https://doi.org/10.17221/120/2020-CJAS

The assignment success for 22 horse breeds registered in the Czech Republic: The machine learning perspective

Lenka Štohlová Putnová[1], Radek Štohl[2]*

[1]Department of Animal Morphology, Physiology and Genetics, Faculty of AgriScience, Mendel University in Brno, Brno, Czech Republic
[2]Department of Control and Instrumentation, Faculty of Electrical Engineering and Communication, University of Technology, Brno, Czech Republic

**Corresponding author: putnova@email.cz*

Citation: Štohlová Putnová L., Štohl R. (2021): The assignment success for 22 horse breeds registered in the Czech Republic: The machine learning perspective. Czech J. Anim. Sci., 66: 1–12.

Abstract: The paper demonstrates the dependability of assignment testing in the identification of an appropriate breed to monitor comprehensive genetic information from molecular markers to analyse the collection of real population data covering 22 horse breeds registered in the Czech Republic, including native breeds and genetic resources. If 17 microsatellites are used, the mean number of alleles per locus corresponds to 10.4. The count of alleles at the individual loci ranges between five (*HTG07*) and 17 (*ASB17*). The loci *ASB02*, *ASB23*, *HMS03*, *HTG10*, and *VHL20* exhibit the highest gene diversity and observed heterozygosity (both above 80%), with the mean value of 0.77 and 0.73, respectively. The moderate total inbreeding coefficient (5.2%) is estimated across all the loci and breeds. The levels of apparent breed differentiation span from zero between the Czech Warmblood and Slovak Warmblood to 0.15 between the Shetland Pony and Standardbred. The phylogenetic breed relationships are revealed via the NeighbourNet dendrogram constructed from Reynolds' genetic distances, which clearly separate the Coldblood draught, Hot/Warmblood, and Pony horses. Our results reveal that the Bayesian approach (the Rannala and Mountain technique) provides the most intensive prediction power (83.6%) out of the GENECLASS tools and that the Bayes Net algorithm exhibits the best efficiency (78.4%) from the WEKA machine learning workbench options, considering the use of the five-fold cross validation technique. The algorithms could be trained on large real reference data sets, and thus there appears another viable perspective for machine learning in horse ancestry testing. In this context, it is also important to stress the fact that innovated computational tools will potentially lead towards structuring a novel web server to allow the identification of horse breeds.

Keywords: accuracy; GENECLASS analyses; individual breed assignment; DNA markers; WEKA algorithms

Funded by a project (NAZV QH92277) of the National Agency for Agricultural Research of the Ministry of Agriculture of the Czech Republic, utilising the institutional research and development support at Mendel University in Brno. Furthermore, the completion of this article was made possible by the grant (FEKT-S-20-6205) – "Research in Automation, Cybernetics and Artificial Intelligence within Industry 4.0" financially supported by the Internal Science Fund of Brno University of Technology.

https://doi.org/10.17221/120/2020-CJAS

Machine learning (ML), also known as statistical learning, is a subfield of artificial intelligence dedicated to the study of prediction and inference algorithms. An insight into how such instruments can be applied to solve pressing problems within animal science was proposed in a recent study by Morota et al. (2018). Importantly, ML will have an increasing impact on breeding as unstructured genomic and phenotype data expand. Further, the data sets in animal breeding have traditionally been larger than those in many related contemporary biological sciences, and the development of efficient algorithms has embodied a major and relevant activity in the field (Perez-Enciso 2017). Although an operational methodology could constitute a valuable tool for conservation and breed improvement programs, the verification and validation of population assignment methods are still required; moreover, the real-world performance of assignment tests that utilize machine learning remains an unexplored problem, despite the wide popularity of these testing means.

Livestock breeds have been formed through centuries of human and natural selection to fit a wide range of environmental conditions and human needs. At this point, let us note that all Warmbloods from former Czechoslovakia originated during the Habsburg Monarchy and Austro-Hungarian Empire, and their development involved multiple gene pools originating from Czech, Polish, Austrian and Hungarian stud farms. In Moravia, the discussed breed was a steady type, denoted as the Moravian Warmblood between 1920 and 1971. Unfortunately, Western European breeds were imported in 1970, and Czech Warmblood breeds became significantly influenced by Trakehner, along with other, mostly German, and later on also French and Dutch Warmblood breeds. Crossbreeding produces the genetically very heterogeneous population of the Czech Warmblood, which has not resulted in a stable type to date. Since 1971, the Moravian Warmblood has been included in the stud book of the Czech Warmblood but with a specific focus on character and stature. In 1996, a group of enthusiasts started their efforts to save and regenerate the remaining horses of this origin. Eight years later, the authorities restored the Moravian Warmblood stud book, thereby recognizing the breed's existence. By the end of the 20th century, the Kinsky Horse had been largely absorbed into the Czech Warmblood, too. The stud book of the Kinsky Horse was recognised by the Ministry of Agriculture of the Czech Republic only recently, in 2005. Predictably, the historical development of the Slovak Warmblood markedly resembled that of the Czech native Warmbloods. After the dissolution of Czechoslovakia, the stud book of the Slovak Warmblood bred in the Czech Republic was founded (in 1995), under the auspices of the Ministry of Agriculture. Some of the animals are nevertheless registered in both stud books (for the Czech and Slovak Warmbloods). The above-mentioned events probably affected the composition of the gene pools in Czech-raised horse breeds. Breed descriptors were developed to identify a breed, but these cover only "pure breed" or true to breed animals, excluding undefined or admixed populations. Molecular marker polymorphisms revolutionised breed identification through using small samples of biological tissue or germplasm, including specimens of blood, embryos, ova, semen, and carcass that do not show any evident phenotype (Iquebal et al. 2014). A breed signature using single nucleotide polymorphism (SNP) or microsatellite DNA markers has the potential to discriminate between pure breeds and breed crosses, an aspect important for population management and the conservation of animal genetic resources (Bjornstad and Roed 2002).

Most breed registries require DNA testing to verify the animal identity and parentage before registration or breeding. The International Studbook Committee (ISBC) relies on the International Society of Animal Genetics (ISAG). The DNA markers used by the Czech laboratories are based on the internationally recognised microsatellite testing panel operated by the ISAG; moreover, they have been endorsed by the Food and Agriculture Organization of the United Nations (FAO) as a tool to support diversity studies (Funk et al. 2020).

In the above-outlined sense, this research paper is designed to demonstrate, extensively and in detail, the dependability of the assignment success rate; the task is then performed by using different methods for selecting the appropriate breed. An allelic microsatellite profile was established to determine the distance-based phylogenetic relationships between the investigated Czech-registered horse breeds. The WEKA machine learning workbench, equipped with a broad collection of machine learning algorithms and data pre-processing techniques, was used to classify

https://doi.org/10.17221/120/2020-CJAS

and predict the desired indicators in the individual breeds. We also analysed the standard assignment methods (Bayesian, frequency, and distance) using GENECLASS and STRUCTURE, one of the most extensively used individual assignment programs available for analysing microsatellites (Pritchard et al. 2000); the software is also based on a Bayesian methodology. Generally, our study provides interesting results relevant to different assignment-based approaches to estimate breed identification in horses with stud books maintained in the Czech Republic; these outcomes may also find application in the building of a novel web server.

MATERIAL AND METHODS

Experimental data set

We collected applicable information relating to 7 588/1 742 (full/randomly reduced) individuals; this step was carried out via employing nuclear microsatellite DNA polymorphisms to describe 22 horse breeds for which stud books are available in the Czech Republic: THO = Thoroughbred (214/100), ARAB = Arabian (100/100), CZW = Czech Warmblood (3 295/100), CMB = Czech-Moravian Belgian Horse (127/100), CSP = Czech Sport Pony (74/74), HAF = Haflinger (243/100), HUC = Hucul (387/100), IRI = Irish Cob (43/43), STA = Standardbred (142/100), KIN = Kinsky Horse (100/100), LIP = Lipizzan (10/10), MIN = Miniature Horse (24/24), MOW = Moravian Warmblood (36/36), NOR = Noriker (83/83), SHA = Shagya (180/100), SHP = Shetland Pony (95/95), SNOR = Silesian Noriker (104/100), SLW = Slovak Warmblood in the Czech environment (1 888/100), KLA = Old Kladruber Horse (266/100), TRA = Trakehner (32/32), WPB = Welsh Part Bred (87/87), and WPC = Welsh Pony and Cob (58/58). Eight of these are unique native populations (CZW, MOW, KIN, CSP, CMB, HUC, KLA, and SNOR), and four rank among the genetic resources (CMB, HUC, KLA, and SNOR).

Genomic DNA isolation

The total genomic DNA was extracted from the blood, hair roots, and semen straws, using QIAamp® Blood/Tissue Kit (Qiagen, Valencia, CA, USA); JETQUICK Blood/Tissue DNA Spin Kit (Genomed GmbH, Löhne, Germany); or Geneaid™ DNA Isolation Blood/Tissue Kit (Geneaid Biotech Ltd., Taipei, Taiwan), invariably according to the manufacturers' instructions. The extracted DNA was visualised on 1% tris-acetate-EDTA (TAE) agarose gel stained with ethidium bromide. The samples were then denatured at 95 °C for 7 min and kept at –20 °C until the microsatellite analysis.

Allele detection and genotyping

Seventeen dinucleotide repeat microsatellite loci – *AHT04*, *AHT05*, *ASB02*, *ASB17*, *ASB23*, *UCDEQ425*, *HMS01*, *HMS02*, *HMS03*, *HMS06*, *HMS07*, *HTG04*, *HTG06*, *HTG07*, *HTG10*, *LEX3*, and *VHL20* – dispersed over 13 different chromosomes were subjected to an analysis. All microsatellite markers except *HMS01* and the X-chromosomal *LEX03* are among those recommended by the FAO for diversity studies. The DNA samples were diluted to a working concentration of 1–10 ng/µl by adding an appropriate amount of sterile water. The co-amplification of the 17 microsatellites was performed via a multiplex PCR reaction using the commercially available StockMarks® for Horses 17-Plex Genotyping Kit (Applied Biosystems, Foster City, CA, USA) and Equine Genotypes™ Panel 1.1 (Thermo Fisher Scientific Inc., Waltham, MA, USA) as recommended by the manufacturers. The PCR reactions were carried out employing a GeneAmp® PCR System 9700 or a Veriti® Thermal cycler (Thermo Fisher Scientific Inc., Waltham, MA, USA). Each 1 µl of the PCR product and 0.5 µl of a GeneScan™ 500 LIZ™ dye Size Standard (Life Technologies, Carlsbad, CA, USA) were loaded in 11.5 µl of Hi-Di™ Formamide (Life Technologies, Carlsbad, CA, USA). The samples were then denatured at 95 °C for 5 min and cooled down for another 5 minutes. The genotype scoring was performed on an ABI PRISM 310™ Genetic Analyzer (Thermo Fisher Scientific Inc., Waltham, MA, USA) with five-colour detection via fluorescent fragment analysis and then detected using dedicated software packages. The allele size was determined in bp by comparing the length with an internal-lane size standard. An alphabetical nomenclature was used for the allele size designation in accordance with the ISAG. Further, a sample of known genotypes from the ISAG comparison

https://doi.org/10.17221/120/2020-CJAS

test was employed as the reference and positive control to standardize the allele sizes in every run. The laboratory where our research was conducted is regularly accredited according to the ISO/IEC 17025 standard and has been involved in the ISAG comparative testing since 2001.

Statistical analyses

We used PowerMarker v3.25 (https://brcwebportal.cos.ncsu.edu/powermarker/index.html) to compute the frequencies and count of the alleles, polymorphic information content (PIC), observed heterozygosity, expected heterozygosity (or gene diversity), and *f* (total inbreeding coefficient due to subdivision). The Reynolds' genetic distances (D_R), a measure based on pairwise population differentiation (F_{ST}) values [$D_R = -\ln(1 - F_{ST})$], were computed for the Czech registered breeds using the same software and visualised by a NeighbourNet dendrogram via SplitsTree v4.13.1 (http://www.splitstree.org). The Hardy-Weinberg equilibrium test was performed with the Genepop v4.2.1 software (https://kimura.univ-montp2.fr/~rousset/Genepop.htm). A global test subsuming the diverse loci and populations, i.e., the multisample score test, was used together with the hypothesis of the heterozygote deficit and excess. When carrying out multiple tests, we adjusted the standards of significance via the sequential Bonferroni method. The loci were examined in terms of null alleles, large allele dropout, and stutter peaks in MICRO-CHECKER v2.2.3 (http://www.nrp.ac.uk/nrp-strategic-alliances/elsa/software/microchecker/) at a 95% confidence interval, utilizing 10 000 permutations according to Brookfield's and Chakraborty's approach.

The dataset was narrowed at random by MS Excel 2016 to 100 or fewer individuals per breed (n = 1 742) before using the program FSTAT v2.9.4 (http://www2.unil.ch/popgen/softwares/fstat.htm). The same program was exploited to estimate the linkage disequilibrium (LD) between pairs of loci and the fixation index F_{ST} (reduction in the heterozygosity of a subpopulation due to random genetic drift). Low F_{ST} values denote similarity between the allele frequencies within each population.

The DNA signature of a breed is a "statistical signature" based on the allele type, relative frequency, and its relative distribution. The probabilistic assignments of animals to the pre-determined populations were tested through several processes, specified as follows:

1) distance-based approaches – Nei's (1972) standard distance, Nei's (1973) minimum distance, Nei et al.'s (1983) D_A distance, Cavalli-Sforza and Edwards' (1967) distance, and Goldstein et al.'s (1995) distance;

2) the frequency-based method by Paetkau et al. (1995), which assigns an animal to the population where the individual's genotype will occur the most probably;

3) the Bayesian statistics-based technique of Rannala and Mountain (1997) and Baudouin and Lebrun (2001), incorporated in the GENECLASS v2.0.h software (http://www1.montpellier.inra.fr/CBGP/software/GeneClass/).

The WEKA v3.8.1 machine learning software (https://www.cs.waikato.ac.nz/ml/weka/) with a large corpus of machine learning algorithms was employed in the allelic data to predict and classify breeds registered in the Czech Republic, including Czech native breeds and genetic resources. Four classification methods were used: Bayes Network and Naive Bayes (probabilistic classification algorithms); JRip (rule learning algorithm); and multilayer perceptron (MLP) as a class of the artificial neural network (backpropagation as an iterative algorithm based on gradient descent). An algorithm is superior to a different one if it has more correctly and less incorrectly classified instances. The five-fold cross validation technique with default parameters was used to derive the classification performance of each option. At the final stage, the breed assignment was assessed by using Bayesian clustering methods in STRUCTURE v2.3.4 (Pritchard et al. 2000). An admixture model and a correlated allele frequency model with default parameters were used to analyse the data set. Ten independent runs were performed for number of populations K = 22, with the burn-in period of 50 000 steps, and followed by 150 000 Markov chain Monte Carlo iterations. The CLUMPP v1.1.2 (Jakobsson and Rosenberg 2007) was used to maximize the measure of similarity in the Q-matrices of the ten replicates.

We employed the Circos visualization software (Krzywinski et al. 2009), which exploits a circular ideogram, to facilitate global distribution of the prediction power detected in each horse breed via GENECLASS, WEKA, and STRUCTURE tools and arising from a comparison of the breeds' microsatellite data.

https://doi.org/10.17221/120/2020-CJAS

RESULTS AND DISCUSSION

Genetic variability among the loci

Genetic diversity is a necessary precondition for the successful and sustainable breeding of a population; one of the goals of population management therefore consists in maintaining genetic diversity at a high level, with inbreeding suppressed. The statistics overview relating to marker loci analysed across the data set (n = 7 588) is presented in Table 1. The mean number of alleles per locus was 10.4. The count of alleles at the loci ranged between five (*HTG07*) and 17 (*ASB17*). Certain alleles exhibited the highest occurrence across all the tested samples: *UCDEQ425* – allele *N* (0.463), *HMS01* – allele *M* (0.451), *HMS06* – allele *P* (0.444), *HTG04* – allele *M* (0.435), and *HTG07* – allele *O* (0.468). The observed heterozygosity ranged from 0.49 (*LEX03*) to 0.85 (*HMS03* and *ASB23*). The loci *ASB02*, *ASB23*, *HMS03*, *HTG10*, and *VHL20* showed the highest gene diversity and observed heterozygosity (both above 80%), with the mean values of 0.77 and 0.73, respectively. The *LEX03* locus yielded the highest inbreeding coefficient value (f = 0.414) due its localization on the X chromosome. Further, a moderate total inbreeding coefficient (f = 0.052) was estimated across all the loci and breeds. The observed heterozygosity had a strong negative correlation (Pearson's coefficient of r = –0.75, P = 0.000 526) with the total inbreeding due to subdivision. The number of alleles was moderately positively correlated with the gene diversity and the polymorphic information content, respectively (r = 0.71, P < 0.01). Significant interaction of these two factors was found previously (Fan et al. 2005).

MICRO-CHECKER did not reveal the presence of null alleles in any of the populations, and neither evidence of scoring errors due to stuttering nor a proof of large allele dropout were found. The horse breeds examined in terms of each locus via heterozygote deficit and excess exhibited genotypic frequencies in agreement with the Hardy-Weinberg equilibrium. The tests focused on the LD between all pairs of loci across all the populations, and they were carried out by using FSTAT; no significant

Table 1. The statistics summary for the marker loci in the horse breeds with stud books kept in the Czech Republic (n = 7 588)

Locus	Chromosome	MAF	Genotype No.	Allele No.	Gene diversity	Heterozygosity	PIC	f
AHT04	24	0.300	50	10	0.773	0.751	0.738	0.029
AHT05	8	0.236	36	9	0.804	0.793	0.775	0.014
HMS01	15	0.451	27	7	0.635	0.622	0.566	0.022
HMS02	10	0.358	51	11	0.771	0.741	0.740	0.040
HMS03	9	0.249	38	11	0.818	0.854	0.793	–0.044
HMS06	4	0.444	22	7	0.713	0.687	0.674	0.037
HMS07	1	0.317	35	8	0.796	0.758	0.768	0.048
HTG04	9	0.435	27	9	0.680	0.630	0.627	0.073
HTG06	15	0.328	35	10	0.736	0.679	0.689	0.078
HTG07	4	0.468	15	5	0.666	0.640	0.609	0.038
HTG10	21	0.252	62	11	0.839	0.825	0.819	0.017
VHL20	30	0.244	53	11	0.834	0.815	0.814	0.023
ASB02	15	0.250	60	13	0.832	0.803	0.811	0.035
ASB17	2	0.275	106	17	0.828	0.798	0.808	0.036
ASB23	3	0.226	66	15	0.830	0.845	0.807	–0.019
UCDEQ425	28	0.464	51	11	0.714	0.693	0.681	0.029
LEX03	–	0.229	64	12	0.844	0.495	0.826	0.414
Mean	–	0.325	46.941	10.412	0.771	0.731	0.738	0.052

f = moment estimator or maximum likelihood estimator of within-population inbreeding coefficient; MAF = major allele frequency; PIC = polymorphic information content

https://doi.org/10.17221/120/2020-CJAS

genotypic disequilibrium was found ($P > 0.05$ for all after Bonferroni corrections).

Breed relationships

The pairwise population differentiations (F_{ST}) and Reynolds' D_R genetic distances revealed close relationships between some of the populations. The calculated D_R distances (n = 7 588) and paired F_{ST} values (n = 1 742) between 22 Czech-registered breeds estimated from 17 microsatellite loci are presented in Table S1 in electronic supplementary material (ESM).

In most cases, the D_R showed the same pattern as the F_{ST}. The Reynolds' genetic distances pointed to close relationships as regards the CZW and SLW breeds (D_R = 0.002 7). Further, the members of the combination SNOR-NOR exhibited the highest connectivity to each other (D_R = 0.010 6). Common historical origin or past breeding strategies possibly led to genetic admixture occurring as a result of the relatively intensive gene flow in these breeds. By contrast, we determined the pair SHP-LIP (D_R = 0.167 6) to be the most distinct one in terms of the Czech-registered breeds. The genetic differentiation among the SHP and the STA (F_{ST} = 0.151 1) was the highest from all the breed combinations. Zero F_{ST} value was detected between the SLW and CZW, with these breeds exhibiting the highest similarity. We found that 35 values (15%) of the 231 pairwise F_{ST} indices were ≤ 0.05. The breeds' relationships were analysed also via NeighbourNet visualization of the Reynolds' D_R genetic distances. A NeighbourNet dendrogram of the horse breeds with stud books maintained in the Czech Republic was designed (Figure 1). The network topology clearly separated the Cold-blooded draught, Hot/Warm-blooded, and Pony horses.

Breed assignment

We studied the assignment success in 22 horse breeds with stud books kept in the Czech Republic (n = 1 742), including the Czech native populations (n = 710) and genetic resources (n = 400). Eight different methods (Bayesian, frequency, and distance) from the GENECLASS software, four classifiers (Bayes Network, Naive Bayes, JRip, and multilayer perceptron) associated with WEKA, and STRUCTURE software were all used in the data set to facilitate the horse breed identification. Applying different machine learning algorithms to a real horse database may prove useful for comparative breed analysis. The X-linked locus *LEX03* was not included in the following breed assignment analyses.

Table S2 in ESM summarizes the individual assignment success rates for each horse breed, obtained by using eight different assignment methods in GENECLASS. The Bayesian, frequency, and Nei-D_A distance techniques performed more or less similarly. The Pearson correlation analysis demonstrated only a weak relationship between the sample size

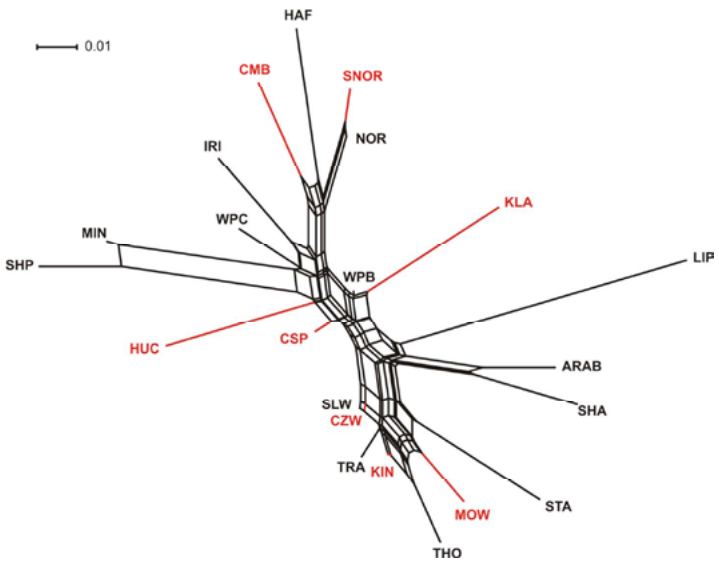

Figure 1. The neighbour network built from the D_R distances between 22 horse breeds registered in the Czech Republic, including the Czech native breeds and genetic resources (n = 7 588)
ARAB = Arabian; CMB = Czech-Moravian Belgian Horse; CSP = Czech Sport Pony; CZW = Czech Warmblood; HAF = Haflinger; HUC = Hucul; IRI = Irish Cob; KIN = Kinsky Horse; KLA = Old Kladruber Horse; LIP = Lipizzan; MIN = Miniature Horse; MOW = Moravian Warmblood; NOR = Noriker; SHA = Shagya; SHP = Shetland Pony; SLW = Slovak Warmblood in the Czech environment; SNOR = Silesian Noriker; STA = Standardbred; THO = Thoroughbred; TRA = Trakehner; WPB = Welsh Part Bred; WPC = Welsh Pony and Cob
The eight unique native breeds inclusive of the endangered gene reserve (CMB, HUC, KLA, and SNOR) are indicated in red

https://doi.org/10.17221/120/2020-CJAS

and the assignment accuracy. The overall percentage of correct assignment for the 14 different methods in the 22 registered breeds is summarised in Table 2. For the breeds registered in the Czech Republic, the breed-wise accuracy in GENECLASS and WEKA ranged from 83.6% to 35.8% and from 78.4% to 46.2%, respectively. Our results discussed above indicate that the Bayesian method (the Rannala and Mountain approach) provided a prediction power superior to the other GENECLASS methods and that the Bayes Net, then, surpassed all similar algorithms within the WEKA ML family. Similarly, high prediction power was obtained from the STRUCTURE. To ensure such a capability, however, we had to supply a priori information about the sample origin; without these data, there was only 58.03% accuracy yielded from the Bayesian approach that assigns individuals to source populations or facilitates the proportional assignment of ancestry to multiple (22 in this case) populations. Very large data sets are highly challenging for cluster analyses, especially when populations with complex genetic histories are investigated (Funk et al. 2020). STRUCTURE works best with a small number of discrete populations (Pritchard et al. 2000), but the performance with large data sets was also evaluated (e.g. Putnova et al. 2019; Funk et al. 2020).

The confusion matrices to show the prediction power of these three best performing tools, as calculated through GENECLASS, WEKA, and STRUCTURE, are proposed in Tables S3, S4, and S5 in ESM. Circular breed data visualization was carried out to allow individual identification and analysis of the similarities and differences arising from comparison of the microsatellite variation between the pairs of breeds. Summarizing the distribution of the assignment success, appropriate Circos plots were created for each of the confusion matrices (Figures 2, 3 and 4).

Breed assignment tests can be performed by means of diverse statistical tools based on the genetic distances and differences in allelic frequencies. According to Cornuet et al. (1999), the procedures exploiting the Bayesian approach yield the best results, but the populations must be in the Hardy-Weinberg and linkage equilibria; this necessity, however, can be eliminated by distance-based methodologies. Factors having an impact on the individual assignment success, such as the number and divergence of populations, sample size, and number of loci, were assessed from actual or simulated data and are widely reported in the literature (Bjornstad and Roed 2002; Koskinen 2003; Fan et al. 2005; Talle et al. 2005; Putnova and Stohl 2019; Funk et al. 2020).

In our study, the Goldstein $\delta\mu^2$ distance and the neural network MLP model based on gradient descent algorithm exhibit overall inferior capabilities in breed classification. However, the performance of each method varied, with the accuracy affected by genetic differentiation between the breeds, the number and allelic diversity of loci, the proportion of sampled animals, and the number of baseline populations (Bjornstad and Roed 2002; Koskinen

Table 2. The assignment accuracy of 14 different methods for 22 horse breeds registered in the Czech Republic

Software	Method	Assignment accuracy (%)
STRUCTURE v2.3.4	Using prior population information	89.74
GENECLASS v2.0.h	Bayesian method – Rannala and Mountain approach	83.59
GENECLASS v2.0.h	Bayesian method – Baudouin and Lebrun approach	80.69
GENECLASS v2.0.h	Frequency method	80.39
WEKA v3.8.1	Bayes Network	78.36
GENECLASS v2.0.h	Nei's D_A distance	77.55
WEKA v3.8.1	Naive Bayes	75.13
GENECLASS v2.0.h	Nei's standard distance	73.32
GENECLASS v2.0.h	Cavalli-Sforza and Edwards' distance	73.30
GENECLASS v2.0.h	Nei's minimum distance	73.13
STRUCTURE v2.3.4	Using only genetic information	58.03
WEKA v3.8.1	JRip	50.11
WEKA v3.8.1	Multilayer perceptron	46.21
GENECLASS v2.0.h	Goldstein's $\delta\mu^2$ distance	35.81

https://doi.org/10.17221/120/2020-CJAS

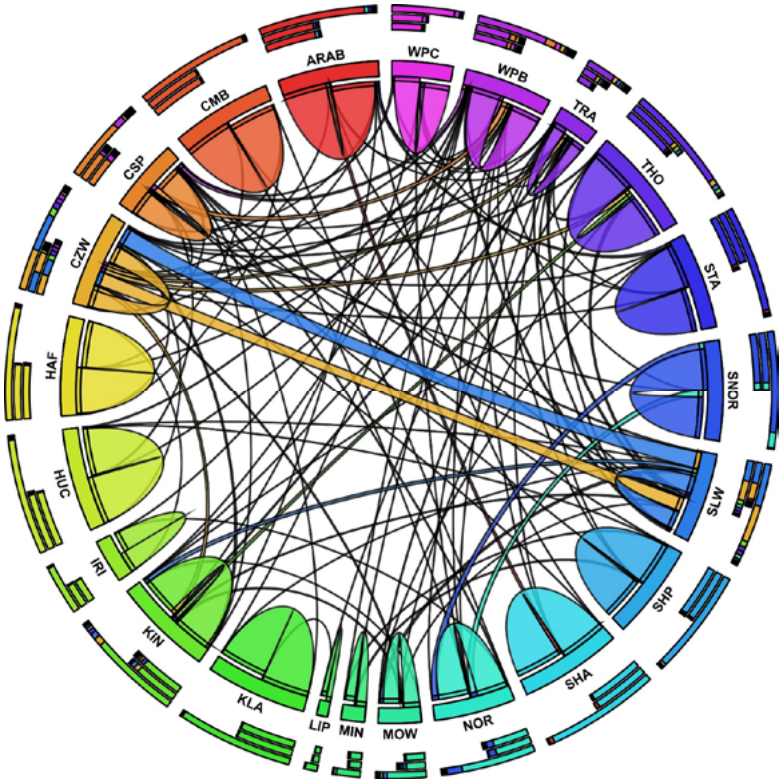

Figure 2. A Circos plot displaying the variation in the microsatellite structure and summarizing the genetic prediction power of the Bayesian method (the Rannala and Mountain approach), as calculated by GENECLASS for each Czech-registered horse breed (n = 1 742; accuracy 83.6%)
ARAB = Arabian; CMB = Czech-Moravian Belgian Horse; CSP = Czech Sport Pony; CZW = Czech Warmblood; HAF = Haflinger; HUC = Hucul; IRI = Irish Cob; KIN = Kinsky Horse; KLA = Old Kladruber Horse; LIP = Lipizzan; MIN = Miniature Horse; MOW = Moravian Warmblood; NOR = Noriker; SHA = Shagya; SHP = Shetland Pony; SLW = Slovak Warmblood in the Czech environment; SNOR = Silesian Noriker; STA = Standardbred; THO = Thoroughbred; TRA = Trakehner; WPB = Welsh Part Bred; WPC = Welsh Pony and Cob

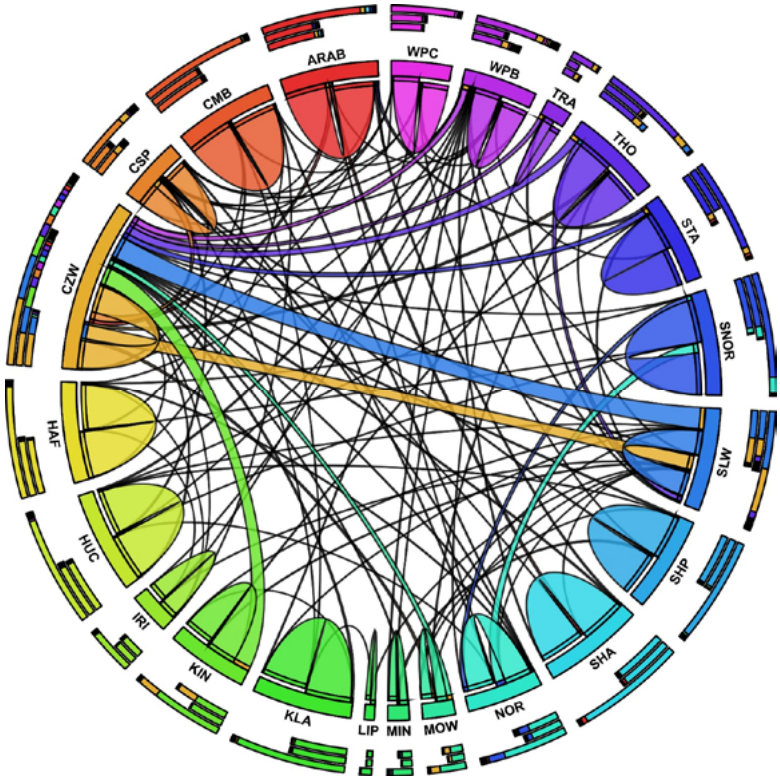

Figure 3. A Circos plot displaying the variation in the microsatellite structure and summarizing the genetic prediction power of the Bayes Network classifier, as calculated by WEKA for each Czech-registered horse breed (n = 1 742; accuracy 78.4%)
ARAB = Arabian; CMB = Czech-Moravian Belgian Horse; CSP = Czech Sport Pony; CZW = Czech Warmblood; HAF = Haflinger; HUC = Hucul; IRI = Irish Cob; KIN = Kinsky Horse; KLA = Old Kladruber Horse; LIP = Lipizzan; MIN = Miniature Horse; MOW = Moravian Warmblood; NOR = Noriker; SHA = Shagya; SHP = Shetland Pony; SLW = Slovak Warmblood in the Czech environment; SNOR = Silesian Noriker; STA = Standardbred; THO = Thoroughbred; TRA = Trakehner; WPB = Welsh Part Bred; WPC = Welsh Pony and Cob

2003; Fan et al. 2005; Talle et al. 2005). The assignment success increased with rising numbers of loci, higher levels of divergence, and lower amounts of source breeds (Putnova and Stohl 2019). As for the fourteen assignment methods used herein, most animals were correctly classifiable into their

https://doi.org/10.17221/120/2020-CJAS

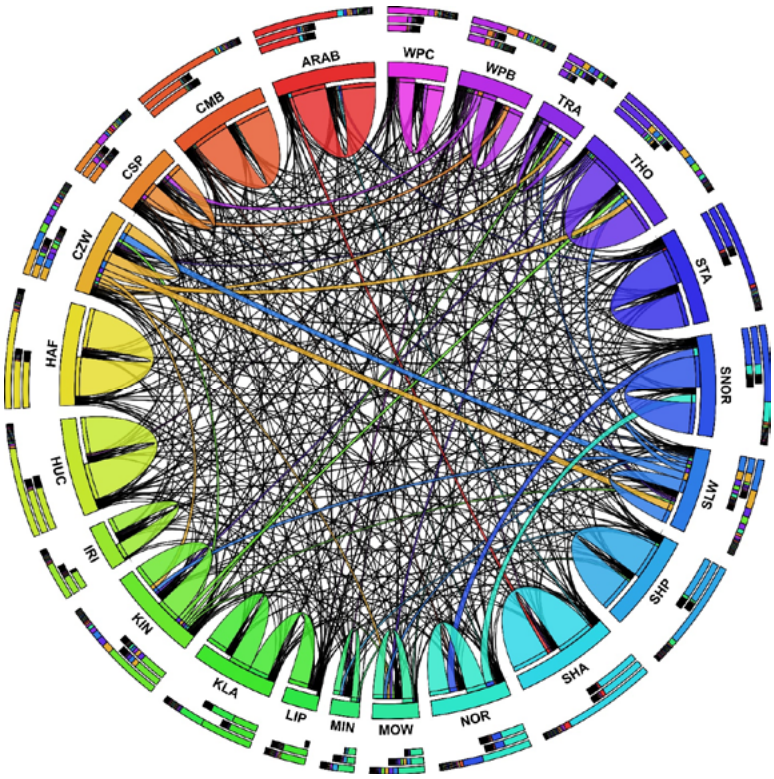

Figure 4. A Circos plot displaying the variation in the microsatellite structure and summarizing the genetic prediction power of the Bayesian clustering method, as calculated by STRUCTURE for each Czech-registered horse breed (n = 1 742; accuracy 58.03%)
ARAB = Arabian; CMB = Czech-Moravian Belgian Horse; CSP = Czech Sport Pony; CZW = Czech Warmblood; HAF = Haflinger; HUC = Hucul; IRI = Irish Cob; KIN = Kinsky Horse; KLA = Old Kladruber Horse; LIP = Lipizzan; MIN = Miniature Horse; MOW = Moravian Warmblood; NOR = Noriker; SHA = Shagya; SHP = Shetland Pony; SLW = Slovak Warmblood in the Czech environment; SNOR = Silesian Noriker; STA = Standardbred; THO = Thoroughbred; TRA = Trakehner; WPB = Welsh Part Bred; WPC = Welsh Pony and Cob

populations of origin. Regarding the Czech native breeds, the best individual assignment results were achieved in the HUC (98%), KLA (98%), and CMB (97%). Conversely, and as expected, a lower breed assignment level was obtained in the CZW (35%) and SLW (36%). The CZW and the SLW exhibit higher admixture between each other, as there has been a major exchange between these two breeds in recent years. Using the Bayesian statistical technique, Rannala and Mountain (1997) and Van de Goor et al. (2011) obtained an assignment accuracy for the Welsh (82.2%) and Warmblood breeds (85.4%) similar to our data (80.2% in the Welsh and 80.1% in the Warmblood breeds).

Out of the 1 742 individuals considered, 83.6/81.5/94.8% were correctly assigned to their populations of origin with respect to the Czech registered breeds/native breeds/gene reserves (the Rannala and Mountain approach); by another definition, the indicated values correspond to the average assignment values in the relevant groups. When the same technique was applied to 22 breeds including eleven breeds of Warmblood, seven of Pony, and four of Cold-blooded horses, the following respective breed identification accuracies were achieved: 80.1%, 86.2%, and 88.7%. Thus, the highest breed genomic prediction accuracy was found in the Coldblood horses. Overall, the Bayesian algorithms implemented in WEKA and GENECLASS outperformed the other approaches. These statistical tools, inclusive of machine learning, could be employed to assess the breed traceability system.

Similarly to our study, relevant papers by Fan et al. (2005) and Talle et al. (2005) associate the top percentages of correct assignment with the Bayesian, gene frequency, and distance-based methods. However, when using a ten loci combination, Koskinen (2003) and Fan et al. (2005) reported slightly higher values in dogs and pigs, respectively. Based on multilocus genotypes, the Bayesian method provided 98% assignment success in pigs. Most strikingly, the Bayesian genetic assignment analysis did not indicate any gene flow pattern between the assessed canine breeds, as all of the individuals were 100% assigned to their reference populations. In cattle, the assignment success rate ranged from 55% to 70%, involving eight purebred populations and 27 loci (Talle et al. 2005). Nevertheless, the assignment accuracy value for MLP observed in this study (46.2%) is not in good agreement with the value found by Iquebal et al. (2014) for goat breeds (96.6%).

Today, well-defined breed descriptors based on morphology are used to categorize breeds. These phenotypic tools have certain limitations, as they

https://doi.org/10.17221/120/2020-CJAS

cannot identify semen, ova, embryo, or a breed product; moreover, they are unable to predict a breed in an admixture, namely, a non-descriptive population (Iquebal et al. 2014).

In the year 2000, Pritchard, Stephens, and Donnelly released a model-based clustering method (Pritchard et al. 2000) that was to become known as STRUCTURE, one of the most important and widely recognised software tools to serve the given purpose. A clustering, microsatellite-based analysis of large datasets is nevertheless very time-intensive, and we thus employed STRUCTURE only marginally. The software finds use especially in off-line analyses; conversely, the algorithms contained in WEKA and GENECLASS are trainable on big data accessible in laboratories, can be utilised for on-line web applications, and the assignment of an individual to a breed does not require excessive time. Future work should focus on testing how a higher number of loci (microsatellites and/or SNP) would increase the breed assignment success, and whether the change of the large-scale cluster analysis of populations into smaller units reflecting the horse breed history performs better.

In horses, the population structure resulting from selective breeding is characterised by high interbreed and low intrabreed genetic diversity, and reflected by a huge array of morphological and behavioural traits (Librado et al. 2016). Using ancient genomes, Fages et al. (2019) found that while past horse breeders maintained diverse genetic resources for millennia after they first domesticated the horse, this diversity dropped by ~16% within the last 200 years. Modern breeding practices have changed the natural evolutionary trajectory of horses by favouring the reproduction of a limited number of animals showing traits of interest. Reduced breeding stocks hampered the elimination of deleterious variants by means of negative selection, ultimately inflating mutational loads (Orlando and Librado 2019). The authors confirm that deleterious mutations segregated at low frequencies during the last 3 500 years have only spread and incremented their occurrence in the homozygous state during modern times, owing to inbreeding.

We hope that the genetic monitoring cycles outlined herein will continue in the future to preserve the unique breeds kept in the Czech Republic. The genetic diversity and assignment accuracy found in domestic horse breeds allow farmers to develop new characteristics in response to changes in the environment, diseases, or market conditions. Microsatellite DNA genetic testing instruments are available to advantageously facilitate the correct classification of different groups and to potentially assist managers in the decision-making steps that relate to breeding and registration, mainly in genetically differentiated breeds.

CONCLUSION

Microsatellites and/or SNP chip-based data are usable in horse breed ancestry testing. While breed identification in purebred horses can be performed rather easily, determining breed contributions in crossbred individuals embodies a significantly more complex problem. In Czech husbandry practices, the animals are not tested for breed confirmation, because such a task requires additional testing, and certain limitations are imposed on breed composition tests to render them more informative and accurate. As regards the functional strategies, this paper shows that the WEKA Bayes network classification algorithms, despite not being centred on genetic analyses, exhibit results comparable to those provided by GENECLASS. At dedicated laboratories, large sets of real data are available, containing sufficient amounts of genetic profiles from individual horses belonging to different breeds. Interestingly, horse ancestry testing could be based upon comparing the DNA genotype of a horse to a reference data set of various horse breeds, and thus there appears another viable perspective for machine learning, which allows the algorithm to be trained on big data. In this context, it is also important to stress the fact that innovated computational tools will potentially lead to establishing a novel web server to allow the identification of horse breeds. Such means and instruments based on DNA profile will then enable breeders to verify not only an animal's but also a breed's identity.

Acknowledgement

The authors acknowledge Irena Vrtková, Ph.D., of the Laboratory of Agrogenomics, for her unwavering support and intensive cooperation over the years; professors Eva Tůmová and Antonín Stratil (the editor and co-editor, respectively), for their

https://doi.org/10.17221/120/2020-CJAS

time, effort, and close attention to the manuscript; and the anonymous reviewers, for the comments and recommendations that markedly helped to refine the final text.

Conflict of interest

The authors declare no conflict of interest.

REFERENCES

Baudouin L, Lebrun P. An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. Acta Hortic. 2001;546:81-93.

Bjornstad G, Roed KH. Evaluation of factors affecting individual assignment precision using microsatellite data from horse breeds and simulated breed crosses. Anim Genet. 2002 Aug;33(4):264-70.

Cavalli-Sforza LL, Edwards AWF. Phylogenetic analysis: Models and estimation procedures. Am J Hum Genet. 1967 May;19(3 Pt 1):233-57.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics. 1999 Dec;153(4):1989-2000.

Fages A, Hanghoj K, Khan N, Gaunitz C, Seguin-Orlando A, Leonardi M, McCrory Constantz C, Gamba C, Al-Rasheid KAS, Albizuri S, Alfarhan AH, Allentoft M, Alquraishi S, Anthony D, Baimukhanov N, Barrett JH, Bayarsaikhan J, Benecke N, Bernaldez-Sanchez E, Berrocal-Rangel L, Biglari F, Boessenkool S, Boldgiv B, Brem G, Brown D, Burger J, Crubezy E, Daugnora L, Davoudi H, de Barros Damgaard P, Chorro y de Villa-Ceballos MA, Deschler-Erb S, Detry C, Dill N, do Mar Oom M, Dohr A, Ellingvag S, Erdenebaatar D, Fathi H, Felkel S, Fernandez-Rodriguez C, Garcia-Vinas E, Germonpre M, Granado JD, Hallsson JH, Hemmer H, Hofreiter M, Kasparov A, Khasanov M, Khazaeli R, Kosintsev P, Kristiansen K, Kubatbek T, Kuderna L, Kuznetsov P, Laleh H, Leonard JA, Lhuillier J, von Lettow-Vorbeck CL, Logvin A, Lougas L, Ludwig A, Luis C, Arruda AM, Marques-Bonet T, Silva RM, Merz V, Mijiddorj E, Miller BK, Monchalov O, Mohaseb FA, Morales A, Nieto-Espinet A, Nistelberger H, Onar V, Palsdottir AH, Pitulko V, Pitskhelauri K, Pruvost M, Sikanjic PR, Papesa AR, Roslyakova N, Sardari A, Sauer E, Schafberg R, Scheu A, Schibler J, Schlumbaum A, Serrand N, Serres-Armero A, Shapiro B, Seno SS, Shevnina I, Shidrang S, Southon J, Star B, Sykes N, Taheri K, Taylor W, Teegen WR, Trbojevic Vukicevic T, Trixl S, Tumen D, Undrakhbold S, Usmanova E, Vahdati A, Ialenzuela-Lamas S, Viegas C, Wallner B, Weinstock J, Zaibert V, Clavel B, Lepetz S, Mashkour M, Helgason A, Stefansson K, Barrey E, Willerslev E, Outram AK, Librado P, Orlando L. Tracking five millennia of horse management with extensive ancient genome time series. Cell. 2019 May 30;177(6):1419-35.

Fan B, Chen YZ, Moran C, Zhao SH, Liu B, Zhu MJ, Xiong TA, Li K. Individual-breed assignment analysis in swine populations by using microsatellite markers. Asian-Aust J Anim Sci. 2005 Dec 2;18(11):1529-34.

Funk SM, Guedaoura S, Juras R, Raziq A, Landolsi F, Luis C, Martinez AM, Musa Mayaki A, Mujica F, Oom MD, Ouragh L. Major inconsistencies of inferred population genetic structure estimated in a large set of domestic horse breeds using microsatellites. Ecol Evol. 2020 May;10 (10):4261-79.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW. Genetic absolute dating based on microsatellites and the origin of modern humans. PNAS. 1995 Jul 18; 92(15):6723-7.

Iquebal MA, Ansari MS, Dixit SP, Verma NK, Aggarwal RA, Jayakumar S, Rai A, Kumar D. Locus minimization in breed prediction using artificial neural network approach. Anim Genet. 2014 Dec;45(6):898-902.

Jakobsson M, Rosenberg NA. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007 Jul 15;23(14):1801-6.

Koskinen M. Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. Anim Genet. 2003 Aug;34(4):297-301.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: An information aesthetic for comparative genomics. Genome Res. 2009 Sep 1;19(9):1639-45.

Librado P, Fages A, Gaunitz C, Leonardi M, Wagner S, Khan N, Hanghoj K, Alquraishi SA, Alfarhan AH, Al-Rasheid KA, Der Sarkissian C, Schubert M, Orlando L. The evolutionary origin and genetic makeup of domestic horses. Genetics. 2016 Oct 1;204(2):423-34.

Morota G, Ventura RV, Silva FF, Koyama M, Fernando SC. Machine learning and data mining advance predictive big data analysis in precision animal agriculture. J Anim Sci. 2018 Apr;96(4):1540-50.

Nei M. Genetic distance between populations. Am Nat. 1972 May-Jun;106(949):283-91.

Nei M. The theory and estimation of genetic distance. In: Morton NE, editor. Genetic structure of populations. Honolulu: University of Hawaii Press; 1973. p. 45-51.

https://doi.org/10.17221/120/2020-CJAS

Nei M, Tajima F, Tateno Y. Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol. 1983 Mar; 19(2):153-70.

Orlando L, Librado P. Origin and evolution of deleterious mutations in horses. Genes. 2019 Aug 28;10(9): [16 p.].

Paetkau D, Calvert W, Stirling I, Strobeck C. Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol. 1995 Jun;4(3):347-54.

Perez-Enciso M. Animal breeding learning from machine learning. J Anim Breed Genet. 2017 Apr;134(2):85-6.

Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun 1;155(2):945-59.

Putnova L, Stohl R. Comparing assignment-based approaches to breed identification within a large set of horses. J Appl Genet. 2019 Apr 8;60(2):187-98.

Putnova L, Stohl R, Vrtkova I. Using nuclear microsatellite data to trace the gene flow and population structure in Czech horses. Czech J Anim Sci. 2019 Feb;64(2):67-77.

Rannala B, Mountain JL. Detecting immigration by using multilocus genotypes. PNAS. 1997 Aug 19;94(17):9197-201.

Talle SB, Fimland E, Syrstad O, Meuwissen T, Klungland H. Comparison of individual assignment methods and factors affecting assignment success in cattle breeds using microsatellites. Acta Agric Scand. 2005 Aug;55(2-3):74-9.

Van de Goor LH, van Haeringen WA, Lenstra JA. Population studies of 17 equine STR for forensic and phylogenetic analysis. Anim Genet. 2011 Dec;42(6):627-33.

Received: May 6, 2020
Accepted: November 28, 2020