Topic 2: Analysis of the AnGR biodiversity status using genomic data
Lecture

In the context of the evaluation of genomic data, it is first necessary to briefly describe how such information can be obtained. To analyse the genome, we can use a variety of tools, including single genetic markers, genotyping chips or whole genome sequencing.
The difference between these methods is related both to the laboratory procedure for their determination and to the amount of genome data that we obtain with them.
What can we understand by a genetic marker. It is any characteristic trait or manifestation of an organism that can be used to identify a specific chromosome, cell or individual. The term genetic marker may refer to a gene, a short segment of DNA, or other manifestations of genotype, chromosomes or karyotype. However, it is important to remember that a genetic marker is usually a polymorphic variant that shows mendelistic inheritance and is correlated with variation in a phenotypic trait (character) that is of importance, for example, from a breeding point of view.
In terms of livestock production traits, both candidate genes are monitored because their alleles and genotypes influence the formation of quantitative traits and, at the same time, loci for quantitative traits.
The advantage of DNA markers is mainly that they are directly detectable in nucleotide sequences, show an increased level of polymorphism and dominant or codominant inheritance. DNA markers are relatively common in the genome and can be tested relatively easily and rapidly with a high degree of repeatability.
The most commonly used genetic markers are nowadays single nucleotide polymorphisms, called SNPs. SNPs are usually generated by a point mutation, e.g. a nucleotide substitution in the DNA at a particular site. Compared to other types of genetic markers, they occur frequently in the genome, every 100 to 300 base pairs. Mutations that occur at a frequency of more than 1% in a given population, i.e. a minor or less frequent allele is present in the genotype of at least one 1% of individuals belonging to that population, are usually considered SNPs. It is a biallelic marker, i.e. within a population we recognize only two alleles for SNP, namely dominant and recessive. The term dominant indicates that it is the predominant allele in individuals in a given population. A minor allele is, on the other hand, an allele that occurs less frequently in the genotypes of individuals within a population. However, it is important to note that even if an allele is dominant in one population, it may not be dominant in another population with different genetic origin. SNP markers have a wide range of applications, from biodiversity evaluation to genomic selection.
Whole-genome sequencing is the term used to refer to the process of determining the exact order of nucleotides in a strand of a DNA molecule, i.e. determining its primary structure. The classical methods are Maxam-Gilbert and Sanger methods, from which the currently used next-generation sequencing methods are derived. Within the NGS methods there are several platforms which, even though they differ in their technological approach, yield comparable outputs.
Even though the cost of whole-genome sequencing has decreased significantly compared to the previous period, it is still high if we want to obtain whole-population data, especially for high-coverage sequencing. For this reason, SNP genotyping chips are now being used in population-wide studies, which allow to obtain information on a large number of SNP markers uniformly distributed across the genome at a lower cost. SNP chips allow genotyping from a few thousand up to 700 000 SNP markers. They are available for most livestock and companion animal species. The information obtained in this way can be used for a variety of purposes, including testing parentage, genomic diversity status, genome-wide association studies or estimation of genomic breeding values.

In the following slides, we will discuss indicators that are used to estimate the biodiversity status of animal genetic resources based on genomic data analysis. The first indicator is genome homozygosity and genomic inbreeding. In the context of genome homozygosity, two terms you will often find in the literature: autozygosity and runs of homozygosity, abbreviated as ROH. Basically, autozygosity reflects all alleles or chromosomal segments of DNA that are identical by descent, i.e., coming from a common ancestor. Runs of homozygosity are considered to be all genomic regions with a specific number of consecutive homozygous genotypes or, when talking about SNP marker testing, all homozygous SNP markers. The distribution, number and length of runs of homozygosity depend on various factors affecting the livestock genome. The most significant in this context can be considered to be artifical selection and the intensity of inbreeding. The length of the ROH segments in an individual's genome itself corresponds to the distance of the ancestors in the individual's pedigree. If the parents of an individual have a common ancestor, their genome will share the same genetic variants in certain regions, i.e. such parents will be identical by descent. If both parents transfer the same region to the offspring, then the offspring will be homozygous for the genetic variants, thus creating an ROH region in the offspring's genome. This assumption is the basis of the approach for estimating the genomic inbreeding coefficient through the coverage of the genome by runs of homozygosity.

However, information about the occurrence and length of ROH segments in the genome can be used not only to estimate genomic inbreeding but also to test the impact of artificial selection on specific regions in the genome or to identify causal variants involved in the control of preferred phenotypic traits and characteristics.

In this slide, you can see in the first part the formula for estimating genomic inbreeding, referred to as FROH, where the numerator expresses the total length of homozygous segments in an individual's genome and the denominator the total genome length derived from the physical position of the markers tested. FROH allows to establish the trend of inbreeding, where segments longer than 4 Mb reflect autozygous regions derived from ancestors approximately 12 generations ago, segments longer than 8 Mb are derived from ancestors 6 generations ago, and segments longer than 16 Mb correspond to the proportion of autozygosity inherited from ancestors from the last 3 generations. Similar to pedigree inbreeding, the genomic inbreeding values range from 0 to 1 or, in percentage terms, from 0 to 100%. Information on the increase in inbreeding per generation and the overall inbreeding coefficient is important both in terms of the occurrence of inbreeding depression and at the same time the survival of the population in the long term. One of the reasons is that the accumulation of inbreeding across generations leads to a reduction in genetic diversity. It is generally accepted that the increase in inbreeding per generation should not exceed 1% in small populations and 4% in large populations. The most commonly used programs to estimate genomic inbreeding coefficients include detectRUNS, plink or cgaTOH. In the figure you can see the results from a comparative analysis of the inbreeding coefficient in 15 cattle breeds based on ROH segments longer than 4 and 8 Mbp.

Another indicator of biodiversity that we can estimate by testing genomic markers is the linkage disequilibrium between SNP markers in the genome and, consequently, the effective population size based on it. The term linkage disequilibrium essentially refers to a non-random relationship or association between alleles of different SNP markers in the genome of the evaluated population, which is likely to be due to selection, mating system, recombination, or genetic drift. As a result, this means that such genetic variants can produce specific combinations of genotypes in a population, also called haplotypes. Information on the level of linkage disequilibrium can be used to assess the evolutionary shaping of populations, to estimate effective size, or, as in the case of ROH segments, to test for the occurrence of specific genetic variants that have been strongly influenced by artificial or natural selection.

The most commonly used formula for calculating linkage disequilibrium between SNP markers is shown in the slide. However, in addition to this formula proposed by Hill and Robertson, there are other modifications of it that take into account, for example, the mutation rate or the nature of the genetic markers tested (biallelic or more multiallelic). The range of values in the case of linkage disequilibrium ranges from 0 to 1, with 0 indicating linkage equilibrium between markers and 1 indicating complete linkage disequilibrium.

Effective population size essentially reflects the number of individuals that are active in reproduction in a given population, i.e. can provide individuals for the next generation. Estimation of this parameter in the case of genomic data is most often based on its relation to the degree of linkage disequilibrium in the genome, where it is possible to test not only the current effective size but also the trend of its evolution in the past.

The effective population size, abbreviated as Ne, can be determined using, for example, the formula proposed by Corbin et al. shown in this slide. This formula takes into account the inheritance model, the physical distance between SNP markers, or the intensity of mutations. In this case, the historical effective size is estimated as a function of time and the physical distance between the two markers, assuming a constant linear growth of Ne with time expressed by past generations. The figure on the right shows representative results of the analysis of the effective population size trend in two cattle breeds, the Slovak Spotted and the Slovak Pinzgau. Similar to pedigree information, the effective population size can range from 0 to n. It is generally accepted that the effective population size should not be less than 50 individuals in the case of small populations or 100 individuals in the case of large populations. In terms of long-term sustainability, the effective population size should be at least 500 individuals. In the case of genomic data, programs such as SneP or GONE can be used for the calculation.

In animal genetic resources, indicators describing population structure at intra- and interpopulation level are often evaluated. In this context, genetic distances are most often analysed, as they reflect the degree of genetic differences between individuals, populations or species. The most commonly discussed in the literature are Nei's genetic distances, Wright's fixation index FST, principal component analysis, or methods quantifying the degree of genetic admixture and gene flow between populations.

Nei's genetic distance theory assumes that if two populations showing low genetic distances are similar, they share common ancestors with a high degree of confidence. For this reason, this indicator can also be considered as the molecular equivalent of the relatedness coefficient calculated on the basis of pedigree information. You can see the formula for calculating the standard Nei's genetic distance on the slide. The minimum value that the Nei genetic distance can take is 0. This value means that individuals or populations have the same variants (alleles or genotypes) in the genome, i.e. they are genetically identical. The maximum value that the Nei's genetic distance can take is 1. This value reflects the fact that due to completely different genetic variants, individuals or populations are genetically different and, we can say, unrelated. To calculate Nei's genetic distances we can use the R package StAMPP, Poppr or other programs.

Compared to Nei's genetic distance, Wright's FST fixation index only allows to estimate the level of diversity at the population level. This index is essentially an indicator of the intensity of population fragmentation, expressed as a decrease in heterozygosity in subpopulations due to the effect of genetic drift. Hence, to calculate this index we need to have information about the expected heterozygosity within the metapopulation and the average heterozygosity within the subpopulations as you can see in the formula on the slide. Wright's fixation index FST takes values from 0 to 1, and the interpretation of the values is similar to that of Nei's genetic distances. If the value of the index is equal to 0 the populations are genetically identical and opposite if the value is equal to 1 the populations are genetically distinct. In real livestock populations, the value of this index usually ranges from 0 to 0.5, of course, if we are testing a

single species. Populations with an FST value greater than 0.25 are considered to be genetically differentiated.
Other commonly used approaches to evaluate population structure and genetic relationships between populations include principal component analysis and Bayesian analysis of genetic admixture. Principal component analysis is a popular multivariate statistical method that has found applications in various scientific fields, including population genetics. Simply said, this analysis is used to represent high-dimensional data, e.g. genomic information about individuals or populations, in fewer dimensions. Bayesian statistics is a method that is used in other scientific disciplines as well. This statistic operates with conditional probability and allows the probability of the initial hypothesis to be refined in sequence as other relevant facts appear.
This slide shows representative results of testing the proportion of genetic admixture, principal component analysis and gene flow. In the case of the first figure, this is a Bayesian analysis of admixture within 15 cattle breeds, with the proportion of admixture within breeds represented by different colours of the lines. The second figure on the left shows representative results of the principal component analysis, with the degree of admixture being best seen in part D through the overlapping peaks of different colours. In the third figure on the right, we can see the results of the genetic admixture analysis of the 4 breeds and the numerical representation of the gene flow between their gene pools.
As I mentioned before in the case of ROH segments and linkage disequilibrium, in addition to standard indicators such as Ne and F, we can also evaluate the effect of selection on the genomic structure or identify specific genetic variants under strong selection pressure. Unlike whole-genome association studies, this approach does not require access to phenotypic information about individuals. It is essentially the identification of so-called selection signals, the occurrence of which depends on a variety of factors, in livestock mainly artificial selection. Two groups of methods are basically used for this purpose, methods testing differences between populations or breeds and methods analysing intra-population differences.
In terms of interpopulation differences, whole-genome screening of the FST fixation index, analysis of variability in linkage disequilibrium, or calculation of integrated haplotype scores are most commonly used to identify selection signals. A number of programs exist for this purpose, such as PLINK, varLD or the R package rehh. The figure on the right shows the result of the analysis of testing selection signals reflecting differences between Slovak Spotted and Slovak Pinzgau cattle. The strongest selection signals were found in the casein gene family and the KIT and KDR genes responsible for spotting.
Within intra-population differences, selection signals are usually determined based on the distribution of runs of homozygosity or variation in linkage disequilibrium. The same programs as for the previous methods can be used for the calculation. The figure on the right shows the results of testing ROH segments distribution in the genome of the Slovak Spotted and Slovak Pinzgau cattle. The results show that, similar to the previous approach, the selection signals are strongest in the genomic region of casein family genes.
If you have any further questions about the presentation, please contact me at the email address shown in the last slide. Information about the project, including access to other presentations, can be found by scanning the barcode on the last slide. Thank you for your attention.