American Genetic Association

Journal of Heredity, 2018, 1–10
doi:10.1093/jhered/esy061
Original Article
Advance Access publication December 27, 2018

OXFORD

Original Article

Genetic Diversity of Persian Arabian Horses and Their Relationship to Other Native Iranian Horse Breeds

Raheleh Sadeghi, Mohammad Moradi-Shahrbabak, Seyed Reza Miraei Ashtiani, Florencia Schlamp, Elissa J. Cosgrove, and Doug F. Antczak

From the Department of Animal Science, College of Agriculture and Natural Resources, University of Tehran, Karaj, Iran (Sadeghi, Moradi-Shahrbabak, and Miraei Ashtiani); the Baker Institute for Animal Health, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853 (Sadeghi and Antczak); and the Department of Molecular Biology & Genetics, Cornell University, Ithaca, NY 14853 (Schlamp and Cosgrove).

Address correspondence to M. Moradi-Shahrbabak at the address above, or e-mail: moradim@ut.ac.ir.

Address correspondence also to D. F. Antczak at the address above, or e-mail: dfa1@cornell.edu.

Corresponding Editor: Ernest Bailey

Received April 6, 2018; First decision October 28, 2018; Accepted November 13, 2018.

Abstract

The principal aims of this study were to explore genetic diversity and genome-wide selection signatures in Persian Arabian horses and to determine genetic relationship of Persian Arabians with other Iranian horse breeds. We evaluated 71 horses from 8 matrilineal strains tracing to 47 mares from the mid to late 19th century, using the equine 670k single nucleotide polymorphism (SNP) BeadChip. Mean observed and expected heterozygosity were (0.43) and (0.45), respectively, average inbreeding measures (inbreeding estimates based on runs of homozygosity and pedigree information) were low, indicating high genetic diversity in Persian Arabian horses. Analysis of population genetic structure using STRUCTURE and principal component analysis suggested that Persian Arabian horses can be divided into 3 groups, however the groups do not match traditional matrilineal strains. In total, 15 genomic regions were identified by at least 2 of the 3 implemented methods, Tajima's D, H, and H12, as potentially under selection in Persian Arabian horses. Most of these peaks were found on chromosome 9, overlapping with QTLs previously associated with horse temperament. Biological function analysis of identified candidate genes highlighted enrichment of GO term "response to lipopolysaccharide" and KEGG pathway "chemokine-mediated signaling pathway," which are associated with immune responses and may have been targets of selection in Persian Arabian horses. Independent analyses of SNP data from 30 horses of 4 other Iranian breeds suggested distinct population structure between Persian Arabian, and Turkemen and Caspian horse breeds. Overall, the results of this study suggest a rich genetic diversity in the Persian Arabian horses and a clear genetic differentiation with Turkemen and Caspian breeds.

Subject areas: Population structure and phylogeography, Molecular adaptation and selection
Keywords: genetic diversity, Persian Arabian horse, ROH, selective sweeps, SNP, structure

The horse has played an important role in the history of Iran for many centuries (Fotovati 2000), but the adoption of mechanization has reduced the need for horses, and subsequently, their numbers have declined. The Persian Arabian, Caspian, Turkemen, Dareshuri, and Kurdish horse breeds have long-standing husbandry, cultural, and historic importance in Iran (Firouz 1998; Fotovati 2000). Because some of these breeds are considered endangered, it is important to know the state of their genetic diversity. Based on a study of mitochondrial DNA (mtDNA), an ancient origin of the Persian Arabian and Turkemen breeds has been proposed (Moridi et al. 2013). Only the Arabian horse is found widely around the world, with populations in roughly 60 countries and a global census estimated to be more than 1 million horses (World Arabian Horse Organization). Earliest evidence suggests that the Arabian may have been a recognized breed as long as 3000 years ago (Knorr 1912), with scattered small breeding populations in many countries in the Middle East, including Egypt, Saudi Arabia, Syria, and Iran. These isolated breeding populations led to the development of matrilineal strains, in which pedigrees were maintained via oral history. Different phenotypic attributes have been associated with these strains, including height, head shape, and color (Forbis 1976), but there is little evidence for an underlying genetic basis for them. In fact, studies based on mitochondrial DNA (Khanshour and Cothran 2013; Almarzook et al. 2017b) and single nucleotide polymorphism (SNP) analysis (Almarzook et al. 2017a) failed to show concordance between traditional strains designations and molecular markers in several Arabian horse populations.

According to the registered studbook, there are around 1300 purebred Arabian horses in Iran, with an average of 100 foals registered annually. These Persian Arabian horses have not been interbred with other populations of Arabians for several generations. The Persian Arabian studbook includes 9 matrilineal strains: Khersani, Koheilan, Hamdani, Saglawi, Obayan, Moangi, Showayman, Nasmani, and Djelfan. Some of these strains have low registry numbers including Moangi and Showayman, and others are found only in Iran (Khersani and Nasmani). In recent years, Iran has begun importing Arabian horses from other parts of the world, and consequently, the native pure Persian Arabian population is at risk from cross-breeding with imported stock. The inbreeding status of the native strains is not known. A conservation program based on principles of population genetics applied to a closed population could be formulated using information on genetic diversity within or across various strains of Persian Arabian horses to guide future breeding decisions.

Several studies have been conducted in different Iranian native horse breeds addressing the genetic diversity, parentage verification, and the genetic structure of population based on microsatellite markers or mtDNA. Most of these studies were carried out on horses from only a single breed: Caspian (Seyedabadi et al. 2006; Amirinia et al. 2007; Shahsavarani and Rahimi-Mianji 2010), Kurdish (Ala-Amjadi et al. 2017), Turkemen (Rahimi-Mianji et al. 2015), Persian Arabian (Gharahveysi and Irani 2011; Moshkelani et al. 2011), and only one study has been conducted in multiple breeds (Moridi et al. 2013). All studies reported a high genetic diversity in Iranian native horses. To date, there have been no genome-wide SNP-based studies of genetic diversity in Persian Arabian horses. Here, we characterized the level of diversity and putative loci under selection in Persian Arabian horses using the equine SNP 670K platform (Schaefer et al. 2017). In addition, we studied 30 samples from Turkemen (11), Caspian (7), Dareshuri (5), and Kurdish horses (7) to determine the population structure of other native horse breeds of Iran and to evaluate the genetic relationships among these Iranian breeds.

Materials and Methods

Horses and Samples

A total of 101 blood samples were collected from the external jugular vein of horses representing 5 breeds: Persian Arabians (n = 71), Turkemen (n = 11), Caspian (n = 7), Kurdish (n = 7), and Dareshuri (n = 5). This sample collection followed the Cornell Institutional Animal Care and Use Committee protocol #1986-0216. The 71 Persian Arabian samples were collected from 8 different traditional matrilineal strains and 5 Iranian provinces (Isfahan, Karaj, Yazd, Khuzestan, and Kerman). For the Persian Arabians, we sampled horses that were less related and from different bloodlines, as determined by examination of pedigrees in the published studbook of Persian Arabian horses. Using this pedigree information, we collected samples from 65 farms in 5 Iranian provinces. Most sampled horses were not related in the first and second generations. However, the population of Persian Arabian horses is small, and because of this some of the horses were related as paternal half-siblings. We included a small number of such horses when we could not sample the dams of the half-siblings. Other horse breeds samples were collected from Isfahan and Tehran (Turkemen horse), Rasht (Caspian horse), Yazd (Dareshuri horse), and Tehran (Kurdish horse) provinces of Iran. The genomic DNA was extracted using the phenol-chloroform method.

Pedigree Analysis

Data from the pedigrees of horses registered in the Persian Arabian studbook, containing information from 1952 to 2014, were provided by the Iranian Equestrian Federation and the Iran Asil Association. For the early records, only a small proportion of the population was registered. Quality of the pedigree was evaluated using the depth of the pedigree parameter, which was examined using the CFC program (Sargolzaei 2006) by computing the number of generations from the base population to the reference population, assuming each generation preceded discretely (Woolliams and Mäntysaari 1995).

Genotyping and Quality Control

Genotyping was conducted at Affymetrix (Santa Clara, CA) using the Equine SNP670K BeadChip (on Axiom® MNEc670 Arrays), assaying 670 796 single nucleotide polymorphisms (SNPs) (Schaefer et al. 2017). After genotyping, the variant list restricted to PolyHighResolution cluster type with a FLD ≥3.6.

Sample and marker-based quality control were performed using PLINK software (Purcell et al. 2007). SNPs and samples were filtered based on genotyping rate (<100%), minor allele frequency (MAF < 0.05), and Hardy–Weinberg equilibrium (HWE P-value < 10^{-6}). Diversity estimates are sensitive to the ascertainment biases and exclusion of SNPs in linkage disequilibrium (LD) has been reported to minimize the effect of these biases (Malomane et al. 2018). Therefore, SNPs were pruned in PLINK using the indep-pairwise command (SNP window size: 50, SNPs shifted per step: 5, r^2 threshold: 0.5) to minimize the effect of these biases. These filters were applied to several iterations of the sample set: full Persian Arabian sample set (99 483 SNPs and 71 samples), a nonrelated (IBD < 0.25) Persian Arabian sample set (105 341 SNPs and 51 samples), full multibreed sample set (126 262 SNPs and 101 samples), and a nonrelated multibreed sample set (140 072 SNPs and 77 samples). The only exception for multibreed sample sets was that we did not apply HWE filtering.

Genetic Diversity

Estimation of observed heterozygosity (H_O) per horse, within strain, was calculated for full Persian Arabian sample set in Arlequin V3.5.2.2 and compared with the expected heterozygosity (H_E). The pairwise identity-by-descent (IBD) estimates were calculated using the same data set and - - genome flag in PLINK (Purcell et al. 2007).

PGDSpider version 2.0.8.2 (Lischer and Excoffier 2012) was used to convert PLINK files to Arlequin format. The analysis of molecular variance (AMOVA) was used to estimate the variance between strains and among genotypes within strains for full Persian Arabian sample set in Arlequin V3.5.2.2 using 20 000 permutations (Excoffier et al. 2007). Fixation indexes and number of polymorphic loci were also calculated using the same sample set in Arlequin (Excoffier et al. 2007).

Principal Component Analysis

To examine relationships between individual samples, strains, and breeds, we applied principal component analysis (PCA) using EIGENSOFT version 5.0.1 (Price et al. 2006). PCA was conducted for both the full Persian Arabian sample set, and the full multibreed sample set.

Structure Analysis

For the analysis of population structure, a Bayesian clustering analysis was performed with the software STRUCTURE 2.3.4 (Pritchard et al. 2000). Following the findings of Rodriguez-Ramilo and Wang (2012), we removed closely related individuals (IBD > 0.25) before applying Structure. We applied STRUCTURE to 5 randomly thinned SNP subsets (20 000 SNPs selected using - - thin flag in PLINK) of both the nonrelated Persian Arabian and the nonrelated multibreed sample sets, for number of clusters K = 1–10 and K = 1–8, respectively. Three replicates were run for each value of K, using BURNIN = 20 000 and NUMREPS = 30 000 in all runs. Results from each thinned set were nearly identical (data not shown), and a representative set was used for results presented herein.

To determine the optimal value of K, we considered the K value corresponding to maximum mean estimated log probability of data (Pritchard et al. 2000), and also the K value selected using the ΔK statistic (Evanno et al. 2005) using STRUCTURE HARVESTER (Earl and vonHoldt 2012). We used the software CLUMPP (Jakobsson and Rosenberg 2007) to summarize Structure results across replicate runs and plotted results in R.

Inbreeding Coefficients

Two measures of inbreeding were calculated for all Persian Arabian horses. Runs of homozygosity (ROH) was estimated according to the method described in Druml et al. (2017). The SNPs on sex chromosomes and SNPs with unknown chromosome position were omitted for this analysis. SNPs were filtered by more than 10% missing genotypes and we did not apply MAF filtering to consider all homozygous SNPs, as done in Druml et al. (2017).

ROH segments along the genome were calculated using the filtered SNP data in PLINK (Purcell et al. 2007) based on the following parameters: minimum SNP density was set to one SNP per 50 kb, with a maximum gap length of 100 kb. Therefore, the final segments were considered as ROH, if the minimum length of the homozygous segment was greater than 500 kb and comprised more than 80 homozygous SNPs, while 1 heterozygote and 2 missing genotypes were permitted within each segment, as implemented by Druml et al. (2017). The total number of ROH, length of ROH (in Mb), and the sum of all ROH segments (in Mb) of each horse were calculated for the Persian Arabian horses. The ROH segments (LROH) were also divided into the following 5 length groups: 0.5–1, 1–2, >2–4, >4–6, and >6–8 Mb.

The inbreeding coefficient based on ROH (FROH; previously described for horse; Metzger et al. 2015), was estimated as the sum of the length of all ROH per horse as a proportion of the total genome length across autosomes covered by SNPs in the Equine 670K SNP chip (~2.24 Gb).

Pedigree-based inbreeding coefficients (FPED) were computed for the 71 Persian Arabian horses based on the full pedigree (1560 horses) using the package "pedigree" in R (R Core Team 2017). Finally FPED and FROH were compared using linear regression and Pearson correlation coefficients, across all 71 horses by R software (R Core Team 2017).

Effective Population Size (N_e)

To estimate effective population size based on the LD method, we followed the filtering criteria as described in Corbin et al. (2010) and Lee et al. (2014). This filtering excludes markers which deviated from HWE (P < 0.0001), markers which were genotyped in less than 95% of samples, individuals with more than 10% of SNPs missing, markers with low MAFs (with a 0.10 threshold), and markers with the physical distances less than 100 bp. Furthermore, only autosomal markers were used because sex chromosomes have different determinism (Purcell et al. 2007). This filtering resulted in a data set of 250 568 SNPs and 71 samples.

Effective population size (N_e) was estimated using the software SNeP version 1.1 (Barbato et al. 2015). The recombination rate was calculated using the Sved and Feldman's mutation rate modifier (Sved and Feldman 1973) and sample size correction was considered for unphased genotypes. Linkage distance between markers was estimated using the assumption of a genome-wide linear relationship such that 1 cM =1 Mb. Minimum and maximum inter-SNP distances of 50 000 and 40 000 000 bp, respectively, were used.

We also performed the "N_e slope analysis" (N_eS) to look into the rate and directionality of N_e changes occurring in generations which helps to identify the subtle changes in the inferred N_e curve when the changes in N_e plot is not clear, as done in Pitt et al. (2018).

Determination of Genomic Regions under Selection in Persian Arabian Horses

In contrast to other analyses herein, the only SNP filters applied before selection scan analysis were to require less than 2% missingness per SNP and less than 5% missingness per sample, resulting in a data set of 71 samples and 319 477 SNPs. This data set was then phased using SHAPEIT with consideration of the pedigree information (Delaneau et al. 2014). The ends of chromosomes 1, 2, 6, and 20 had higher density of SNPs compared to the rest of the genome, so they were excluded from the selection scan analysis. They were trimmed at positions 145, 73.5, 31.5, and 28.7 Mb, respectively. This left a total of 262 647 SNPs used in selection scan analysis.

To identify genomic regions potentially under selection in Persian Arabian horse population, we ran genome-wide scans using 3 popular statistics: Tajima's D (Tajima 1989), H statistic (Schlamp et al. 2016), and H12 (Garud et al. 2015). Using a combination of these statistics allows targets of selection to be identified with greater precision. One methods is based on SNP frequencies (Tajima's D) while the other 2 are based on haplotype frequencies (H statistic and H12). Tajima's D (TD) compares estimators of the number of segregating

sites (s) in a population sample with levels of heterozygosity (π) to detect genomic regions that have a significant excess of low- or high-frequency SNPs when compared with neutral expectations (Tajima 1989). *H* statistic measures the average length of pairwise haplotype homozygosity tracts around each SNP. The *H* statistic was estimated using the program H-SCAN (version 1.3), downloaded from: https://messerlab.org/resources/ (April 2015), following the methods described in Schlamp et al. (2016). H12 and TD were all calculated over windows of a fixed number of SNPs on our genotyping chip (251 and 351) and the estimated values of each statistic were then assigned to the position of the center SNP of the window, as done in Schlamp et al. (2016) and as suggested in Voight et al. (2006). In Schlamp et al. (2016), the window sizes were 25, 51, 101, and 201 neighboring sites in their dog chip, which had a total of ~153k informative sites. In our horse chip we have ~262.6k informative sites, therefore we empirically increased the number of neighboring sites in the windows to account for the higher SNP density in our chip.

Gene Ontology Analysis

First, we filtered the list of candidate regions under selection to only those identified by at least 2 of the 3 selection scan methods, and that passed a length cutoff (<10 Mb). The latter filter was applied to exclude broad regions where the selection signal is less specific. For the remaining regions, Ensembl gene IDs were retrieved with the Ensembl genome browser using the *Equus Caballus* genome assembly EquCab2. Gene ontology (GO) and KEGG pathway analysis were performed on the candidate gene sets using the functional annotation cluster tool from the database for annotation, visualization and integrated discovery (DAVID) v6.8 (Huang Da et al. 2009) to determine significantly enriched biological functions or processes positively selected in Persian Arabian horses. To identify KEGG enriched clusters, we used an enrichment score of 1.3, equal to Fisher exact test *P*-value of 0.05, as a threshold as suggested by the software, and as done in Bahbahani et al. (2017). In addition, the Equine QTL animal database (http://www.animalgenome.org) was used to find the overlapping QTLs with the candidate genome regions.

Results

Pedigree Structure

The pedigree records of Persian Arabian horses constitute a total of 4393 horses born between 1952 and 2014. The current population of Persian Arabian horses is approximately 1300. Pedigrees of these horses were traced back to the earliest recorded ancestors in the mid-1800s. In the current population of Persian Arabians the following matrilineal lines are represented: Khersani (n = 476), Koheilan (n = 256), Hamdani (n = 180), Saglawi (n = 108), Obayan (n = 94), Nasmani (n = 91), Djelfan (n = 57), Moangi (n = 18), and Showayman (n = 8). In this study, we genotyped 71 Persian Arabians of 8 strains that could be traced by pedigree in the direct maternal line to 47 mares in the mid-19th century (Supplementary Table S1). An average yield of 481.49 (ng/μL) of DNA with a 260/280 ratio of 1.81 and a 260/230 ratio of 2.03 were obtained using the phenol-chloroform extraction DNA protocol. These values show high concentration and purity of extracted DNA as measured using NanoDrop Spectrophotometers.

SNP Polymorphism and Genetic Diversity Metrics

DNA from the 71 Persian Arabians in this study was applied to the 670k SNP chip and analyzed for population genetic diversity metrics using Arlequin software. The overall mean expected and observed heterozygosity was 0.45 and 0.43, respectively, for all Persian Arabian horses. The average relatedness, based on pairwise IBD between all Persian Arabian horses was small, with an overall average value of 2.49% (Table 1). The overall percentage of polymorphic loci was 85.63% and a high number of polymorphic loci were observed for most strains, except for Nasmani and Moangi strains with small sample size.

The AMOVA for the 8 populations (strains) revealed that the majority of the genetic variation was found within individuals (Vc = 97.61%, FIT = 0.024), rather than among individuals within populations (Vb = 1.59%, FIS = 0.016), and among populations (Va = 0.8%, F_{ST} = 0.008) (Supplementary Table S2). The overall F_{ST} value among all strains was low (0.008) and not significant (P value = 1).

Population Structure

Principal Component Analysis

PCA was applied to examine relationships between strains and between individuals. The first 2 principal components (PCs) for the full Persian Arabian sample set, including 8 Persian Arabian horse strains, are plotted in Figure 1, with PC1 and PC2 explaining 3.64% and 2.14% of the total variance, respectively. We did not observe separation of different strains on the PCA plot, but the plot highlighted a few potential outlier individuals in the upper left corner of the plot. These 4 individuals (3 Koheilan and 1 Moangi) are related with mean IBD = 0.18, and share a common ancestor within 2–4 generations from the Hetli substrain of Persian Arabian horses.

We also applied PCA to the full multibreed sample set, which included 5 different Iranian horse breeds. Plotting the first 2 principal components, we observed clear separation of the Persian Arabian

Table 1. Number of samples (*N*), expected and observed heterozygosity (H_E and H_o) (mean ± standard deviation), polymorphic loci, and mean pairwise identity-by-descent (IBD) for different Persian Arabian horse strains using the full data set (71 samples and 99 483 SNPs)

Strain	*N*	H_E	H_o	Polymorphic loci	IBD (%)
Khersani	26	0.35 ± 0.14	0.34 ± 0.16	99 130 (99.65%)	3.94
Koheilan	15	0.37 ± 0.13	0.36 ± 0.16	98 383 (98.89%)	2.25
Hamdani	11	0.37 ± 0.14	0.36 ± 0.18	96 025 (96.52%)	3.06
Saglawi	6	0.40 ± 0.13	0.40 ± 0.20	92 051 (92.53%)	1.65
Obayan	5	0.41 ± 0.13	0.40 ± 0.21	87 156 (87.61%)	1.52
Djelfan	3	0.47 ± 0.11	0.47 ± 0.25	78 514 (78.92%)	2.40
Nasmani	3	0.46 ± 0.11	0.47 ± 0.25	68 303 (68.66%)	11.71
Moangi	2	0.56 ± 0.08	0.51 ± 0.29	61 985 (62.31%)	0.00
All	71	0.45 ± 0.11	0.43 ± 0.23	85 193 (85.63%)	2.49

horses from the Turkemen and Caspian horses, while there was some overlap between the Persian Arabian, Kurdish, and Dareshuri horses (Figure 2).

Structure

To infer the population structure from genotype data among the 8 strains of Persian Arabian horses, we used the STRUCTURE program package (Figure 3). Both maximum mean estimated log probability of data [LnP(D)] and Evanno's Δ*K* method supported the optimal number of clusters *K* = 3 (Supplementary Figure S1). Supplementary Table S3 presents the proportion of each strain belonging to each of the 3 clusters. At least 50% of Khersani, Koheilan, Djelfan, Hamdani, Saglawi, and Obayan strains were assigned to cluster 1, while 78% of Nasmani and 51% of Moangi were in clusters 2 and 3, respectively. When we labeled each sample by predominant STRUCTURE cluster assignment and replotted the PCA figure reflecting these labels, we observed concordance between STRUCTURE and PCA results (Supplementary Figure S2). We also applied STRUCTURE to the multibreed sample set. Evanno's Δ*K* method suggested optimal number of clusters *K* = 4, but maximum mean estimated LnP(D) was nearly equivalent for both *K* = 3 and *K* = 4, so we considered results for both of these *K* values. Using either value of *K*, we found that the Caspian and Turkemen horses clustered separately from the Persian Arabian, Dareshuri, and Kurdish horses (Supplementary Figure S3). This is consistent with PCA results for the multibreed sample set. The proportions of each breed belonging to each of the clusters are reported in Supplementary Table S4.

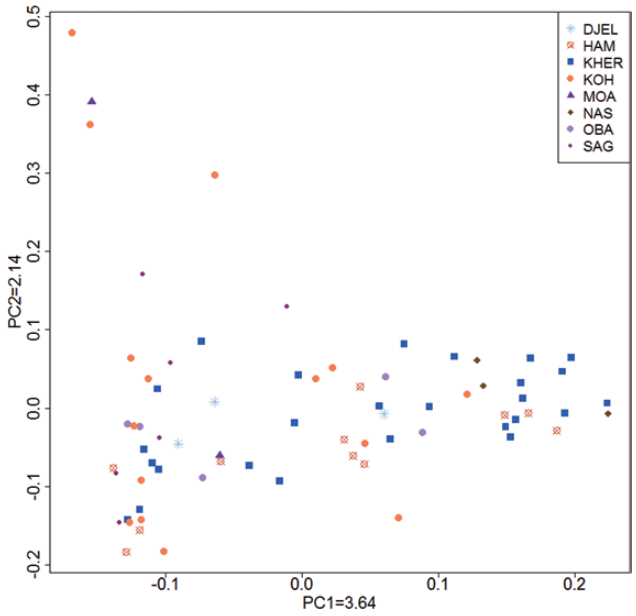

Figure 1. Principal component analysis, PC1 (3.64%) and PC2 (2.14%) for different strains of Persian Arabian horses (DJEL: Djelfan, HAM: Hamdani, KHER: Khersani, KOH: Koheilan, MOA: Moangi, NAS: Nasmani, OBA: Obayan, SAG: Saglawi).

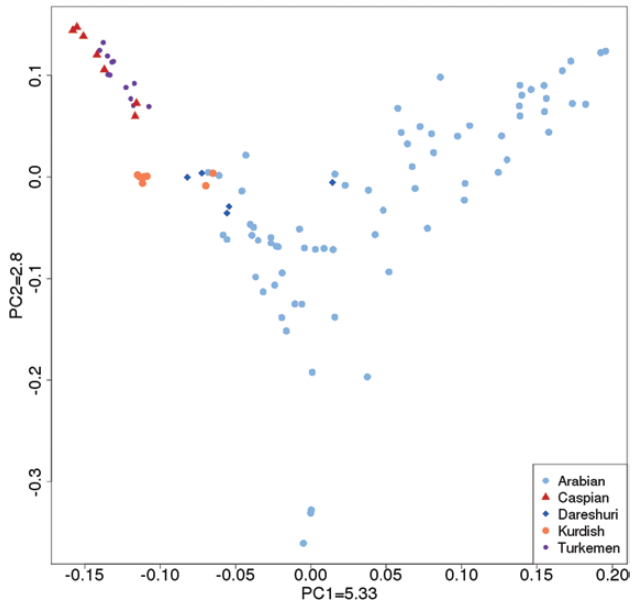

Figure 2. Principal component analysis, PC1 (5.33%) and PC2 (2.8%) for Iranian horse breeds.

Correlation between ROH and Inbreeding Coefficient

A total of 71 samples and 320 857 SNPs covering overall length 2 241 026 899 bp of the genome were used to estimate ROH in Persian Arabian horses. The total genome length covered by ROH (Mean LROH) comprised 317.5 Mb (max. 762.4 Mb; min. 141.9 Mb), the mean number of ROHs (nROH) was 164.9 per horse (max. 214 ROHs; min. 110 ROHs), and the F_{ROH} value was 14.16% (max. 34.02; min. 6.33). The main part (71.83%) of ROH segments in our sample set of Persian Arabian horses had a mean length between 1 and 2 Mb, followed by 2–4 (23.94%). We also calculated FROH considering ROHs longer than 2 Mb. The mean FROH for ROHs longer than 2 Mb was 22.47% (max. 34.02; min.15.65).

Given that the pedigree-based inbreeding coefficients depend on the quality of the pedigree information, we first assessed the quality of the pedigree based on the average number of discrete generation equivalents. This value was 6.79 with maximum of 9.93, which shows good pedigree completeness. A comparison of the ROH and pedigree-based estimates of inbreeding demonstrated a clear linear relationship (R = 0.75; Supplementary Figure S4). The intercept of the regression of F_{ROH} on the pedigree inbreeding coefficient was greater than zero (0.10), suggesting that the pedigree-based inbreeding coefficient may underestimate the levels of ancestral genomic relatedness.

Effective Population Size (N_e)

As shown in Figure 4, the effective population size of Persian Arabian horses declined over time. N_e was estimated to be about 113 horses one generation ago. The estimate of N_e at 75 generations ago was about 539 horses (Figure 4). To investigate the change in slope of the inferred N_e obtained from LD-based, we used the new method, "NeS," which offers more detailed information about population changes 1–75 generations ago (Figure 5). The values above 0 means that with respect to the previous generation, the N_e increases. This analysis highlighted a marked reduction in N_e about 6 generation ago in Persian Arabian horses.

Determination of Genomic Regions under Selection

To capture genomic regions under selection in Persian Arabian horses, we applied genome-wide scan using 3 statistics (Tajima's *D*, *H*, and H12). Figure 6 summarizes the results of the 3 estimated statistics in Persian Arabian horses. In total, we identified 29 regions

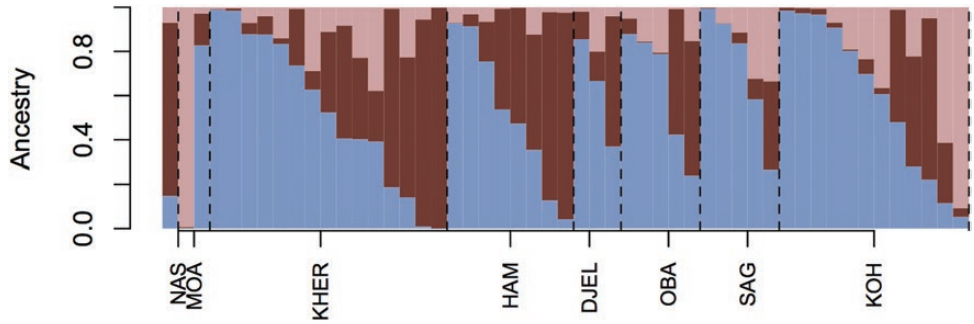

Figure 3. Inferred population structure for 8 Persian Arabian horse strains using Structure software. Structure bar plot for the genetic clusters representing ($K = 3$) using filtered SNPs data (99 483 SNP data) (DJEL: Djelfan, HAM: Hamdani, KHER: Khersani, KOH: Koheilan, MOA: Moangi, NAS: Nasmani, OBA: Obayan, SAG: Saglawi).

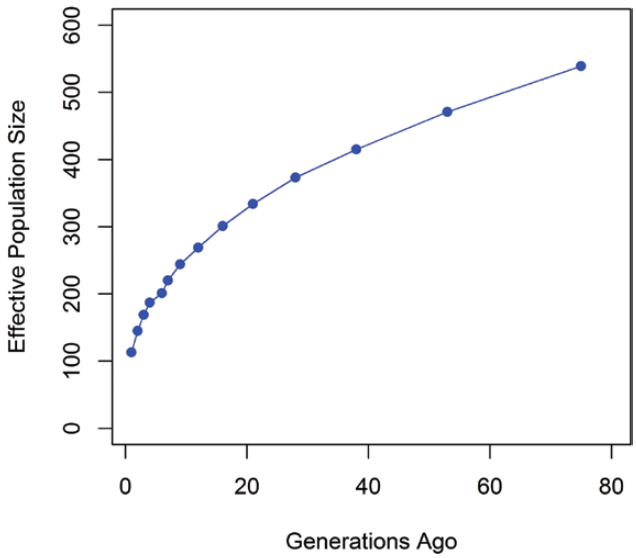

Figure 4. Plots of effective population size change between 1 and 75 generations in Persian Arabian horse population.

of the horse genome under selection using a 0.99 quantile threshold. Fourteen of these regions were identified with only one statistic, while 15 other regions were identified by at least 2 of the 3 statistics. These putative selection signature regions are represented mostly on chromosome 9. Table 2 shows top 2 identified regions by each statistic and corresponding window size, and how all these regions overlap with regions identified by at least one other statistic. This table shows, for example, how the first and second highest H peaks observed were on chromosome 9 (28 584 615–43 112 754 bp) and chromosome 11 (25 122 493–32 197 635 bp), and how the peak in chromosome 9 was also identified by H12 (d = 251 and 351) and Tajima's D (d = 251 and 351), while the peak in chromosome 11 was also identified by H12 (d = 251 and 351) and Tajima's D (d = 351). The positional information of all candidate selected regions and list of genes found in each region are shown in Supplementary Table S5. Within the candidate regions obtained from 3 statistics (H, H12, and Tajima's D), a total of 832 unique candidate genes were retrieved from Ensembl genes database (Supplementary Table S5). Some of candidate regions under selection overlapped with known horse QTLs, such as temperament in chromosome 9 (Table 2 and Supplementary Table S5). We further investigated the functions associated with the genes found within putative regions under selection by analyzing over-represented annotations and pathways using DAVID 6.8 (https://david.ncifcrf.gov/). Fourteen regions spanning 382 genes were considered in the DAVID analysis, which identified 17 significantly enriched terms ($P < 0.05$), including 9 functional terms for biological processes (BP), 6 for molecular function (MF), and 2 for cellular component (CC) (Supplementary Table S6). Most of the functional terms were involved in regulation of metabolic processes such as regulation of cell proliferation, and cellular response to steroid hormone stimulus. Some of the significantly enriched terms were also involved in immune response, such as response to lipopolysaccharide and chemokine-mediated signaling pathways. Additionally, we identified 6 KEGG clusters (Supplementary Table S7) with enrichment score greater than 1.3 and P <0.05 relative to the whole equine genome. These include enriched clusters associated with "Chemokine signaling pathway," and "Hippo signaling pathway."

Discussion

The artificial selection applied to livestock species by humans during the process of domestication and afterward has changed the livestock genome and as a result also altered the genetic variation that determines phenotypic differences (Andersson and Georges 2004). Arabian horse populations have been subjected to distinct breeding objectives and classified in a traditional system based on strains. With the availability of a new SNP genotyping platform (Schaefer et al. 2017), the aim of this study was to determine the genetic diversity, population structure, and selection signatures of the Persian Arabian horse population and compare with other Iranian horse breeds using the Equine SNP670k platform.

Observed heterozygosity (H_o) ranged between 0.34 and 0.51, expected heterozygosity (H_E) between 0.35 and 0.56, indicating high genetic diversity in the matrilineal strains of Persian Arabian horses that we studied. These results are in agreement with a recently reported study related to genetic diversity in 3 different strains of Arabian horses from Syria (Almarzook et al. 2017a), for which the observed heterozygosity in Hamdani (0.30) was also lower than Saglawi (0.32) strain. Moridi et al. (2013) reported a high matrilineal diversity in Persian Arabian horses. This diversity was attributed to the dispersed distribution of the horses in different parts of Iran, and to distinctive selection pressures. The authors further hypothesized that the diversity was indicative of an ancient origin of this breed. Using the same sample set reported here, we have also found a very high number of major histocompatibility complex (MHC) haplotypes in the Persian Arabian population, determined using a

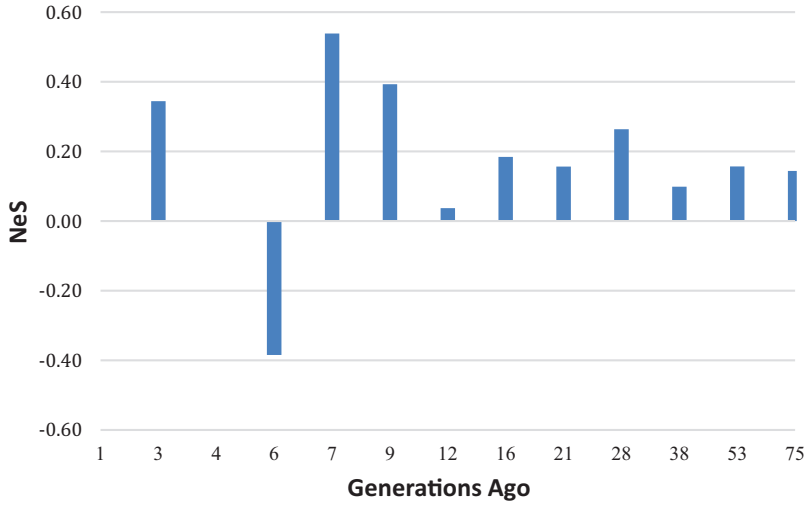

Figure 5. N_e slope analysis between 1 and 75 generations ago.

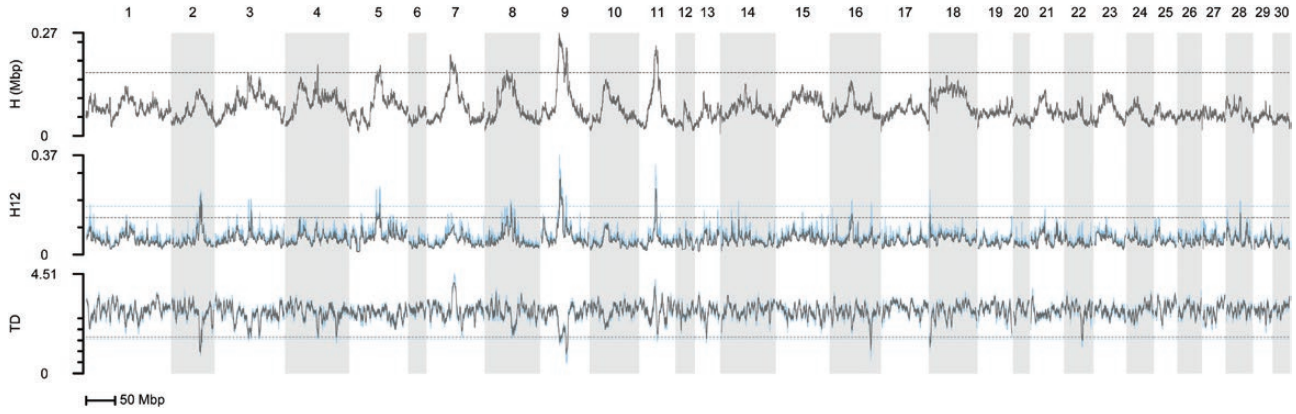

Figure 6. Plots of single population selection statistics *H,* H12, and Tajima's *D* (TD) in Persian Arabian horse breed. For H12 and TD the blue and gray lines show the results for a window size of 251 and 351 SNPs, respectively. Higher values of *H* and H12, and lower values of TD show the signals of positive selection. The horizontal dashed lines (one for each windows size) show the value cutoff at 99% quantile, which was estimated genome-wide at each statistic and window size.

panel of highly polymorphic intra-MHC microsatellites (Sadeghi et al. 2017).

The population structure of the Persian Arabian horse population was analyzed using both PCA and STRUCTURE. The PCA results matched the AMOVA results, with no clear separation for different strains. Our results agree with the previous studies (Bowling and Ruvinsky 2000; Khanshour and Cothran 2013; Hudson 2017) reported no evidence of clear subdivisions for different traditional maternal-based strain classification system using whole mtDNA D-loop sequence in Arabian horses.

It has been suggested that using high number of markers, STRUCTURE can correctly infer the number of subpopulations in a data set when clusters were not well-differentiated and genetic differentiation among groups is low (Colonna et al. 2009). The STRUCTURE for different strains of Persian Arabian horses resulted in 3 different clusters, and horses of the different strains were scattered across the different clusters. We observed concordance between STRUCTURE and PCA results when we labeled each sample based on predominant STRUCTURE cluster assignment. The results of the analysis of STRUCTURE underlines the admixture and relationship among the Persian Arabian strains that was suggested by the PCA. This demonstrates that these strain-based populations have high levels of gene flow, which is due to a mating system in which stallions are selected independently of their strain origin. Our results are in agreement with a previously published study of different strains of Syrian Arabian horses, in which different strains did not cluster separately, and STRUCTURE analysis identified 3 clusters in their sample set (Almarzook et al. 2017a). It would be of interest in the future to compare the horses from Syria and Iran in a single experiment to determine whether the 3 clusters are the same or different in horses from the 2 countries.

The extent and frequency of ROH have been used to infer ancestry of an individual and the breed population history. Long ROH results from recent inbreeding, whereas short ROH capture ancient inbreeding which derived from old ancestors and ancient haplotypes (Kirin et al. 2010). Here, we estimated ROH and compared ROH-based measures of inbreeding with pedigree-based inbreeding coefficients. The ROH estimates reported in this study were less than previously published values from a study of Purebred Arabian horses of French origin (LROH: 317.5 Mb vs. 396.501; nROH: 164.9 vs. 278.5 Mb; FROH: 14.16% vs. 17.7%, for Persian Arabian and Purebred Arabian horses, respectively; Druml et al. 2017), although the genotyping array and methods were similar. Furthermore, 72% of the Persian Arabian horses sampled in this study carry short ROH

Table 2. Table reporting 2 most significant regions identified by statistic as being under selection

Statistic	Chr	Position	Overlap with	No. genes	QTL ID and trait
H	9	28 584 615–43 112 754	H12.251, H12.351, TD.251, TD.351	76	
	11	25 122 493–32 197 635	H12.251, H12.351, TD.351	65	29289, 2990, 37892, "insect bite hypersensitivity"
H12.251	9	31 381 847–39 454 008	H, H12.351, TD.251, TD.351	0	
	11	26 648 659–30 382 384	H, H12.351, TD.251	0	29289, "insect bite hypersensitivity"
H12.351	9	30 572 930–39 000 381	H, H12.251, TD.251, TD.351	0	
	11	27 378 597–29 834 074	H, H12.251	0	
TD.251	9	42 885 399–46 609 413	H, H12.251, H12.351, TD.351	1	119813, "temperament"
	16	68 777 382–70 834 284	H12.351, TD.351	15	
TD.351	9	42 518 790–46 825 887	H, H12.251, H12.351, TD.251	1	119813, "temperament"
	2	47 564 199–50 908 212	H12.251, H12.351, TD.251	0	

These regions overlap with some regions that have been identified by other statistics, as reported in the "Overlap with" column. For each region, we report the number of genes found within, as well as the QTL ID and trait associated with the region when applicable.

with average length of runs 1–2 Mb, which is higher than the value reported for French origin Purebred Arabians (26.3%), indicate the possible ancient demographic history of the Persian Arabian horses (Druml et al. 2017). The high correlation we observed here between the 2 estimates for inbreeding coefficient (pedigree and ROH-based inbreeding) suggests that the fraction of the genome under ROH can be used to infer inbreeding and the breed population history.

The N_e for the Persian Arabian horses was estimated to be 113 animals one generation ago and showed a downward trend from 75 generations ago. This trend is consistent with artificial selection by breeders with an increased focus on particular bloodlines and stallions in recent generations, which can pose a danger for the genetic diversity of the Persian Arabian population, and needs to be examined routinely. African horse sickness which is a highly fatal disease in horses appeared in the summer of 1959 in Iran and caused the death of many Arabian horses (Capinera 2004). This could be a reason for a highly reduced number of Persian Arabian horses about 6 generations ago, based on the change in slope of the inferred N_e.

Tajima's *D*, *H*, and H12 were used to identify both hard and soft sweeps in Persian Arabian horses (Garud et al. 2015). Applying these tests to our data revealed 29 regions under selection with overlaps between some tests. We investigated the candidate genes, QTLs, and biological pathways within the candidate regions under selection which passed the length cutoff (<10 Mb) and were identified by at least 2 of the 3 methods, to find selective forces which might have shaped the Persian Arabian horse genome. Some of the regions identified under selection overlap with previously reported QTLs in horse such as behavior (temperament), reproduction (male fertility), and coat color (white markings) traits (http://www.animalgenome.org). Functional annotation analyses of genes found within candidate regions identified enriched GO term "response to lipopolysaccharide" and KEGG pathway "chemokine-mediated signaling pathway," which are associated with immune responses. These results highlight that genes and QTLs within these categories might be targeted by selection in Persian Arabian horses to adapt to environmental conditions.

In addition to Persian Arabian, Iran is home to other native horse breeds, including the Caspian, Turkemen, Kurdish, and Dareshuri horses. In this research, we have looked at the genetic relationship between these breeds with different analysis. Both PCA and STRUCTURE analyses across different Iranian horse breeds demonstrated that Persian Arabian horses cluster separately from Turkemen and Caspian horse breeds, but have some overlap with Dareshuri and Kurdish horses. This similarity can be explained by cross breeding of some Dareshuri and Kurdish horses with Persian Arabians to improve the former 2 breeds.

Conclusions

In this study, we have applied several methods for estimated genetic diversity to the Persian Arabian horse, and compared this breed with other Iranian native horse breeds. In general, high levels of genetic variability were observed in the Persian Arabian horse population. This diversity may reflect an ancient origin of this strain of Arabian horse. Because the overall size of the population is small, there is a risk of loss of genetic diversity unless these rare horses are sustainably managed. With good management practices, the Persian Arabian horse may represent a valuable genetic resource for the long-term preservation of Arabian horses. We observed the lack of differentiation of matrilineal families (strains) in Persian Arabian horses. We also identified genomic regions putatively under selection using Tajima's *D*, *H*, and H12 statistics which suggest biological pathways related to immune response as potentially under selection in Persian Arabian horses. We confirmed the distinct population structure between Persian Arabian with Turkemen and Caspian horse breeds.

Supplementary Material

Supplementary data are available at *Journal of Heredity* online.

Funding

This project was supported in part by funds from the Dorothy Russell Havemeyer Foundation, Inc. D.F.A. is an investigator of the Havemeyer Foundation.

Acknowledgments

We gratefully acknowledge the horse owners who permitted us to obtain samples for analysis, and the Persian Arabian Horse Association and Equestrian Federation of Iran for providing the Stud-book. We also thank Don Miller for the technical assistance.

Conflict of interest

The authors declare no conflict of interest.

Data Availability

SNP genotype data has been deposited on Dryad (doi: 10.5061/dryad.54vb7f2).

References

Ala-Amjadi M, Yeganeh H, Sadeghi M. 2017. Study of genetic variation in Iranian Kurdish horse using microsatellite marker. *Iran J Anim Sci*. 48:335–342.

Almarzook S, Reissmann M, Arends D, Brockmann GA. 2017a. Genetic diversity of Syrian Arabian horses. *Anim Genet*. 48:486–489.

Almarzook S, Reissmann M, Brockmann GA. 2017b. Diversity of mitochondrial DNA in three Arabian horse strains. *J Appl Genet*. 58:273–276.

Amirinia C, Seyedabadi H, Banabazi M, Kamali M. 2007. Bottleneck study and genetic structure of Iranian Caspian horse population using microsatellites. *Pakistan J Biol Sci*. 10:1540–1543.

Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet*. 5:202–212.

Bahbahani H, Tijjani A, Mukasa C, Wragg D, Almathen F, Nash O, Akpa GN, Mbole-Kariuki M, Malla S, *et al.* 2017. Signatures of selection for environmental adaptation and zebu × taurine hybrid fitness in East African Shorthorn Zebu. *Front Genet*. 8:68.

Barbato M, Orozco-Terwengel P, Tapio M, Bruford MW. 2015. SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. *Front Genet*. 6:109.

Bowling AT, Ruvinsky A. 2000. *Genetic aspects of domestication, breeds and their origins*. Wallingford: CABI Publishing. p. 25–51.

Capinera JL. 2004. *Encyclopedia of entomology:* volume 3 P-Z. Netherlands:Kluwer Academic Publishers.

Colonna V, Nutile T, Ferrucci RR, Fardella G, Aversano M, Barbujani G, Ciullo M. 2009. Comparing population structure as inferred from genealogical versus genetic information. *Eur J Hum Genet*. 17:1635–1641.

Corbin LJ, Blott SC, Swinburne JE, Vaudin M, Bishop SC, Woolliams JA. 2010. Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Anim Genet*. 41:8–15.

Delaneau O, Marchini J, 1000 Genomes Project Consortium. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun*. 5:3934.

Druml T, Neuditschko M, Grilz-Seger G, Horna M, Ricard A, Mesarič M, Cotman M, Pausch H, Brem G. 2017. Population networks associated with runs of homozygosity reveal new insights into the breeding history of the Haflinger horse. *J Hered*. 109:384–392.

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.

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.

Excoffier L, Laval G, Schneider S. 2007. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online*. 1:47–50.

Firouz L. 1998. Original ancestors of the Turkoman and Caspian horses. *First International Conference on Turkoman horses*; Ashgabat, Turkmenistan.

Forbis J. 1976. *The classic Arabian horse*. New York: Liveright Publishing Corporation.

Fotovati A. 2000. Persian horse breeds from ancient time to present and their rules in development of world horse breeds. *Asian-Aus J Anim Sci*. 13(Suppl): 401.

Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLOS Genet*. 11:e1005004.

Gharahveysi S, Irani M. 2011. Inbreeding study on the Iranian Arab horse population. *World J Zool*. 6:01–06.

Huang Da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 4:44–57.

Hudson W. 2017. Whole-loop mitochondrial DNA D-loop sequence variability in Egyptian Arabian equine matrilines. *PLoS One*. 12:e0184309.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*. 23:1801–1806.

Khanshour AM, Cothran EG. 2013. Maternal phylogenetic relationships and genetic variation among Arabian horse populations using whole mitochondrial DNA D-loop sequencing. *BMC Genet*. 14:83.

Kirin M, Mcquillan R, Franklin CS, Campbell H, Mckeigue PM, Wilson JF. 2010. Genomic runs of homozygosity record population history and consanguinity. *PLoS One*. 5:e13996.

Knorr F. 1912. A history of the Arabian horse and its influence on modern breeds. *J Hered*. 3:174–180.

Lee Y-S, Woo LEE J, Kim H. 2014. Estimating effective population size of thoroughbred horses using linkage disequilibrium and theta (4Nμ) value. *Livestock Sci*. 168:32–37.

Lischer HE, Excoffier L. 2012. PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*. 28:298–299.

Malomane DK, Reimer C, Weigend S, Weigend A, Sharifi AR, Simianer H. 2018. Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies. *BMC Genomics*. 19:22.

Metzger J, Karwath M, Tonda R, Beltran S, Agueda L, Gut M, Gut IG, Distl O. 2015. Runs of homozygosity reveal signatures of positive selection for reproduction traits in breed and non-breed horses. *BMC Genomics*. 16:764.

Moridi M, Masoudi AA, Vaez Torshizi R, Hill EW. 2013. Mitochondrial DNA D-loop sequence variation in maternal lineages of Iranian native horses. *Anim Genet*. 44:209–213.

Moshkelani S, Rabiee S, Javaheri-Koupaei M. 2011. DNA fingerprinting of Iranian Arab horse using fourteen microsatellites marker. *Res J Biol Sci*. 6:402–405.

Pitt D, Bruford MW, Barbato M, Orozco-Terwengel P, Martinez R, Sevane N. 2018. Demography and rapid local adaptation shape Creole cattle genome diversity in the tropics. *Evol Appl*. 58:238. Advance online publication. doi:10.1111/eva.12641

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38:904–909.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics*. 155:945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 81: 559–575.

Rahimi-Mianji G, Nejati-Javaremi A, Farhadi A. 2015. Genetic diversity, parentage verification, and genetic bottlenecks evaluation in iranian turkmen horse. *Russ J Genet*. 51:916–924.

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Available from http://www.R-project.org/

Rodriguez-Ramilo ST, Wang J. 2012. The effect of close relatives on unsupervised Bayesian clustering algorithms in population genetic structure analysis. *Mol Ecol Resour*. 12:873–884.

Sadeghi R, Moradi-Shahrbabak M, Miraei Ashtiani SR, Miller DC, Antczak DF. 2017. MHC haplotype diversity in Persian Arabian horses determined using polymorphic microsatellites. *Immunogenetics*. 70:305–315.

Sargolzaei M, Iwaisaki H, Colleau JJ, 2006. *Contribution, Inbreeding F, Coancestry (CFC): a tool for monitoring genetic diversity.* In: Comm in Proceedings of the 8th World Congr. Genet. Appl. Livest. Prod; Belo Horizonte, Brazil. p. 27–28.

Schaefer RJ, Schubert M, Bailey E, Bannasch DL, Barrey E, Bar-Gal GK, Brem G, Brooks SA, Distl O, Fries R, *et al.* 2017. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. *BMC Genomics*. 18:565.

Schlamp F, Van Der Made J, Stambler R, Chesebrough L, Boyko AR, Messer PW. 2016. Evaluating the performance of selection scans to detect selective sweeps in domestic dogs. *Mol Ecol*. 25:342–356.

Seyedabadi H, Amirinia C, Banabazi M, Emrani H. 2006. Parentage verification of Iranian Caspian horse using microsatellites markers. *Iran J Biotechnol*. 4:260–264.

Shahsavarani H, Rahimi-Mianji G. 2010. Analysis of genetic diversity and estimation of inbreeding coefficient within Caspian horse population using microsatellite markers. *Afr J Biotechnol*. 9:293–299.

Sved JA, Feldman MW. 1973. Correlation and probability methods for one and two loci. *Theor Popul Biol*. 4:129–132.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 123:585–595.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. 4:e72.

Woolliams JA, Mäntysaari EA. 1995. Genetic contributions of Finnish Ayrshire bulls over four generations. *Anim Sci*. 61:177–187.

World Arabian Horse Organization. List of registering authority members [Internet]. Arabian Horse Association; [cited 2014]. Available from: http://www.waho.org/list-of-registering-authority-members-2/