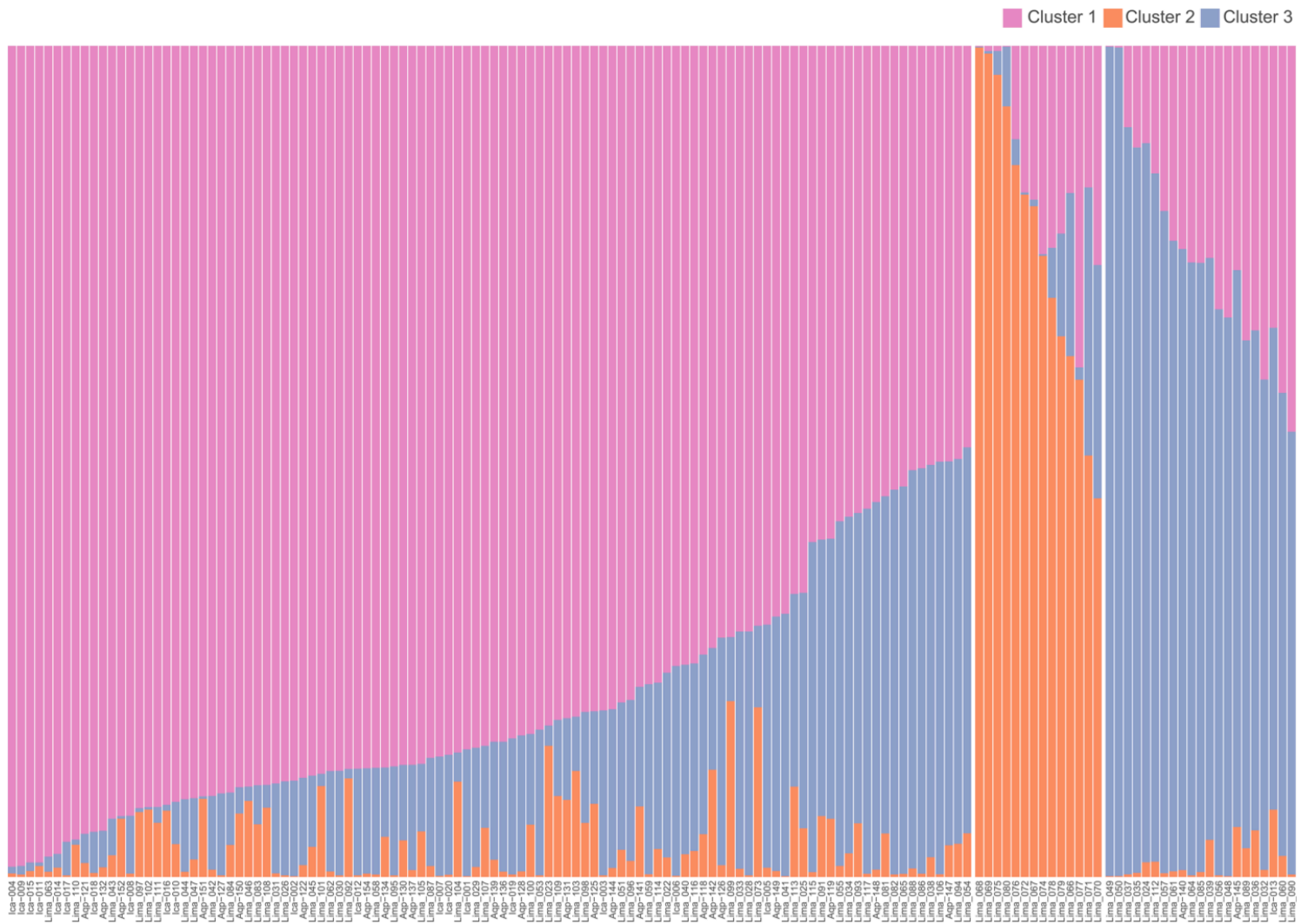


Figure 1

Population structure of the southern PPH inferred. Each vertical bar represents an individual horse, and colors indicate the proportion of genomic ancestry assigned to each inferred genetic cluster. The presence of admixed individuals reflects genetic connectivity among breeding lineages

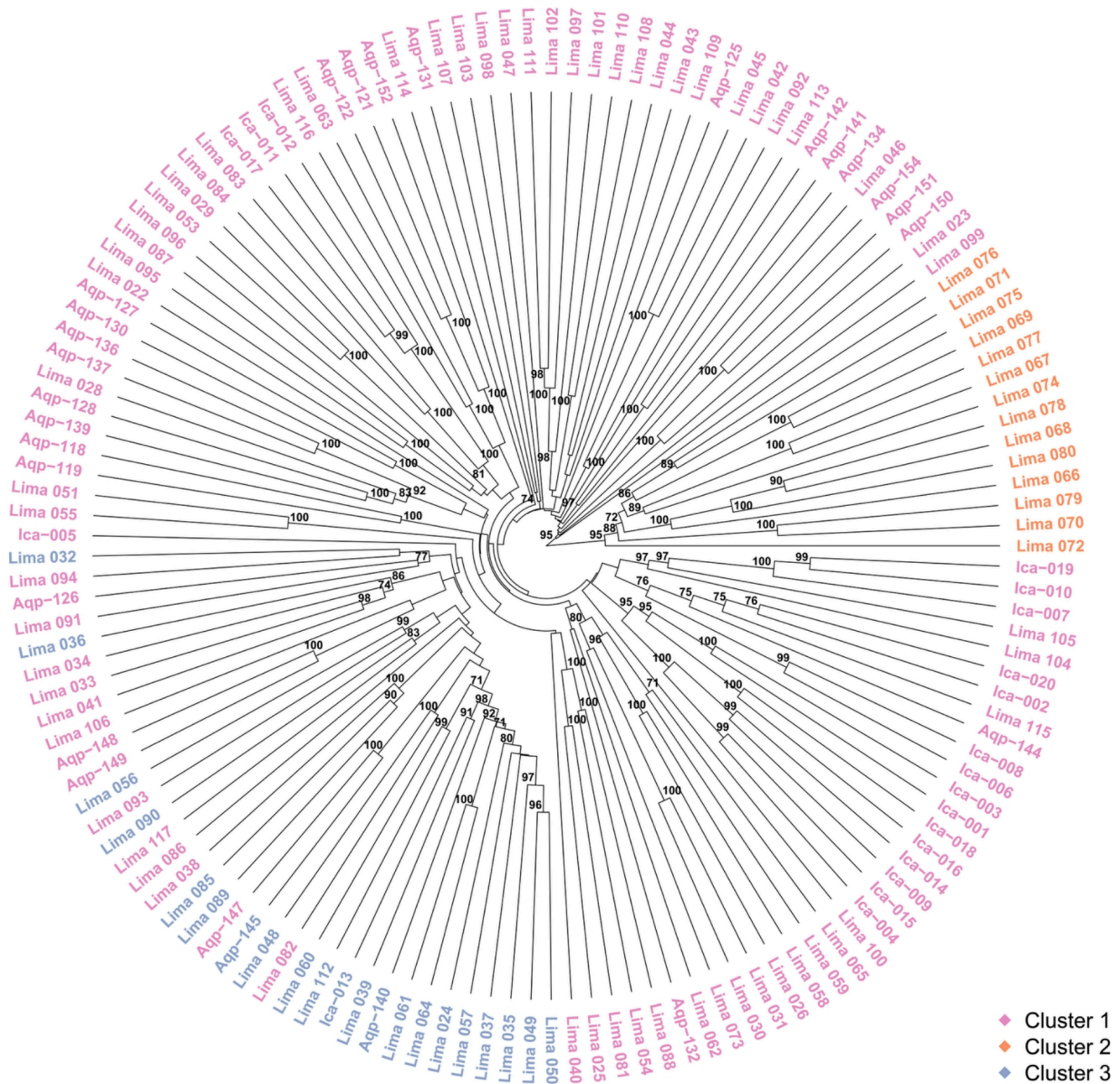


Figure 3

UPGMA dendrogram showing the genetic relationships among 141 southern Peruvian Paso horses based on Nei's genetic distance. Tip colors represent the three genetic clusters identified ($K = 3$). Bootstrap support values were estimated from 1,000 replicates, and only values above 70% are shown

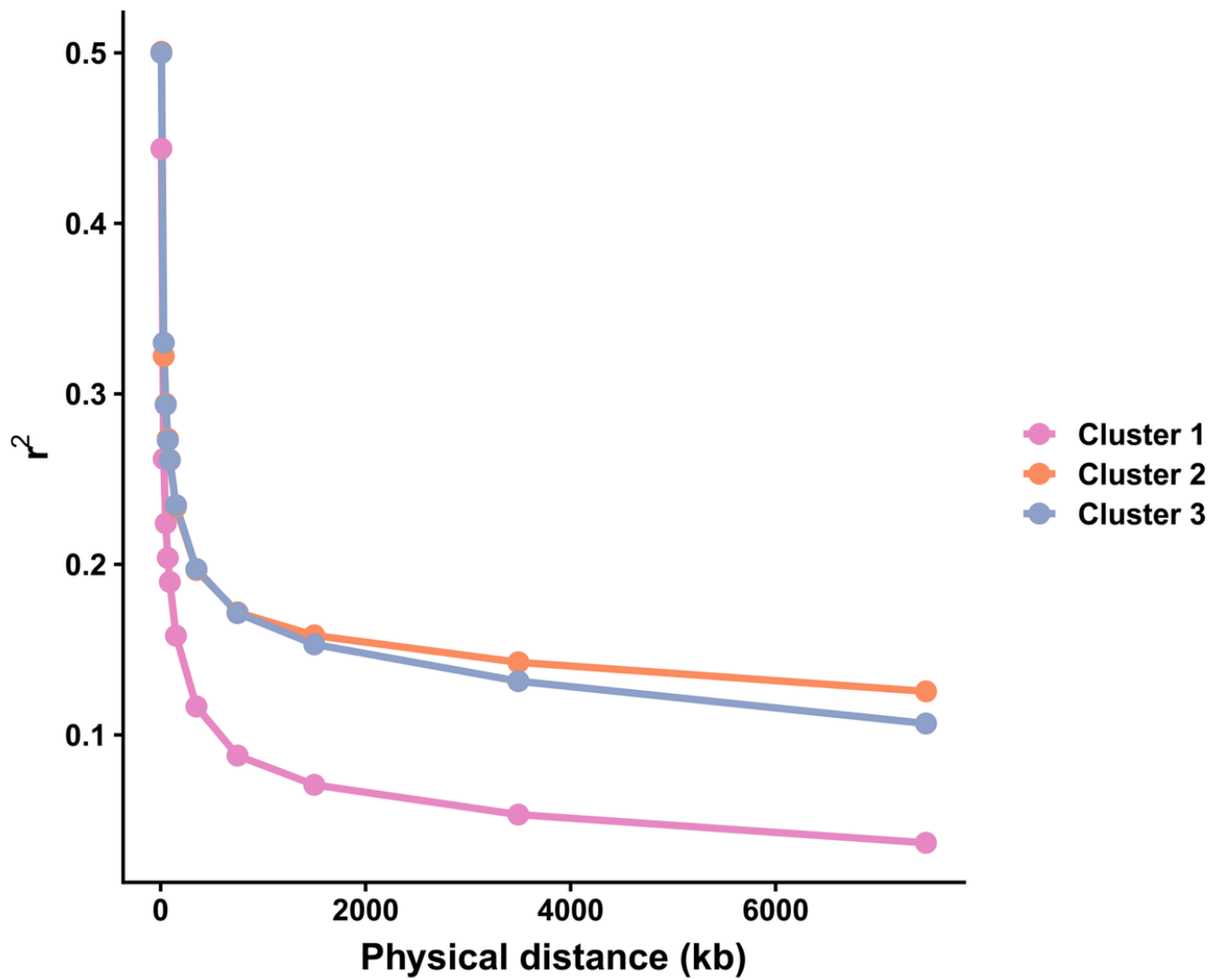


Figure 4

Linkage disequilibrium decay in southern PPH clusters

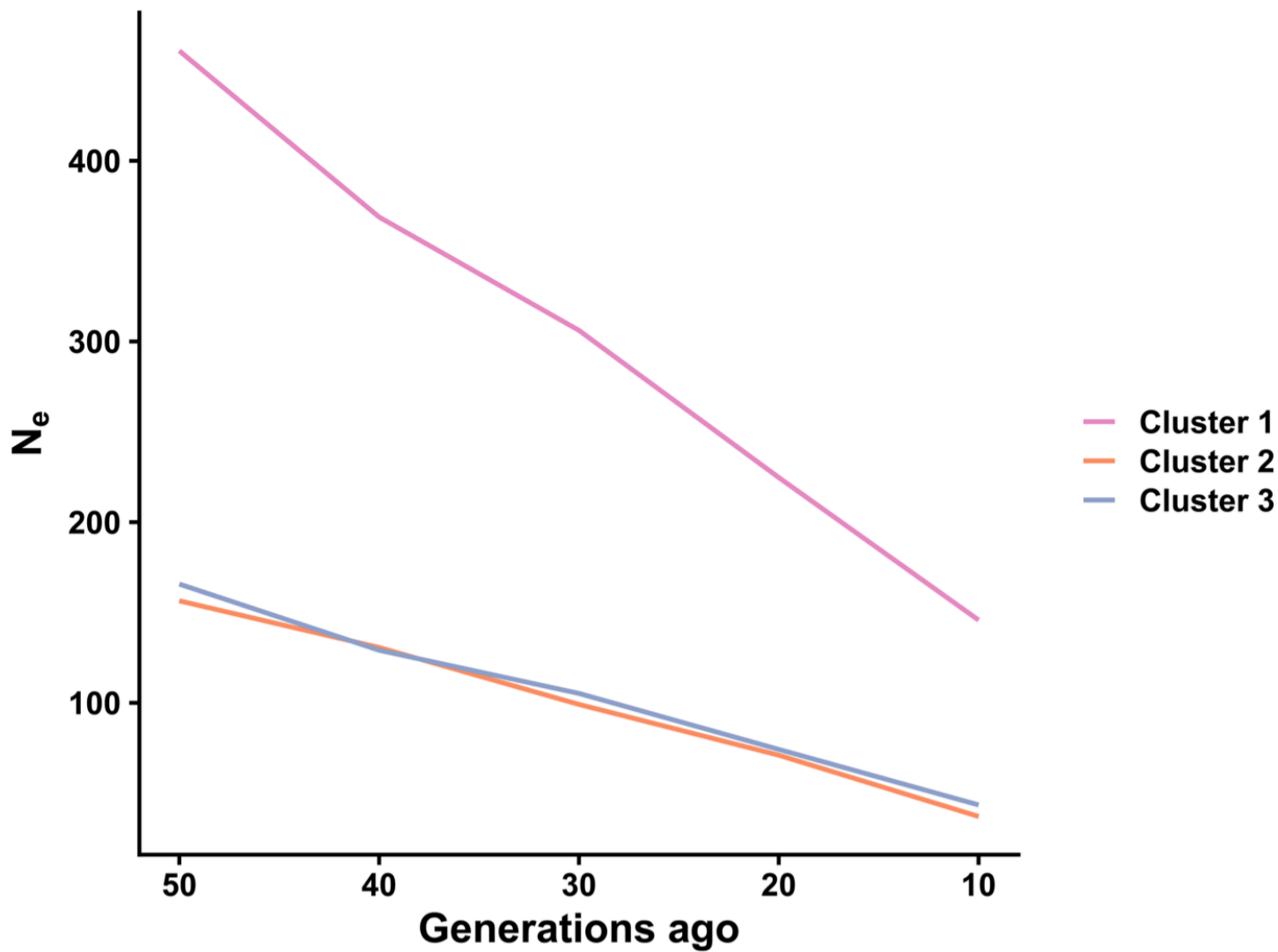


Figure 5

Historical effective population size (N_e) inferred from linkage disequilibrium patterns for the three genetic clusters identified ($K = 3$) in southern PPH clusters

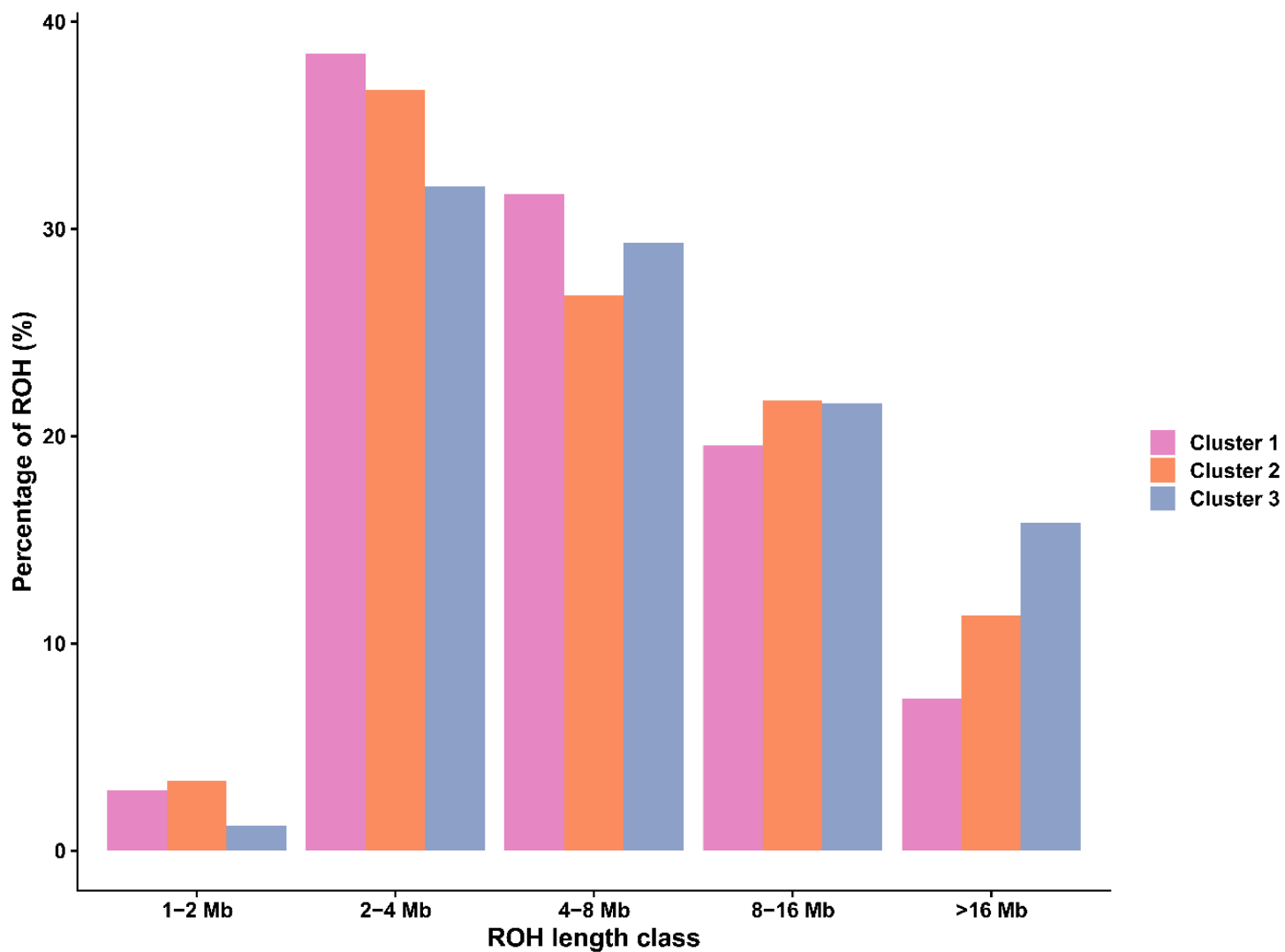


Figure 6

Distribution of runs of homozygosity (ROH) length classes among the three genetic clusters inferred from population structure analysis

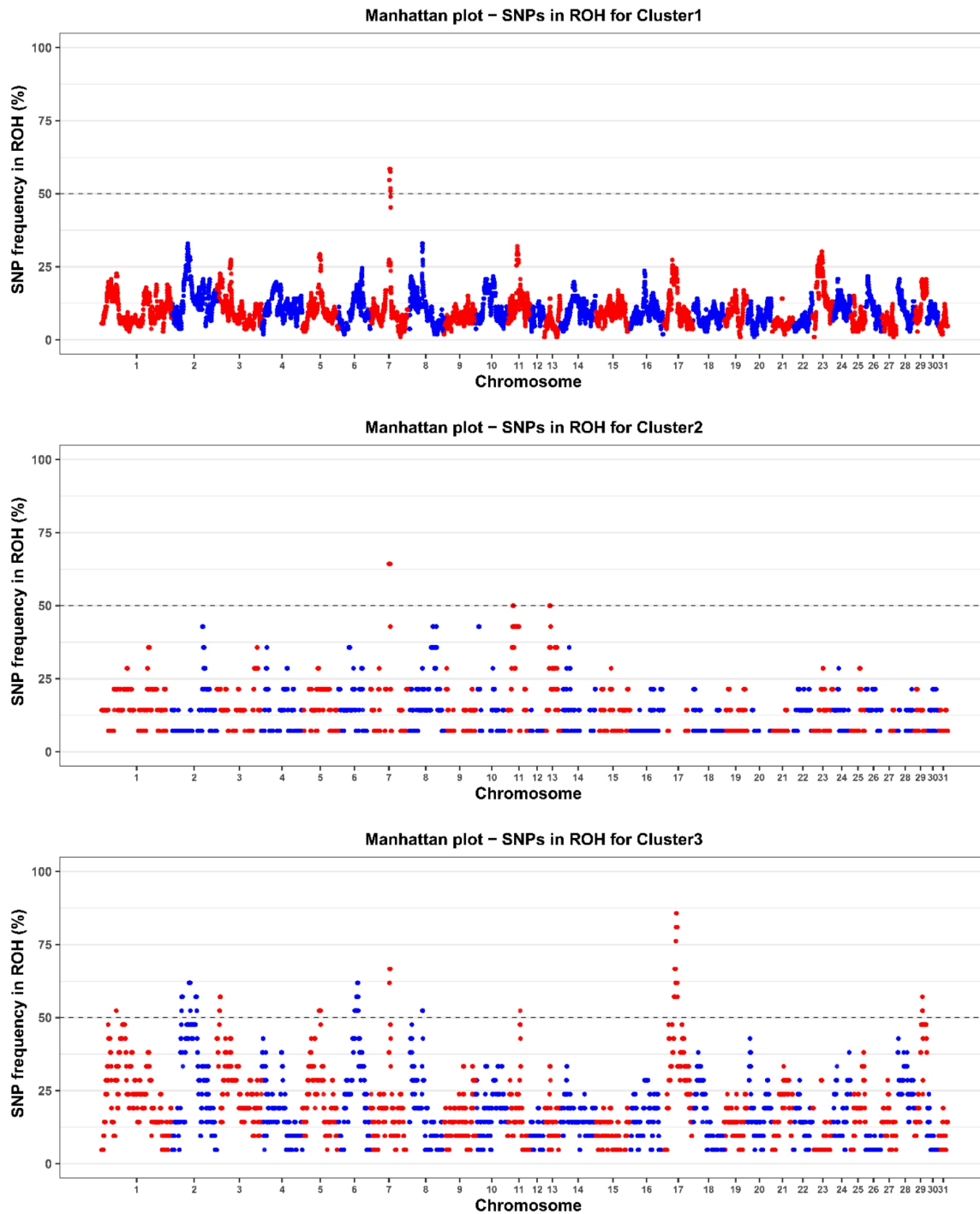


Figure 7

Manhattan plots showing the frequency of SNP occurrence within ROH segments across the genome for Cluster 1, Cluster 2, and Cluster 3. The dashed horizontal line indicates the 50% threshold used to define ROH hotspot regions

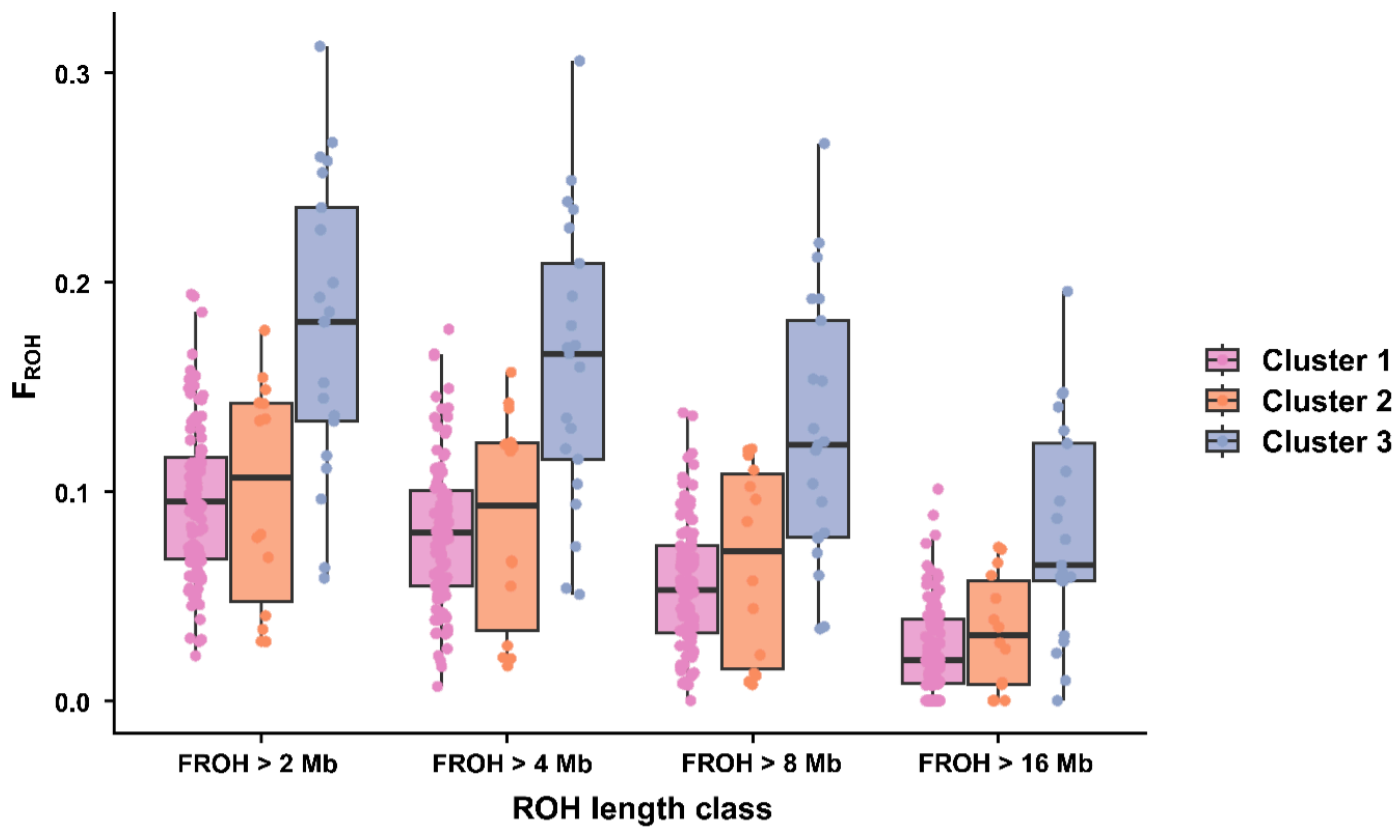


Figure 8

Distribution of genomic inbreeding coefficients (F_{ROH}) across ROH length thresholds in the three genetic clusters of the southern PPH population.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SaldaaetalSupplementalMaterial.zip](#)