Study of genetic diversity and population structure using mtDNA and nuclear microsatellite markers in the honey bee

ISAGREED MENDELU PhD module 1 ENG

Hello, in this lecture for PhD students we will focus on the possibilities of assessing genetic diversity and population structure using mitochondrial DNA and nuclear microsatellite markers applied to the honey bee. The lecture is part of Module 1 - Animal Genetics. The creation of this presentation was supported by an Erasmus+, KA2 grant within the ISAGREED project Innovation of the content and structure of study programs in the field of management of animal genetic and food resources using digitization.

This lecture will cover topics such as genetic data acquisition, assessment of genetic variability using mitochondrial DNA sequences and assessment of genetic variability using nuclear STRs or microsatellite markers.

Why is genetic diversity important? Genetic diversity is important for the health of populations. It is a key source of the ability to build tolerance or resistance to current and future diseases, pathogens and predators. The current state of bee populations can be attributed in part to a reduction in diversity. Bee diversity has been assessed using morphometric traits such as wing parameters, pigmentation, etc.

The honey bee (Apis mellifera) is now known to comprise 31 subspecies, breeds or races. DNA analysis, particularly of mitochondrial origin, has facilitated the description of evolutionary lineages including the Western Mediterranean type M, the Northern Mediterranean type C, the African lineage A and the Oriental lineage O.

The honey bee genome has been completely sequenced on multiple occasions. The individual chromosomes are visible, and the penultimate column illustrates the size of each chromosome in terms of the number of base pairs. A total of 12,398 genes have been described or are estimated to exist in the honey bee genome. Of these, 9,935 genes code for some kind of protein, while 2,421 genes do not code for proteins, but rather code for other RNAs, such as transfer RNAs or other small nuclear RNAs.

The mitochondrial DNA of most species is estimated to be within the range of approximately 16 to 20 kilobases. The mitochondrial genome of the western honey bee is estimated to comprise approximately 16,500 base pairs. In the reference genome NC001566, the mitochondrial DNA is 16,343 base pairs in size. In the complete genome of the Carpathian mitochondrial DNA, the size of the mitochondrial DNA is 16,358 base pairs.

The mitochondrial DNA of the honey bee contains 13 genes that encode proteins, as well as 22 genes that encode tRNA and 2 genes that encode ribosomal RNA. In particular, the barcoding sequence, which is the cytochrome oxidase 1 sequence, and the so-called intergenic region, which includes part of the tRNA gene for leucine and cytochrome oxidase 2, are employed for phylogenetic and phylogeographic analyses.

What degree of variability can be observed at the DNA level, and which molecular genetic markers exist? At present, the most frequently utilized are biallelic single nucleotide

polymorphisms, which can be identified in both coding and non-coding regions. Additionally, data on insertions and deletions, defined as the presence or absence of a base, can be employed. These are commonly referred to as indels. The images on the right illustrate this: the top row depicts SNP markers, while the bottom row displays deletions of cytosine in the ACA sequence region.

Following the isolation of the DNA from the bee sample and the amplification of a specific small section, for example one of the two genes mentioned above, sequencing is conducted using a capillary electrophoresis-based sequencer. The resulting identification of the individual bases in the sequence is illustrated in the figure.

The individual peaks are represented by the colors used to identify the individual bases.

Two mitochondrial DNA sequences were employed for the purpose of identifying subtypes, mitotypes or haplotypes within the lineage. These are the aforementioned intergenic region, tRNA leucine-cox2, and cytochrome oxidase 1 region. This section comprises two mitochondrial genes: the transfer RNA for leucine and the cytochrome oxidase 2. This sequence is distinguished by a high mutation content and notable variations in nucleotide length and composition across honey bee populations. This amplicon is cleaved by the restriction endonuclease DraI, which specifically recognizes the TTTAAA sequence to identify each lineage. The second sequence is the sequence for the barcoding region and this is part of the cytochrome oxidase 1 (cox1) gene. This sequence is compared to sequences stored in databases such as the BOLT system or GenBank. The DNA fragment is highly conserved within taxa and is often used to distinguish taxa and species.

The tRNA-leucine-cox2 sequence structure allows for the identification of distinct evolutionary lineages within the honey bee. The C lineage, which encompasses the honey bee (Apis mellifera) as well as the A. m. ligustica, A. m. macedonica, and other related species, is characterized by the presence of a single copy of the Q sequence. The aforementioned lines contain one to two copies of the aforementioned Q sequence, in addition to the so-called P0 segment. The M lineage, which is the original black bee (Apis mellifera mellifera), which is no longer found in the Czech Republic, may contain one, two, or three repeats of the aforementioned Q sequence, in addition to the so-called P sequence. By sequencing and comparing individual sequences, it is possible to identify an evolutionary lineage in each bee.

The identification of a particular lineage is possible through cleavage with the restriction enzyme *Dra*I, which recognizes the altered TTTAAA sequence. Once this change occurs as a result of a mutation, this enzyme is unable to recognize this sequence, instead it is unable to cleave it. Using a classical PCR reaction, based on the length of the individual fragments, it is possible to distinguish between variants such as C and A1 or A4.

The second option is to obtain the entire sequence of a given segment by sequencing and subsequently analyzing it. The following example illustrates the sequencing and subsequent analysis of tRNA-leucine-cox2 sequences from several individuals.

Some software, such as UniPro Ugene, enables the user to perform the cleavage with DraI restriction enzyme in silico, that is, on a computer. The following example illustrates the

cleavage of a sequence belonging to the C lineage, which contains three cleavage sites, resulting in three fragments of specific lengths.

In another sample, only two cleavage sites were identified, and the length of the fragments suggests that this is a bee belonging to the A lineage, or African lineage.

Subsequently, the sequences obtained from the larger population are compared using the method of multiple sequential alignment, MSA. This process could be completed manually; however, software has been developed with algorithms that facilitate this task. Some of these programs are accessible online, for example on the European Bioinformatics Institute servers. For our purposes, Kalign was the most suitable. However, there are other tools, such as Clustal Omega, MAFFT and so on. Additionally, there are programs that can be downloaded and installed to perform this analysis, such as MEGA or the Unipro Ugene.

The DnaSP program was employed to identify DNA polymorphisms and haplotypes in both mitochondrial DNA regions, utilizing all sequences from multiple sequence alignments in FASTA format.

Moreover, nucleotide substitutions and insertions/deletions for each haplotype were compared with the reference genome. To identify specific haplotypes in the tRNA Leucine-cox2 lineage C and A, reference sequences with 100% identity were further searched using BLAST local pairwise alignment tools against sequences found in the National Center for Biotechnology Information (NCBI) database at the US GenBank. BLAST was also employed to verify the cox1 haplotypes, with the sequences subsequently validated using the BOLD database based on multiple alignments using Kalign, necessitating additional manual refinement.

A total of 13 haplotypes were identified, 3 of which belonged to the A lineage and the rest to the C lineage. The most prevalent haplotype was C1a, which is typical for Apis mellifera ligustica, the Italian bee.

The table illustrates the classification of individual haplotypes into C and A lineages based on DraI spectrum cleavage and sequencing. The individual haplotypes and their sequences have been uploaded to the GenBank database on the NCBI server. In the third column, the reference sequences are displayed. Additionally, the numbers and lengths of the fragments produced by the cleavage are shown, which also demonstrate considerable variability. In the last two columns, the identification or comparison with other sequences in the GenBank database is presented, where sequences with 100% identity to our sequences have been selected.

It is notable that all haplotypes belonging to the C lineage have been previously described in Apis mellifera carnica. Additionally, three distinct African haplotypes were identified as Apis mellifera iberica. However, a single sequence exhibiting complete identity was not assigned to any particular subspecies.

This table illustrates the identification of the most significant polymorphic sites, indicating the bases present in each haplotype. Ten positions were found to exhibit mainly single-nucleotide polymorphisms (SNPs) and deletions. It is notable that position 50 displays polymorphism, with the standard allele identified as C. The remaining haplotypes at this position exhibited a deletion.

Similarly, the cox1 sequence was analyzed, whereby 13 different haplotypes for barcoding were identified. As with the previous analysis, individual SNP mutations were observed. No insertions or deletions were identified within the barcoding sequence; only SNP substitutions were present.

The tables below present the results of the haplotype frequencies in the tRNA leucine-cox2 gene and in the cox1 gene in the Czech Republic. It can be observed that there are a number of haplotypes that are relatively well represented, such as C1A, C2L, C2E, and C2C. In contrast, there are haplotypes that have been identified in only one or a few individuals. A similar situation was observed in the cox1 genes, where the first four haplotypes were the most frequently represented, while the others were present in minority or individual samples.

From a population of bees in the Czech Republic comprising over 300 samples, certain characteristics were calculated, including haplotype diversity parameters in the tRNA Leucine-cox2 and cox1 sequences. The genetic diversity indices, namely haplotype diversity (HD), molecular diversity (π) and Tajima's D, were estimated. These indices were evaluated using the Pegas package in R program, although alternative software such as DnaSP, MEGA or Arlequin can also be employed.

Additionally, the so-called haplotype networks were determined. We used the Randomised Minimum Spanning Tree (RMSAT) method, which takes into account frequencies and relationships between haplotypes. These haplotype networks were processed using the Pegas package in R. However, other programs such as PopArt and others can be used.

Here we see the result of the haplotype network analysis for 308 individuals in the tRNA Leucine-cox2 sequence. The most common haplotype is C1A, followed by C2E, which has two point mutations, and C2C, which is the third most common haplotype in the Czech Republic. Each color in the circle represents a specific region from which the bee was obtained. The aim was to cover the whole Czech Republic, with sampling evenly distributed across all regions.

This slide presents the analysis of the haplotype network for the cox1 sequence. As observed previously, if a haplotype was sufficiently abundant, it occurred in almost all regions. This is exemplified by HpB02, HpB03, HpB01 and HpB04. Conversely, the other haplotypes were present in a few individuals or in only one individual.

Subsequently, further phylogenetic analysis was conducted on these sequences. Following the completion of the MSA and multiple sequence alignment, phylogenetic tree generation was conducted using the Maximum Likelihood method and the Tamura-Nei model in Mega X software. This method entailed the construction of a bootstrap consensus tree based on 10,000 replicates. The individual branches correspond to partitions produced in less than 50% of bootstrap replicates, as well as the percentage of replication trees that clustered related haplotypes in the bootstrap test.

The initial tree for the heuristic search was obtained automatically by applying the Neighbor Joining and BioNJ algorithms to the pairwise distance matrix estimated by the Tamura-Nei

model and then selecting the topology with the highest log-likelihood value. As a result, the following phylogenetic trees were obtained.

The phylogenetic tree based on the analysis of the tRNA leucine-cox2 sequence is displayed on the left. The phylogenetic tree based on the cox1 sequence is displayed on the right. It can be observed that the bees from lineage A cluster together, in contrast to the other bees, namely those from lineage C, which are marked in red.

The second type of markers used to assess genetic variability are microsatellites. These are polymorphisms that occur exclusively in nuclear DNA. These polymorphisms are characterized by the repetition of a particular motif, such as GC, in a series of units called tandem repeats. Each allele is referred to by the length of this repeat.

To illustrate, we have an allele with eight repeats, another with three repeats, and the last with ten repeats. The designation of the allele is dependent on the region in which the microsatellite is located, which is bounded by primers. The length of the segment containing the microsatellite can be determined using a sequencer and fragmentation analysis. The figure below illustrates the genotyping for a particular microsatellite, which in this case is characterised by three alleles, numbered 156, 152 and 142.

We can see that microsatellites are very polymorphic, they can have not just three alleles, but 20 alleles, which in a population can mean a large number of different combinations of genotypes. So they are useful for assessing diversity in populations. Here is an example of variability in two populations. The population on the left has little variability, containing only three types of alleles and a large number of homozygous individuals. The second population on the right contains a large number of alleles and may often contain heterozygous genotypes.

Why are microsatellite markers still used when we have whole genome sequences? Microsatellite markers are relatively inexpensive, and their identification provides multilocus genotypic information. They can easily be used to estimate the genetic diversity of populations and structures, which is also important in conservation genetics and breeding.

The only method used to determine genotypes is fragmentation analysis, which is performed in sequencers using capillary electrophoresis. The fragments are separated according to their size, and the sensor detects the passage of the molecules, their colour and signal intensity over time, providing information about the length of the fragments.

The instrument used is a genetic analyser. In our case, we used the ABIPrism 3500 genetic analyser. Fragment sizes were accurately determined using GeneScan software and genotypes were determined using GeneMapper software.

As we were looking at 22 microsatellite loci, we grouped certain microsatellites in a single multiplex reaction. We were able to identify several microsatellites under the same conditions, distinguished by different colors.

The result of the analysis is a display where, for a particular microsatellite locus, we see the color-coded peaks, which are identified fragments of a particular length.

The figure shows the evaluation of the variability of the genotyping of microsatellite loci in three individuals. We can see that different alleles can be present at certain positions of the region, which allows easy discrimination of individuals.

After determining the genotypes at all loci and for all individuals in the population, the next step is to perform a diversity analysis, i.e. to determine diversity parameters such as the number of alleles Na, the effective number of alleles Ne, the Shannon information index I, the observed and expected heterozygosity, Ho and He, respectively, and the unbiased expected heterozygosity uHe, and the so-called fixation index F. We used the GenAlEx program, which runs in Microsoft Excel, but the data and parameters can also be calculated using the diveRsity package in R.

We see that the expected heterozygosity averaged over all loci combined is 0.579 and the actual observed heterozygosity is 0.556. This is a relatively high heterozygosity which characterizes this bee population as sufficiently divergent.

Since we knew which area (district) of the Czech Republic each individual came from, we divided the population of the Czech Republic into 77 districts, which characterize the geographical areas. This allowed us to calculate the so-called Wright's F-statistics, FST, FIS and FIT, and the so-called analysis of molecular variance, which determines the proportion of variability between populations, between individuals within populations and within individuals.

The table on the left characterises the individual loci and the mean values of the F-statistics. The FST is most interesting because it determines the degree of variation between subpopulations, i.e. districts. The value of 0.086 is not very high, but it shows some diversification.

In the table on the right we see the result of the analysis of molecular variance, where in the last column the variation between regional populations (districts) accounts for only 1% of the total variation. But even 1-3% of variability between geographical areas are common figures according to other publications. Variability between individuals within populations is expressed as 6%, and within-individual variability accounts for most of the variability within populations.

Paired Nei's and paired FST genetic distances were calculated using the GenAlEx program. Paired values of these distances were used for principal component analysis (PCA) calculations. The top graph depicts the distances between the first and second components for each district. The bottom graph presents a calculation comparing the first and second components using paired Fst distances. We can see that, for example, the district of Děčin DC or Frydek Místek FM and similar other districts are a little more distant from the central cluster, and there are some distances between them, but they are not that significant. So there are areas that are more distinct from other areas.

The Bayesian clustering method STRUCTURE ver, 2.3.4 was used to analyze the genetic diversity and admixture rate of honey bee populations. Ten independent simulations were run, each involving 10,000 burn-in steps followed by 100,000 iterations of Markov chain Monte

Carlo (MCMC). We then used the Clumpak and Structure Selector programs, which implement Evann's method and Puechmaille's method, respectively, to determine the optimal number of clusters (K) that best fit the data (ΔK, MedMeadK, MaxMeadK, MedMeanK and MaxMeanK). In the case of this population, both methods determined that there are three genetically distinct populations in the honey bee population in the Czech Republic based on these 22 microsatellite markers and that each individual can be assigned to one of these three groups with some high degree of probability.

Other methods can also be used to determine the genetic structure, such as the discriminant analysis of the principal components, DAPC, programmed in the adgenet package in R. This method, in turn, can be used to determine the structure of the population and to speak of clusters, groups that are genetically distinct from each other and to which individuals can be unambiguously assigned. In this population of honey bees in the Czech Republic, five groups, five clusters, were estimated to be different.