Topic 2: Fine-scale analysis of population structure based on genomic data and quantification of selection effect on livestock genome
Study material

Advances in biotechnology and DNA sequencing of livestock have enabled the optimization of breeding strategies to preserve diversity and the productive potential of animal genetic resources without pedigree analysis. Among the most widely used genetic markers today in this context are single nucleotide polymorphisms (SNPs). SNP markers are classified as biallelic markers, which means that three groups of individuals can be expected to occur in the population: dominant homozygotes, heterozygotes and recessive homozygotes. If diversity in the population has not declined, the ratio of these groups should be 25:50:25. Loss of diversity usually occurs as a result of inbreeding, where the proportion of homozygous versus heterozygous animals in a population increases, while at the same time, the overall level of biodiversity decreases.

Currently, SNP chips with low (at least 3,000 – 6,000 SNPs) or high (50,000 – 777,000 SNPs) marker density and whole genome sequences (2 million or more bases) are primarily used to quantify the biodiversity level in livestock, companion animals, and, in some cases, wild animals. SNP chips with at least 50,000 SNP markers cover a substantial portion of the genome because the SNP markers themselves are distributed evenly across all chromosomes in the genome. SNP chips are available for nearly all livestock species and some companion animals (cattle, horses, sheep, goats, pigs, poultry, dogs, and cats). A major proportion of SNP markers included in these chips are selectively neutral; however, a certain proportion is located directly in regions encoding proteins and regulatory sequences (Hayes and Goddard, 2010). Because SNP chips provide a comprehensive overview of genome structure, this technology has become the most popular tool worldwide for determining the genetic predisposition of individuals and populations, increasing genetic gain and preserving the gene pool of animal genetic resources. The current cost of genotyping using 50K SNP chips (50,000 SNP markers) is comparable to paternity testing through several microsatellite markers (for example, approximately 25 euros per individual for cattle and sheep), making this type of analysis also financially efficient.

Compared to other types of genetic markers (e.g., microsatellites, RFLP, or AFLP markers), SNP chips provide much more robust information about the genetic makeup of individuals, allowing more precise estimation of biodiversity indicators in populations and their relationships to each other (Lenstra et al., 2012). Genomic information obtained using SNP chips can be successfully used to estimate various genomic diversity parameters, including the genomic inbreeding coefficient (Moravčíková et al., 2017; Kasarda et al., 2019), the level of linkage disequilibrium, trends in effective population size (Kukučková et al., 2017a,b; Moravčíková et al., 2017; Kasarda et al., 2021a,b), genetic differentiation, and degree of population admixture (Moravčíková et al., 2015; Kukučková et al., 2017a; Moravčíková et al., 2021), or to identify selection signals in the genome (Moravčíková et al., 2019a,b,c).

Genetic Structure of Populations

The genetic structure of populations represents their diversity at both intra- and inter-population levels. Calculation of genetic distances is commonly used to determine differences between individuals and populations, representing the molecular equivalent of the relatedness coefficient derived from pedigree data. Genetic distance measures the genetic difference (genomic variation, e.g., in allele frequencies) between species or populations, which can be quantified using various statistical approaches (Kasarda et al., 2021a). Mathematically, genetic distances can be simply determined using the calculation of Nei's genetic distance (Nei, 1972), Wright's F_{ST} index (Wright, 1940), or visualized via principal component analysis (Jombart and

Ahmed, 2011), or by constructing phylogenetic trees based on IBD matrices (Neuditschko et al., 2012).

Another method for determining genetic differentiation within and between populations is cluster analysis, which also allows for estimating the degree of genetic admixture between populations or breeds. Genetic admixture is a phenomenon that can result from introgression and hybridization of individuals, populations, or species. Introgression describes the "flow" of alleles from one population of a species (typically non-native) or subspecies into another population (native to a given locality). The proportion of genetic variants defining the degree of uniqueness of a population or the degree of admixture with other populations is usually determined by a Bayesian approach (Toro et al., 2014; Jombart and Collins, 2015).

Bayesian statistics is a branch of modern statistics that works with conditional probability and allows to refine the probability of the initial hypothesis in the sequence as other relevant facts appear. The basis of its mathematical model is Bayes' theorem. While classical statistics determines the probability of an event based on known facts from the past, Bayesian statistics is used wherever this is not possible. In population genetics, Bayesian statistics is most often used to estimate the degree of genetic admixture between populations by comparing the frequencies of alleles that define certain populations or specific groups with the frequencies found in individuals. However, its use is much broader because it is fundamentally an estimate of the probability of a given phenomenon occurring in the test population. The accuracy of this method in terms of estimating population differentiation and determining the degree of admixture is often limited by the low number of individuals in the population or the low number of genetic markers. Genetic markers used for this type of analysis should be selectively neutral, not affected by mutation, and in linkage equilibrium with each other. The most commonly used statistical program in this context is Structure (Pritchard et al., 2000), which analyzes genetic differences between populations using a Bayesian approach and Markov Chain Monte Carlo (MCMC) estimation. The MCMC process begins by randomly assigning individuals to a predetermined number of groups in which individuals potentially share similar types of variations. The MCMC process starts by randomly dividing individuals into a predefined number of groups in which individuals potentially share similar types of variation. Then, the frequencies in each group are estimated and based on their values, individuals are divided into clusters. This process is repeated several times, usually in 10,000 to 100,000 replications, leading to reliable estimates of the membership of individuals in each population as well as the degree of genetic admixture between them (Kadlečík et al., 2017).

Impact of Selection on Genome Structure

In the genomic information era, the impact of natural and artificial selection on the genome of animal genetic resources can be quantified without access to phenotypic information. Selection signals can be identified in coding and non-coding regions of genes, depending on the statistical approach used (Qanbari and Simianer, 2014). The choice of method for detecting regions affected by selection pressure depends on the nature of the selection signals as well as the time over which selection has acted on the genome. The number of identified selection signals is influenced by various factors, including the intensity of selection, recombination rate, and the relative age of neutral alleles located near loci affected by selection pressure.

Methods for detecting loci under selection are divided into several groups based on the methodological approach applied. Testing for selection signals reflecting differences between populations due to different types of selection typically uses the determination of Wright's F_{ST} index, analysis of linkage disequilibrium variability, or calculation of the integrated haplotype score (Kukučková et al., 2017a,b; Moravčíková et al., 2019b). Determination of signals resulting from selection pressure on the genome of a specific population in order to achieve a

breeding standard or breeding goal is based in most cases on screening homozygous regions in the genome and determination of the integrated haplotype score (Moravčíková et al., 2019c; Kasarda et al., 2021b; Moravčíková et al., 2021).

Wright's F_{ST} index

One of the most commonly used approaches for detecting selection signals is calculating the genome-wide fixation index, Wright's F_{ST}, which reflects differences in allele frequencies of tested genetic markers between two or more populations. Wright's F_{ST} index also quantifies the level of genetic differentiation or fragmentation of the total population, expressed by a reduction in heterozygosity within individual subpopulations due to genetic drift. It is also referred to as the coefficient of relatedness and is defined as the correlation between gametes of subpopulations relative to randomly selected gametes from the total population (Weir and Cockerham, 1984).

Theoretically, Wright's F_{ST} index can range from 0 to 1, where extreme values represent complete genetic identity of populations (F_{ST} = 0) or, conversely, their complete genetic differentiation (F_{ST} = 1). Two types of signals may arise when using the F_{ST} index. In the first type, the value of the F_{ST} index increases in the selection signal region represented by multiple loci located close to each other due to the "hitchhiking" effect. This type of selection signal corresponds to differences in the direction of selection utilized within individual breeds. On the other hand, when loci with very low F_{ST} values are located within the selection signal region, they represent genomic regions subject to the same type of selection across breeds (Qanbari et al., 2011). Although several modifications of this method have been developed, the F_{ST} index is considered one of the most appropriate indicators of genomic signals of positive selection resulting from genetic differentiation between populations (Fumagalli et al., 2013).

Principal Component Analysis

Positive natural selection or local adaptation are driving forces enabling individuals to adapt to the environmental conditions in which they live. Genome screening to detect genetic variants potentially involved in this process typically relies on genetic differentiation between populations, similar to the fixation index F_{ST}, assuming that extreme values correspond to candidate regions within the genome (Duforet-Frebourg et al., 2016). While high levels of differentiation between populations may result from various processes, it is assumed that individual adaptation to environmental conditions (in the case of livestock, production-related) is one possible explanation for these changes, specifically within certain genomic regions. An alternative method of detecting selection signals based on this theoretical basis utilizes principal component analysis (Duforet-Frebourg et al., 2014).

Principal component analysis (PCA) is arguably the most widely used multivariate statistical method and has applications across various scientific fields, including genetics. PCA represents a method to visualize high-dimensional data, such as genomic information on individuals or populations, in smaller dimensions. It has become popular in genomics, particularly as a tool to reduce the number of initially correlated variables (allele frequencies) to a smaller set of linearly uncorrelated (independent) variables, also known as principal components, which explain variance in the dataset and thus can be used to represent relationships among individuals or populations (Duforet-Frebourg et al., 2016).

In the context of detecting selection signals, PCA has three main advantages over other approaches: it operates at the individual level, the computation time is relatively short compared to methods that use MCMC algorithms, and candidate loci potentially associated with the local adaptation of populations to environmental conditions correspond to individual principal components (Duforet-Frebourg et al., 2016; Luu et al., 2017). For example, screening of

selection signals through PCA was applied in cattle (Moravčíková et al., 2018; Kasarda et al., 2021a).

Variation in Linkage Disequilibrium and Integrated Haplotype Score

Several statistical tests have been developed to identify selection signals resulting from changes in linkage disequilibrium (LD) within population genomes (Sabeti et al., 2002; Kim and Nielsen, 2004; Voight et al., 2006; Kimura et al., 2007). However, this type of selection signal tends to be temporary, as recombination may alter the sequence of the selected locus before it becomes fixed in the population's gene pool in some cases.

One approach to detecting selection signals reflecting LD changes is the long-range haplotype (LRH) test, which assesses the relationship between allele frequencies and LD levels. This test is based on identifying target haplotypes through SNP marker genotyping within short genome segments where no recombination occurs. Subsequently, additional SNP markers are analyzed with increasing distance from the target haplotypes to assess the decline in LD as genetic distance increases. The overall LD level, with increasing distance from the target haplotypes, is evaluated by calculating the extended haplotype homozygosity (EHH) value, representing the probability that two chromosomes carrying specific target haplotypes remain homozygous for the entire region from the target distance to distance x. The relative EHH (REHH) value is used to compare the decline of EHH for a specific target haplotype with the decline of EHH across all other haplotypes. Selection signals are then identified by comparing REHH value and frequency of each target haplotype with REHH values and frequencies of other target haplotypes. A target haplotype with a high REHH value and a high population frequency can be considered a signal of positive selection (Sabeti et al., 2002). Another test to identify selection signals is based on the integrated haplotype score (iHS). This test has been developed especially with regard to the increasing genotyping of populations using SNP chips. The iHS value can simply be defined as the extent to which a haplotype composed of specific SNP markers differs from the rest of the genome. In this approach, each SNP is scored as a target and the test starts by calculating the EHH value for each SNP marker. SNP markers as biallelic loci can be either inherited (ancestral) or derived. The calculation determines the integral of the observed decrease in EHH of the target SNP marker until it reaches an EHH value of 0.05. This value is considered the integrated EHH (iHH) and is identified as $_{iHHA}$ or $_{iHHD}$ depending on whether it was calculated from the ancestral (A) or derived allele (D) of the target SNP marker. The obtained value is standardized for direct comparison with other SNP markers regardless of their allele frequencies (Voight et al., 2006). Due to its applicability to robust genomic data, the iHS score has been successfully applied to various breeds and livestock species (Kukučková et al., 2017b; Moravčíková et al., 2019c).

Because LRH and iHS tests are based on allele frequencies in target haplotypes, they are limited in terms of detecting selection signals, especially if the allele under selection is fixed in the genome in a homozygous form. If such an allele is fixed in homozygous form in one population but remains polymorphic in another, the LRH test can only be based on a comparison between these populations. The XP-EHH statistic is defined as the normalized log ratio between $_{IA}$ and $_{IB}$, where $_{IA}$ is the integral of the observed decrease in EHH from the target SNP to SNP X (which has an EHH value as close as possible to 0.04 in both populations) in population A, and $_{IB}$ expresses the same calculation, but for population B. A very similar principle is used in the method referred to as the *ln(Rsb)* statistic (Sabeti et al., 2007; Tang et al., 2007).

Another method for detecting selection signals reflecting LD changes is based on determining haplotype allelic categories (HAC). This metric is defined as the sum of allelic differences between reference allelic categories and individual haplotypes within the sample. Positive values indicate positive selection in the specified genomic region (Hussin et al., 2010).

Distribution of Runs of Homozygosity (ROH) in the Genome

Screening for selection signals derived from ROH distribution in the genomes of livestock and companion animals is based on the assumption that genomic regions exhibiting strong selection signals result from increased local homozygosity due to intensive breeding for traits defined in the breed standard (Curik et al., 2014; Kim et al., 2017). The resulting ROH segments located close to each other within the genome are composed of alleles inherited from common ancestors and passed down through generations in an unchanged form (Biscarini et al., 2014). Identifying and analyzing these segments provides insights into the changes that have shaped the genome of breeds or populations and serves as a valuable tool for assessing the demographic history of each breed's development. This approach to tracking the effect of selection on population genome structure has been used to screen selection signals in many livestock and companion animal species (e.g., Moravčíková et al., 2019c; Kasarda et al., 2021b; Moravčíková et al., 2021).

Literature

BISCARINI, F. et al.. 2014. Applying runs of homozygosity to the detection of associations between genotype and phenotype in farm animals. In Proceedings of the 10th World Congress on Genetics Applied to Livestock Production, vol. 675, pp. 1-3.

CURIK, I. – FERENČAKOVIĆ, M. – SÖLKNER, J. 2014. Inbreeding and runs of homozygosity: A possible solution to an old problem. In Livestock Science, vol. 166, pp. 26-34.

DUFORET-FREBOURG, N. – BAZIN, E. – BLUM, M.G.B. 2014. Genome scans for detecting footprints of local adaptation using a bayesian factor model. In *Mol Biol Evol*, vol. 31, pp. 2483-2495.

DUFORET-FREBOURG, N. et al. 2016. Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data. In *Mol Biol Evol*, vol. 33, pp. 1082-1093.

FUMAGALLI, M. et al. 2013. Quantifying population genetic differentiation from next-generation sequencing data. In *Genetics*, vol. 195, pp. 979-992.

HAYES, B.J. – GODDARD, M.E. 2010. Genome-wide association and genomic selection in animal breeding. In Genome, vol. 53, pp. 876-883.

HUSSIN, J. et al. 2010. Haplotype allelic classes for detecting ongoing positive selection. In BMC Bioinformatics, vol. 11, e65.

JOMBART, T. – AHMED, I. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. In *Bioinformatics*, vol. 27, pp. 3070–3071.

JOMBART, T. – COLLINS, C. 2015. *A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0.0*. London : MRC Centre for Outbreak Analysis and Modelling. 43 p.

KADLEČÍK, O. – MORAVČÍKOVÁ, N. – KASARDA, R. 2017. *Biodiverzita populácií zvierat.* Nitra : Slovenská poľnohospodárska univerzita, 285 s. ISBN 978-80-552-1763-5.

KASARDA, R. et al. 2019. Level of inbreeding in Norik of muran horse: Pedigree vs. Genomic data. In *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, vol. 67, pp. 1457-1463.

KASARDA, R. et al. 2021a. Food resources biodiversity: The case of local cattle in Slovakia. In Sustainability, vol. 13, pp. 1296.

KASARDA, R. et al. 2021b. The evaluation of genomic diversity and selection signals in the autochthonous Slovak Spotted cattle. In *Czech J Anim Sci*, vol. 66, pp. 251-261.

KIM, Y. – NIELSEN, R. (2004). Linkage disequilibrium as a signature of selective sweeps. In *Genetics*, vol. 167, pp. 1513-1524.

KIM, S.J. et al. (2017a). Cattle genome-wide analysis reveals genetic signatures in trypanotolerant N'Dama. In *BMC Genomics*, vol. 18, pp. 1-18.

KIMURA, R. et al. 2007. A practical genome scan for population-specific strong selective sweeps that have reached fixation. In *PLoS One*, vol. 2, pp. e286.

KUKUČKOVÁ, V. – MORAVČÍKOVÁ, N. – FERENČAKOVIĆ, M. et al. 2017a. Genomic characterization of Pinzgau cattle: genetic conservation and breeding perspectives. In *Conservation Genetics*, vol. 18, pp. 893-910.

KUKUČKOVÁ, V. – MORAVČÍKOVÁ, N. – KASARDA, R. 2017b. Variation in linkage disequilibrium patterns between populations of different production types. In *Agriculturae conspectus scientificus*, vol. 82, pp. 105-109.

LENSTRA, J.A. et al. 2012. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. In *Animal Genetics*, vol. 43, pp. 483-502.

LUU, K. – BAZIN, E. – BLUM, M.G. 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. In *Mol Ecol Resour*, vol. 17, pp. 67-77.

MORAVČÍKOVÁ, N. et al. 2015. Estimation of genomic variation in cervids using cross-species application of SNP arrays. In *Poljoprivreda*, vol. 21, pp. 33-36.

MORAVČÍKOVÁ, N. et al. 2017. Effective population size and genomic inbreeding in Slovak Pinzgau cattle. In *Agriculturae conspectus scientificus*, vol. 82, pp. 97-100.

MORAVČÍKOVÁ, N. et al. 2018b. Genomic signatures of positive selection with respect to the immunity - related genes in cattle. In *11th world genetic applied to livestock production*. Auckland : University of Auckland, 65 p.

MORAVČÍKOVÁ, N. et al. 2019a. Analysis of selection signatures in the beef cattle genome. *In Czech J Anim Sci*, vol 64, pp. 491-503.

MORAVČÍKOVÁ, N. et al. 2019b. Runs of homozygosity as footprints of selection in the norik of muran horse genome. In *Acta Univ Agric Silvic Mendel Brun*, vol. 67, pp. 1165-1170.

MORAVČÍKOVÁ, N. et al. 2019c. Genomic signatures of selection in cattle throught variation of allele frequencies and linkage disequilibrium. In *Journal of Central European Agriculture*, vol. 20, pp. 576-580.

MORAVČÍKOVÁ, N. et al. 2021. Czechoslovakian wolfdog genomic divergence from its ancestors canis lupus, german shepherd dog, and different sheepdogs of european origin. In *Genes*, vol. 12, pp. 832.

NEI, M. 1972. Genetic distance between populations. In Am Nat, vol. 106, pp. 283-285.

NEUDITSCHKO, M. – KHATKAR, M.S. – RAADSMA, H.W. 2012. NetView: a high-definition network-visualisation approach to detect fi-ne-scale population structures from genome-wide patterns of variation. In *PLoS One*, vol. 7, e48375.

PRITCHARD, J.K. – STEPHENS, M. – DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. Genetics portions from molecular data. In *Mol Biol Evol*, vol. 15, pp. 1298-1311.

QANBARI, S. – SIMIANER, H. 2014. Mapping signatures of positive selection in the genome of livestock. In *Livestock Science*, vol. 166, pp. 133-143.

QANBARI, S. et al. 2011. Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. In *BMC Genomics*, vol. 12, pp. 318.

SABETI, P.C. et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. In *Nature*, vol. 419, pp. 832-837.

SABETI, P.C. et al. 2007. Genome-wide detection and characterization of positive selection in human populations. In *Nature*, vol. 449, pp. 913-918.

TANG, K. – THORNTON, K.R. – STONEKING, M. 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. In *PLoS Biology*, vol. 5, pp. e171.

TORO, M.A. et al. 2009. Molecular characterization of breeds and its use in conservation. In *Livest Sci*, vol. 120, pp. 174-195.

VOIGHT, B. F. et al. 2006. A map of recent positive selection in the human genome. In *PLoS Biology*, vol. 4, pp. e72.

WEIR, B.S. – COCKERHAM, C.C. 1984. Estimating F-statistics for the analysis of population structure. In *Evolution*, vol. 38, pp. 1358-1370.

WRIGHT, S. 1940. Breeding structure of populations in relation to speciation. In *Am Nat*, vol. 74, pp. 232-248.