Individual identification and population genetic study of red deer using autosomal tetranucleotide microsatellite markers

THESIS OF PHD DISSERTATION

Zoltán Szabolcsi

DOKTORATE SCHOOL IN BIOLOGY
Head of the school: Dr. Erdei Anna DSc
Member of Hungarian Academy of Sciences

CLASSICAL AND MOLECULAR GENETICS PHD PROGRAM
Head of the program: Dr. Orosz László DSc
Member of Hungarian Academy of Sciences
Supervisor: Dr. Orosz László DSc
Member of Hungarian Academy of Sciences

Budapest
2013
1. INTRODUCTION

The growth of the human population has led to the increase of its negative impact on the environment endangering our natural resources and hence our existence. By solving wildlife crimes, wildlife forensic genetics helps to protect the natural resources and supports efforts for their longtime protection. A basic task of wildlife forensic genetics is to identify the species, population or individual origin of a biological trace serving as evidence during a forensic procedure (Ogden et al., 2009). In the past two decades, the expansion of molecular genetic technologies has induced a dramatic growth in the development of forensic genetic markers. The introduction of the first human multiplex STR kit in 1994 (Kimpton et al., 1994) gave rise to the development of multiplex PCR systems to be used for identification of individuals or geographic origin of wild animal and plant species. Today, there are several multiple STR systems for animal as well as plant species that can be used in forensic investigations. In fact, several poaching cases have been solved using multiple STR systems designed for wild mammal species (Caniglia et al., 2010; Barbanera et al., 2012).

2. MAIN OBJECTIVES

Our objective was to develop, first in Hungary, a multiplex PCR systems consisting of tetranucleotide microsatellites which can serve as DNA-based evidence in forensic procedures (eg. in cases of poaching) (Szabolcsi et al.). In accordance with this, the STRs must be polymorphic so that individuals can be differentiated with the statistical power required for forensic investigations. Allele frequency distribution may vary among populations. As a consequence, the polymorphism of microsatellites within the population had to be evaluated. For this purpose, a population reference databases was created to estimate the identification power of DNA evidence consisting of tetrameric STRs. In population genetic studies of the Hungarian red deer, Hartl detected significant genetic differences between populations of various habitats (Hartl et al., 1990). Since genetic inhomogeneity of the population database may alter the statistical power of the DNA evidence, the genetic structure of our databases had to be examined.

Population genetic studies using STR markers provides a lot of information for wildlife biologists as well as game managers. We wanted to use these STR multiplexes for genetic analysis of red deer stocks in order to identify local populations and define their geographic boundaries. Such information can also aid restructuring the game management units with the
aim to make red deer management principles more sound in ecological terms. This results can contribute to the development of the regional game management system and better allocation of harvest quotas (Csányi et al., 2010).

With these considerations in mind, our main objectives were as follows:

I. Sample collection from different populations. The extension of sampled population to habitats which have not been studied previously.

II. Use of the "zoo cloning" method developed by our team for red deer ortholog determination and DNA sequence analysis of 22 tetranucleotide microsatellite loci isolated from mule deer and vapiti libraries.

III. Polymorphism-based selection of STR markers by the PhastSystem technology.

IV. Construction of multiplex PCR systems capable of parallel amplifying several highly polymorphic STR loci.

V. Development of genotyping by the multiplex PCR systems. Separation of the PCR fragments labelled with DS-33 fluorescent dyes using ABI 3130 capillary gel-electrophoresis instrument. Calculation of fragment sizes by the GeneScan 500 LIZ size standard and GeneMapper v3.2 software.

VI. Evaluation of the polymorphism of STR loci and sequence characterisation of the detected alleles. Use of international nomenclature recommendations in allele type definitions.

VII. For the estimation of identification power of the STR multiplexes, random association of alleles within (HWE) and between loci (LE) have to be tested.

VIII. The identification power of the STR profile can be influenced by the inbreeding and genetic differentiation. These effects have to be measured.
IX. Evaluation of population samples to be used as forensic databases. Estimation of the overall identification power of STR loci, with the correction of genetic variances within and between population databases.

X. Evaluation of the possibility of population origin determination using genetic profiles.

3. MATERIALS AND METHODS

Standard protocols were used to extract DNA from population samples. DNA concentration was measured by spectrophotometry method using NanoDrop instrument. "Zoo-cloning" was used to identify the red deer orthologues of tetrancleotide microsatellites derived from mule deer (*Odocoileus hemionus*) and American vapiti (*Cervus canadensis*). During polymorphism selection using PhastSystem, microsatellites consisting of less than three alleles were excluded from further development. Primers were labelled with DS-33 (Applied Biosystems) fluorescent dyes. The amplified PCR-products were separated and detected on the automatic ABI 3130 capillary-electrophoresis system. Allele counts and frequencies were determined by the Gene pop v4.0 software which was also used to calculate expected and observed heterozygosities at each loci. For testing Hardy-Weinberg equilibrium (HWE) modified Fisher’s exact test was calculated using Gene pop v4.0 and Arlequin v3.11 softwares (*Guo and Thompson, 1992; Excoffier et al., 2006; Rousset, 2008*). Marker location were examined with statistical (LD test) method and othologue mapping based on sequence similarity between red deer and cattle. Arlequin v3.11 was also used to test the linkage disequilibrium (LD) between all pairs of loci using likelihood ratio test (*Excoffier et al., 2006*). The significance level of the multiple tests were corrected using Bonferroni method (*Reiczigel, 2010*).

The genetic structure of the populations were examined with F-statistic parameters estimated by Arlequin software using analysis of molecular variance (AMOVA). The most frequent profile „uniqueness” and the average probability of match/identity (pM/PI_ave) were used to estimate the identification power of the multiplex systems. pM and PI_ave were corrected with significant F-statistic parameters and calculated by using API-CALC software (*NRC, 1996; Ayres and Overall, 2004*). The genetic inhomogeneity of the populations was examined with the probability of population origin of an individual using likelihood based method (LBM).
We estimated the 95 % confidence interval of the multilocus genetic profiles using Chakraborty method (Chakraborty et al., 1993).

4. RESULTS AND DISCUSSION

4.1 Red deer tetranucleotide microsatellite loci

Using the zoo-cloning technique, we managed to characterize the full sequence of 14 red deer orthologs of the 22 tetranucleotide STR markers derived from mule deer (Odocoileus hemionus) and the American vapiti (Cervus canadensis) (Jones et al., 2002; Meredith et al., 2004). During sequence comparisons of the orthologue STR loci we found flanking region mutations for all markers except the T123 locus. Examination of the primer binding sites of the flanking regions revealed sequence differences in four loci. These mutations could have led to the genotyping errors without flanking region sequence analysis. The highest number of mutations compared to the size of the flanking regions was detected in the mule deer derived C01, C143, C229 and the vapiti originated T26 STR markers.

Based on phylogenetic studies (Ludt et al., 2004; Pitra et al., 2004) the high number of sequence differences between red deer and mule deer C01, C143 and C229 orthologs is not surprising as the divergence time is longer than the split time of vapiti and red deer, which fact explains the higher rate of accumulated mutations. The high number of sequence differences in the T26 flanking regions and the high degree of allelic polymorphism of the T26 marker suggests that T26 is the ancient one among the examined microsatellite genes. Earlier formation of this microsatellite gene is also supported by sequence comparison results. This marker showed the lowest sequence similarity (79.37%) to cattle orthologue.

4.2 Polymorphic red deer tetranucleotide microsatellite markers

We collected 100 red deer samples from the south-western counties (DNY – Zala, Somogy, Baranya and Bács-Kiskun (Gemenc)) and the north-eastern regions (EK – Börzsöny, Gödöllő Hills, Cserhát, Bükk and Mátra) of Hungary, which have not been studied previously. After polymorphism selection, 10 polymorphic markers of the 14 red deer orthologs proved to be suitable for the development of a marker set that could be used for individual identification. Allele frequency distribution of populations was estimated for each locus. 135 alleles were detected, 28 of these proved to be intermediate sized alleles. 40 alleles and 10 intermediate
alleles were sequenced. Sequence analysis of microsatellite alleles facilitated the use of the international recommendations of allelic nomenclature. Furthermore, the results of allele sequencing confirmed the existence, and accurate separation of the intermediate alleles differing by one, two or three bases units detected at the T26, T501, T507 and T156 loci. C229 had the lowest number (5) of alleles while T26 and T501, with 21 and 23 alleles respectively, showed the highest degree of polymorphism. Jones et al. (2002) reported that T501 proved to be the most polymorphic marker with a total of 11 alleles identified among 43 vapioti individuals. The high number of T501 alleles are due to the homopolymer A10 stretch in its repeat region. Based on the sequence motif structures of the repeat regions, T108, T156, T172 and T507 were classified as simple, C229, T123 and T193 as compound, while C01, T26 and T501 as complex microsatellite markers. On the basis of stable and polymorphic repeat units and the number of alleles, T26 and T501 are considered as hyperpolymorphic STR loci.

4.3 Genotyping with DeerPlex I-II markers

First in Hungary, 10 polymorphic tetranucleotide STR markers were incorporated in two multiple PCR systems each containing 5 loci (5-Plex), called DeerPlex I and II. The relative quantity of non-allelic PCR products (artifacts) was reduced by optimising the dNTP, Mg²⁺, primer concentrations and annealing the temperature of multiple PCR reactions. To minimise the artifacts near the allelic peaks of the T172 and T501 loci (double peak, shoulder peak), primer modification was applied, the so called "pig-tailing" method. The DS-33 fluorophore dyes were choosen for genotyping. ABI PRISM 3130 capillary electrophoresis system GeneScan™ 500 LIZ (Applied Biosystems) internal size standard and the GeneMapper® ID ver3.2 software were used for PCR fragment size detection and calculation. Allelic fragments of the population samples were grouped into ± 0.5 nt window within a standard deviation of 0.25 nt. The so called manual method of genotyping using DeerPlex I-II consists of two steps. In the first step, the amplified fragments of DeerPlex I-II were measured using the ABI PRISM 3130 instrument and the GeneMapper® ID ver3.2 software for analysis. The allele type depends on which allelic fragment range with ± 0.5 nt size and 0.25 nt SD grouped into and the relative allelic fragment size differences compared to the reference genotype of a female red deer.

4.4 Species specificity of DeerPlex I-II systems
To test the species specificity of the two 5-plex STR sets cross-species amplifications were performed in humans and, furthermore, the most common even-toed ungulates. No allelic PCR product was detected on human, boar, cattle, mouflon and roe deer samples. 9 loci genotype were characterized from fallow deer sample. No PCR products were found in the allele range of the T26 marker. As no T26 marker drop-outs were observed in the genetic profiles of red deer population samples, it is highly unlikely that fallow deer individuals mixed in. On the strength of study outcomes it can be concluded that DeerPlex I-II produced only red deer specific amplicons and population samples were not contaminated either by wild ungulates (roe deer, mouflon, wild boar, fallow deer) or by professional hunters taking part in the sampling processes. Phylogenetic studies suggests that the sika deer belonging to a separate phylogenetic group (Eastern clade), which also includes the American vapiti, can have very similar flanking region sequences; in view of this, a specificity test including sika deer may be recommended. When examining animal samples, admixtures of individuals or even close related species in a sample may cause problems in genotyping. Before individual identification of wild species, it may be worthwhile to analyse the species spectrum of a given sample qualitatively as well as quantitatively using real-time PCR procedure designed for species/subspecies specific in/del mutations.

4.5 Chromosomal localisation of DeerPlex I-II markers

On the basis of Slate’s linkage map (Slate et al., 2002), 4 of our markers had known chromosomal location. The genetic linkage of our STR microsatellites needed to be tested. Two approaches were adopted. As the statistical linkage analysis resulted, no pair of loci has shown LD when tested in all the three possible population samples (NE, SW and WP). Statistical approach to independence studies failed to clarify the marker independence for some of the locus pairs (eg. T108 - C01, C01 - T501).

Using mapping with sequence similarity searches, T507 proved to be the only locus whose location could not be clearly identified. Orthologs of the T26 and C01 markers localised on cattle chromosome 16, which is homologous with the linkage group 14 of the red deer on Slate’s comparative linkage map, suggesting non-independent inheritance. Although the flanking sequence similarity of the cattle T26 sequence is low compared to its red deer orthologue and this loci sequence as a microsatellite gene is also questionable, the potentially linked T26 - C01 pair was taken into account in further tests. This hypothesis can be
confirmed only when the whole genome sequence has become known, but this was not available at the time when the present work was being prepared.

4.6 Genetic diversity and genetic structure of populations

We were the first in Hungary to study the genetic structure of wild red deer stocks using tetrameric microsatellite markers. Maximum values of $H$, $PIC$, $PD$ were detected in the NE population, northeast population can be considered as genetically more variable and the more favourable as far as the efficiency of forensic individual identification is concerned. The results of F-statistics showed a significant genetic difference ($F_{ST}$) between the two subpopulations but this was not high compared to other European red deer populations (Kuehn et al., 2003, 2004). Global AMOVA on the total population produced an average inbreeding value of 0.0373 (average $F_{IS}$), the Wahlund effect was estimated at 0.034 ($F_{ST}$). Compared to $F_{IS}$ and $F_{ST}$ values of human populations, both parameters appear to be high. Correction of the profile frequency (and $LR$) with the common ancestry coefficient (theta) recommended in human and non-human forensic genetics is justified (Linacre et al., 2011). Theta factor is equal to $F_{IT}$ when using the red deer total (T) population database. For the calculation of profile frequency (and of $LR$ as a consequence) using the total population database, the $F_{IS}$ parameter averaged for all loci and the single loci $F_{ST}$ values were used. Genetic inhomogeneity of the subpopulations was analysed by population origin evaluation of an individual using $LBM$ method. The outcome of $LR_{LBM}$ calculations, 17% of the individuals did not belong to the reference population, suggesting further structuring in NE and SW populations. Considering the confidence intervals of profile frequencies, estimation of the population origin is not likely to provide a significant value because of genetic inhomogeneity, i.e. the relatively high percentage of individuals derived from the non-reference populations (17% of northeast and southwest individuals). The identification of subpopulations (demes) would require a more extensive examination of the Hungarian stocks. For the calculation of the $LR$ values of DeerPlex I-II profiles, the allele frequency database containing the data of 100 individuals is recommended. The use of the north-eastern or south-western populations as database, due to their presumable heterogeneity and small size, may provide a biased estimate of $LR$.

4.7 Identification power of the DeerPlex I-II systems
Since our aim was the genetic identification of red deer individuals, the identification power of our two 5-Plex PCR systems consisting of polymorphic microsatellites had to be tested. This was carried out by two-population statistical analysis. Based on the results of linkage tests, the statistical power of DeerPlex I-II for red deer individual differentiation was estimated for 10 red deer STR markers and, with the C01 locus taken out of the multiplex system, 9 markers.

Average matching/identicality probability \( (pM/PI_{ave}) \) of DeerPlex I-II genotypes in the total population was estimated from the allele frequency data of the 100 individuals. Using the API-CALC v1.0 software, the \( pM/PI_{ave} \) value was corrected by the significant average inbreeding parameter within the subpopulation \( (F_{IS}) \) and the subdivision parameter in the total population \( (F_{ST}) \). On the basis of the calculations, the probability that two red deer individuals randomly chosen from the population have the same 10/9 loci genetic profile was \( 2.6236 \times 10^{-15}/1.0056 \times 10^{-12} \). These are very low values, which can also reflect the probability of false identification – if the genotyping error can be excluded.

Another way of quantifying the identification power of the DeerPlex I-II systems is the estimation of the uniqueness of the theoretically "most frequent" genetic profile. Depending on the size of the population, the uniqueness parameter is the probability that a genetic profile is unique. If this genetic profile is the theoretically "most frequent", this probability will give the minimum uniqueness of the DeerPlex I-II in a particular population. In calculating the theoretically "most frequent" profile, inbreeding \( (F_{IS}) \) and substructure \( (F_{ST}) \) parameters were also taken into account. The calculations resulted a value at \( 1 - 6.73 \times 10^{-5} \) for the 10-locus uniqueness probability and \( 1 - 7.93 \times 10^{-4} \) for the 9-locus uniqueness probability. The calculations were based on the estimated 2 million individuals of the European red deer and sika deer population.

Both calculations confirm that, both sets of polymorphic microsatellites (containing 10 or 9) can be safely used for genetic identification and differentiation of red deer individuals in wild populations. Their identification power is similar to those provided by the Identifiler multiplex system uniqueness value and combined PD value calculated for the Caucasian population in the United States (Butler, 2011).

5. SUMMARY

Being a large herbivorous species, the red deer (Cervus elaphus hippelaphus) is an important member of the Hungarian fauna from the point of view of ecology as well as of game
management. The number of individuals is estimated to be more than 100,000. Approximately 40,000 of them are legally hunted each year. Illegal loss by poaching may be as high as 10,000. One of our objectives was to develop a multiple PCR system consisting of tetranucleotide (tetramer) microsatellites (STR) suitable for linking the offender directly with the crime scene without the necessity of catching the perpetrator in the act. To facilitate the linking of a biological trace with an individual, microsatellite loci must be sufficiently polymorphic.

To assess the polymorphisms of loci, a total of 100 samples were collected from the southwestern counties (DNY – Zala, Somogy, Baranya and Bács-Kiskun (Gemenc)) and the northeastern regions (EK – Bőrzsöny, Gödöllő Hills, Cserhát, Bükk and Mátra) of Hungary. The second objective of our investigation was gaining a more extensive knowledge of the genetical structure of the Hungarian red deer stock as a metapopulation.

Using the so called "zoo cloning" method, 14 locus sequences were determined of the 22 tetramer STR orthologs isolated from the American vapiti (Cervus canadensis) and the mule deer (Odocoileus hemionus) available in the literature. Four STR markers were selected by polymorphism screening using the PhastSystem with high resolution PAGE electrophoresis. Of the 10 polymorphic microsatellite loci, two multiple PCR systems each containing 5 STR markers (DeerPlex I-II) were constructed. DS-33 (Applied Biosystems) standard matrix fluorophores were used for labelling. Genotyping was performed using the ABI 3130 capillary gel-electrophoresis instrument and the Genemapper® ID 3.2 software (Applied Biosystems).

To estimate the identification power of the markers, the degree of loci polymorphism was characterised by Nei’s gene diversity, polymorphism information content (PIC) and the power of discrimination (PD). With the exception of the C229 locus, individual values were high. Random association of alleles within the loci was evaluated by the Hardy-Weinberg equilibrium (HWE) test using the Arlequin v3.11 and Genepop v4.0 softwares. At a significance level of 5%, HWE was rejectable for the T501, C01 and T108 markers. Independence tests of the loci did not exclude the genetic linkage of C01 and T26 either by statistics (LD test) or by sequence similarity based mapping, therefore the 9 loci not containing the C01 marker were used for forensic calculations.

To test the individual identification potential of DeerPlex I-II, the genetic structure of the populations needed to be investigated. The Arlequin software was used to perform a molecular variance analysis (AMOVA), which detected a significant genetic differentiation between the two subpopulations ($F_{ST} = 0.034$); mean inbreeding parameter within the
A subpopulation was found to be $F_{IS} = 0.037$. The identification power of DeerPlex I-II was expressed in average probability of identity ($PI_{ave}$) and average matching probability ($pM$) corrected by the $F_{IS}$ and $F_{ST}$ parameters and in the „uniqueness” of the theoretically „most frequent” genetic profile. The probability, that a 9-loci DeerPlex I-II profile is unique not higher than $1 - 7.93 \times 10^{-4}$ and the probability that two randomly chosen individuals have identical 9-locus genotypes was found to be $1.0056 \times 10^{-12}$. The results confirmed a very high identification power, similar to that of the Identifiler kit calculated for the Caucasian population of the United States and used for individual identification.

A good example of the forensic application of the marker set is a red deer poaching case. Biological traces collected from the crime scene and the red deer stag trophy confiscated on the suspect’s estate were successfully originated from one individual using DeerPlex I-II profiles.

The DeerPlex I-II systems were used to study the genetic structure of the subpopulations estimating the population origin of individuals. „Assignment test” was used to estimate the reference population of an individual. As 17% of individuals in both populations were classified as not belonging to the reference population, the populations are considered as heterogeneous and hence cannot be used for statistical interpretation of DNA-based evidence in forensic procedures. For this purpose, the total database is recommended to use for forensic calculations. High proportion of individuals (44%) showed overlapping confidence interval of profile frequency values calculated from the two populations. For this reason, determination the population origin of an individual using assignment test would be biased and therefore are not recommended unless the databases are extended in the future.

REFERENCES


