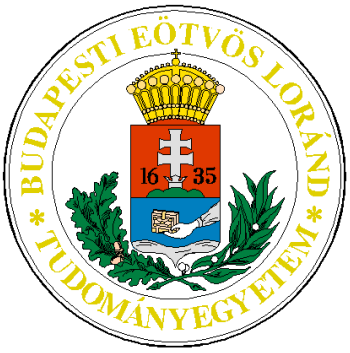

The Genetic Analysis of the Hungarian Population Based on Gonosomal Markers

THESES OF THE PHD DISSERTATION

Written by:

Andrea Vágó-Zalán

Eötvös Loránd University, Faculty of Science, Biology PhD School

Head of the PhD School: Prof. Dr. Anna Erdei

Classical and Molacular Genetics PhD Program

Head of the PhD Program: Prof. Dr. László Orosz

Supervisor: Dr. Horolma Pamjav

2012

1. INTRODUCTION

The vast majority (more than 99,7 %) of the DNA substance is identical in every people, only a fraction of it (approx. 0,3 % or 10 million nucleotides) is different between people, this makes it possible to distinguish each individual with genetic methods. The attributes of one locus itself is not characteristic only for one human person. Therefore alleles determined on proper number of polymorphic loci can be defined as an individual genetic profile characteristic for only one person (genetic identity number). For the use of loci in practice simple, fast and routine typing of alleles is necessary besides the proper high polymorhism.

The first forensic DNA test was carried out in 1985. Nowadays the chriminalistic DNA investigations and paternity tests help effectively the work of criminal investigation and justice. The questions are always raised regarding individuals (for eg. do the two investigated samples originate from the same person, is the alleged father the biological father of the child), but the answers will always be dependant on wich population the questioned persons belong to because allele frequencies can be different between populations.

The investigation of X-chromosomal loci is effective in deficiency paternity cases (the alleged father is dead or disappeared), if the child in the case is a girl. In such cases the female child, with her mother and the mother of the alleged father or legitimate daughter of the alleged father are investigated.

On the Y-chromosome the STR loci are in the focus of forensic genetic applications. They can be applied in deficiency paternity cases, if the child in the case is a boy and paternal relatives of the alleged father (for. eg. brother, father, grandfather, legitimate son) can be investigated. The use of Y-STR loci in chriminalistic DNA investigations is practical if low quantity of male DNA has to be investigated with high female DNA background or the technical disequilibrium (for. eg. “drop outs”) of the autosomal tests are considerable.

The investigation of Y-SNP loci and the Y-SNP haplogroup composition of populations is not significant of a forensic genetic point of view, because the number of Y-SNP haplogroups is considerable smaller than that of Y-STR haplotypes. Y-chromosomal haplogroups have important role in human migration studies, they make possible to identify the essential differences between populations and population groups, to investigate the history and migration of populations.

2. AIMS OF THE STUDY

One of my aims was to introduce new X-chromosomal STR loci in the Hungarian forensic geneticists to help the solving of deficiency paternity cases.

During my former work I carried out the investigation of 4 X-chromosomal STR loci belonging to different linkage groups (HPRTB, DXS7423, DXS7132 ans DXS8378) in the Hungarian population. However, the investiagtion of these 4 loci together could not ensure the sufficient probability of paternity for relationship testing, which is minimal 99,98 % according to the Hungarian methodological regulations and the satisfactory clarfying of exclusions and mutations. To rule out the above mentioned problems I desired to determine the allelefrequencies of 4 new X-STR loci, the haplotype ferequencies of loci-pairs, the populationstatistical values of the loci and to compare my data with other published data.

Regarding that the Y-chromosomal STR and SNP loci are not only of forensic genetic value but their importance in human population genetics is also considerable I desired to participate with the investiagtion of the Hungarian, different Romani and the Malaysian Indian populations in the trailing of history, migration and admixing of human populations

It was my purpose to examine the Hungarian and 4 different Romani populations for 12 Y-chromosomal STR loci and to integrate these data into the international YHRD database. The incorporation in the database makes my data available for international use in forensic cases or for comparison with other population data, wich can facilitate the clearing the relationship of Hungarian and Romani populations with each other or other populations.

I intended to carry out the study for the Y-chromosomal SNP loci combined with haplotyping for the above mentioned populations additionally with the haplogroup typing of Malaysian Indian and Slovakian Romungro Romani males. I wished to examine the relations of the Romani populations in the Carpathian Basin to the Hungarian, other European Romani and non Romani and Indian populations. I also aimed to identify the paternal lineages of the Romani populations. I also intended to send to the YHRD the haplogroup data belonging to the haplotypes of the investigated populations to help the work of the collegaues in international forensic and research DNA laboratories.

3. MATERIALS AND METHODS

I investigated 384 pesons (219 males, 185 females) from the Hungarian population for the X-chromosomal STR loci. For the Y-chromosomal STR loci I investigated 424 males from 5 populations (230 Hungarians, 107 "mixed" Romanies from Hungary, 29 Vlax Romanies from Tiszavasvári, 19 Romungro Romanies from Taktaköz, 39 Vlax Romanies from Tokaj). I investigated 787 males for the Y-chromosomal SNP loci (the above mentioned and 62 Romungo Romanies from Slovakia, 301 Malaysian Indians). The sample types wre the followings: blood samples with EDTA, buccal swabs and dried bloodspots on FTA cards. The DNA from the samples was isolated according to standard protocols. The determining of the DNA concentration happened with Real-Time PCR technology. The multiplex PCR amplification, fragment length analysis and genotyping of the X-STR (Mentype Argus X-8) and Y-STR (PowerPlex Y) loci took place according to the manufacturer's instructions. For the genotyping of 51 Y-SNP loci TaqMan probes were applied according to the manufacturer's instructions. The sequenceing of the rare and microvariant X-STR alleles happened according to standard protocols. I demonstrated the practical applicability of the Mentype Argus X-8 kit through the example of a deficiency paternity case.

The population genetic attributes of the X-STR loci were calculated according to the standard formulae (allele frequencies, haplotype frequencies, PIC, HET^{exp}, HET^{obs}, PD^{male}, PD^{female}, PM^{male}, PM^{female}, $MEC^{Krüger}$, $MEC^{Kishida}$). The testing of the Hardy-Weinberg equilibrium took place with the Arlequin 2.001 software (Fisher's exact test). For the comparison of allele frequency data of population pairs G-test was used.

I calculated the haplotype and haplogroup frequencies, haplotype and haplogroup diversity (Nei's gene diversity) values for the 12 Y-STR and 51 Y-SNP loci located in the non-recombining region of the Y-chromosome. For the comparison of the investigated and published population data F-statistics and AMOVA was used implemented in Arlequin 2.001 software. I analysed the relations of the results with MDS (Multidimensional Scaling, ViSta 7.9.2.4 software). For the visualisation of the relations between the genetic and geographic ditances of the investigated populations I carried out phylogeographical analysis. To visualise the geneological relations of haplotypes belonging to the same haplogroups and to estimate the divergence time of haplotypes with one common ancestor I used the Network 4.2 software.

4. NEW STATEMENTS/OBSERVATIONS (THESES)

4.1. I carried out the investigation and characterization of 384 persons for the 4 X-STR loci (DXS10074, DXS10101, DXS10134, DXS10135) as first in Hungary improving the allele frequency database created earlier containing only 4 X-STR loci [1]. With the investigation of the 219 males I created a haplotype frequency database for 4 closely linked pairs of loci [2]. In 2011 the database was improved with 4 more X-STR loci, now consists of 12 loci [6].

4.1.1. I completed the GeneBank database with the sequence data of the 13 detected (DXS10074: 1 allele, DXS10135: 4 alleles, DXS10101: 3 alleles, DXS10134: 5 alleles) rare and microvariant alleles [2].

4.1.2. There could be no significant difefernce detected between the allele frequencies of males and females, therefore the allele frequencies could be pooled [2].

According to the population genetic analysis the investigated loci are in Hardy-Weinberg equilibrium. There are no considerable differences between the expected (0.8487-0.9419) and observed (0.8485-0.9273) heterosygisity values also supporting the Hardy-Weinberg equilibrium. The polymorphism information content (PIC) values were in the 0.8250-0.9325 range which denotes an acceptable level [2].

The calculated power of discrimination (PD) and probability of match (PM) values are important from the forensic genetic point of view. These values are different for males and females for the 4 X-chromosomal loci– PD^{male} 0.8436-0.9362 and PD^{female} 0.9569-0.9922 – but show an acceptable level for both genders. The combined power of discrimination for the 12 X-STR loci both in males and females is higher than 0.99999999. This means that theoretically one has to investigate more than 100 million males and females to find 2 randomly selected persons in the population whit identical genetic profiles for the 12 X-STR loci [2, 6].

The $MEC^{Kishida}$ (0,8250-0,9325 for the investigated loci) and $MEC^{Krüger}$ (0.6888-0.8715 for the investigated loci) values fall also in the acceptable range, the combined values for the 12 loci ($MEC^{Kishida}$ >0.99999999; $MEC^{Krüger}$ 0.99999876) meet the requirements for the use in forensic genetic casework [2, 6].

4.1.3. Comparisons of the allele frequency data of the Hungarian and non-European populations show significant differences for all the 4 investigated loci between the Hungarian and non-European population data. The comparisons of the Hungarian with other European populations bring various differences for each locus, only the Polish population does not show significant difference from the Hungarian for all the 4 loci.

4.1.4. The aim of investigating X-chromosomal STR loci was the forensic genetic application, which I demonstrated on a deficiency paternity case. The alleged father legitimated 2 daughters in his life but died before the birth of the third girl. Forensic genetic analysis of 4 persons (mother, 2 legitimate daughters and the questioned daughter) on 8 X-STR loci showed no excluding allele combination. The probability of paternity based on the haplotype frequencies turned out to be 99.999728% ("practically proven" verbal category), wich supports the hypothesis that the two legitimated and the questioned daughter originate from the same biological father.

4.2. I carried out the survey of 1 Hungarian and 4 Romani population groups for 12 Y-STR loci, the data were included in the international YHRD database [3, 4, 5]

4.2.1. My initial presumptions were the followings: the Hungarian population is genetically closer to the neighbouring (possible admíxture) and the Finno-Ugric speaking populations, while the geographically further located and non Finno-Ugric speaking populations fall genetically further from the Hungarian one. From the 23 populations included in the comparison the other published Hungarian, the neighbouring (Slovenian, Romanian, Yugoslavian and Bulgarian) and the Norwegian populatuins were genetically closest tothe investigated Hungarian population. The low genetic distances from the neighbouring populations fitted my presumptions; possible cause for it is the admixture of the Hungarian and neighbouring population in the past centuries [3].

The Finnish, Spanish, Belgian and Estonian population data were genetically furthest from the Hungarian. In the case of the Spanish and Belgian populations this can be due to the large geographical distance. In the case of the Finnish and Estonian populations – both belonging to the Finno-Ugric language family together with the Hungarian population – the results contradict the presumptions, however this observation meets other data published earlier [3].

4.2.2. The clusters on the unrooted phylogenetic tree constructed from the genetic distances mirror the geographic and historic relations of the included populations. The Eastern and Western European population data fall into separate clusters. The investigated Hungarian and the neighbouring populations (no linguistic relation) belong to the same cluster, which can be explained with admixture. Based on linguistic relations it could be supposed that the Hungarian population genetically clusters together with the other Finno-Ugric speaking populations. In contradiction the phylogenetic tree shows that the Hungarian and the other Finno-Ugric speaking populations belong to separate clusters. This could be expected knowing the genetic distances, but still contradicts the linguistic studies [3].

4.3. I carried out the population survey of 1 Hungarian, 5 Romani and 1 Malaysian Indian population for 51 Y-SNP loci as first in Hungary contributing also to the research of genetic ancestry. The data were integrated into the international YHRD database [3, 4, 5]

The frequencies of the Y-chromosomal haplogroups show very characteristic geographical distribution. The haplogroups found in the investigated Romani population groups can be linked to 4 main geographic areas: Indian (H1a-M82), Near Eastern/Western Asian (J2a2-M67, J2*-M172, E1b1b1a-M78), European (I1-M253, I2a-P37.2) and Central Asian/Western Eurasian (R1a1-M198, R1b1-P25) [4, 5]

4.3.1. In the comparison of the 7 population data (MDS plot generated from pairwise Fst genetic distances calculated from haplogroup frequencies) the Romani population groups are obviously separated from the Hungarian and Malaysian Indian populations. 3 of the 5 Romani population data cluster together, the genetic distance between them is also low, wich supports that the Slovakian and Tokaj Romanies diverged from each other only lately. The Hungarian "mixed" Romani population data is heterogenous this can explain their low genetic distances to and the clustering with the Slovakian and Tokaj Romani data. The Taktaköz Romungro data is positioned approximately halfway between the cluster of the 3 Romani populations and the Hungarian reference population. This finding supports the theory that the Romungro Romanies arrived to the Carpathian Basin earlier than the Vlax Romanies and had more time to admix with the Hungarian population. The Vlax Romanies form Tiszavasvári are placed distinct from the other 6 populations which can be due to genetic drift or the low haplotype diversity of the population [5].

In a broader comparison of 41 population data (pairwise Fst genetic distances) every Romani population data can be linked with the Balkanian and Anatolian populations. This observation fits the possibility that these areas constituted important geographical link between the Near East, Asia and Europe during the migration of the Romanies. The Hungarian reference population samples show closer genetic relation to the Indian higher castes than to the Balkan populations, which can be due to high frequency of R1a1-M198 in the Hungarians and Indian higher castes. This supports the theory that the R1a1-M198 haplogroup can be traced back to common Central Asian ancestors [5].

The Rst distances calculated from the Y-chromosomal STR haplotypes of 12 Romani and the Malaysian Indian population samples were plotted against the geographical distances of the population pairs in the framework of a phylogeographical analysis. The genetic closeness of populations can be the result of common origin or recent admixture due to the small

geographic distance. Based on the universal relationship geographically close population-pairs should show small genetic distances while geographically distant ones should show high genetic distances. The results of the phylogeographical analysis contradicts the before mentioned relationship, because some Romani population groups are genetically closer to the Malaysian Indian population (geographically distant) than to other Romani populations. This means that in spite of the large geographical distance a part of the Romani populations show clocer genetic relationship to the Malaysian Indian population wich does not contradict the Indian origin of the Romani populations. Some Romani population data pairs are genetically far from each other despite of geographical closeness (for eg. Romungros from Taktaköz – Macedonian Romanies, Romungros from Taktaköz – Lithuanian Romanies, Baranya Romanies – Macedonian Romanies) which can be due to genetic drift and the isolation of the Romani populations [5].

4.3.2. In the MJ network of H1a-M82 Y-chromosomes can be found one common Y-STR haplotype which is present in all investigated Romani populations containing more than half of the H1a-M82 Y-chromosomes. This cluster represents a group of closely-related Y-chromosomes, in this group each haplotype is the descendant of the same central Haplotype (common ancestor from India) independently from its recent geographical position. The H1a-M82 haplogroup can also be found in the Hungarian reference population with low frequency (4.8 %), this is probably due to the fact that the sample donors were not slected for ethnical origin and the ratio of the Romani minority reaches 6-8% in Hungary [5].

The age of the Malaysian Indian H1a-M82 chromosomes based on the STR variance is 8707 ± 1760 years, wich fits the earlier published predictions for the age of H1a-M82 haplogroup. The TMRCA of H1a-M82 haplogroup restricted to the Romani samples (968 ± 336 years) does approximately fit the predicted migration time of Romanies from India to Europe [5].

4.3.3. The J2a2-M67 lineage could be found with relative high frequency in all investigated and the Iberian Romani populations, with lower frequency in the Hungarian population and it is absent from the Malaysian Indian population. This raises the possibility that the Romanies could admix with other non-Indian populations on their route from India to the Carpathian Basin. Taking the fact that the J2a2-M67 haplogroup is also present in the Iberian Romanies in consideration it can be assumed that the J2a2-M67 Y-chromosomes – at least in part - were incorporated into the Romani gene pool before they dispersed from the Balkans in the 15th century [6].

The E1b1b1a-M78 haplogroup is present in all investigated Romani, the Iberian Romani and the Hungarian reference population samples but is absent in the Malaysian – and almost all –

Indian populations. The most of the Romani J2a2-M67 haplotypes compose two clusters in the Mj network. These clusters differ only in one molecular step from each other which indicates that they can be closely related. The results are in accordance with earlier studies which state that the E1b1b1a-M78 Y-chromosomes were incorporated into the Roamni gene pool on their migration route probably in the Balkans. The haplogroup can be found with high frequency in this area which is in agreement with my results [5].

4.3.4. The I1-M253 subhaplogroup of the ancestral European haplogroup I-M170 can be found with relative high frequency in all 6 compared Romani population groups. 61.8% of the men belonging to this haplogroup share a common, central haplotype, which denotes that these Y-chromosomes have a common origin. Taking in consideration that I-M170 and its subhaplogroups are practically absent outside of Europe, the closely related Y-chromosomes from the I1-M253 haplogroup could be incorporated into the Romani gene pool only in Europe before their dispersion within the continent - this means the time of their abstain in the Balkans – despite the fact that I1-M253 has low frequencies in the Balkans in non-Romani populations. This finding is probably the result of genetic drift [5].

The I2a-P37.2 Y-chromosomes are dispersed in the Romani MJ network, the haplogroup can be found only in 3 of the investigated Romani populations. This suggests that these Y-chromosomes could incorporate into the Romani gene pool possibly due to admixture with the host populations in different places and times. It is suppurted also by the fact that the Hungarian I2a-P37.2 Y-chromosomes create a far more diverse network compared to the Romani population groups [5].

4.3.5. The R1a1-M198 haplogroup is present in all Romani population groups excluding the Vlax Romanies from Tiszavasvári. The haplogroup is also absent in the Iberian and Balkanian Romani population groups. The individual haplotypes are dispersed in the MJ network, there can no central R1a1-M198 haplotype be found. If the R1a1-M198 haplotypes from the Hugarian and Malaysian Indian reference populations are also be included in the MJ network it can be seen that the diversity of the haplogroup is in these populations higher than in the Romani samples. Because the R1a1-M198 haplogroup is not present in all Romani population samples, it can be supposed that the haplogroup incorporated in the gene pool of the investigated Romani populations with admixture of populations resident in the Carpathian Basin later than their fragmentation from the Balkans. The analysis of the network also supports this idea, however it is possible that the R1a1-M198 haplogroup incorporated in the Romani gene pool already in India and was lost from some of the Romani populations as a result of genetic drift [5].

The haplogroup R1b1-P25 can be found in the Hungarian and all investigated Romani population samples but is absent from the Malaysian Indian samples. The MJ network in the investigated Romani populations is very diverse; additionally there is only small superimposition in the Hungarian and Romani Y-chromosomes. This supports the possibility that these Y-chromosomes can not be traced back to one common ancestor. Regarding that the R1b1-P25 haplogroup is abundant mainly in Western Europe and is almost absent in India it can be supposed that the R1b1-P25 Y-chromosomes were integrated in the Romani gene pool through population admixtures at different places and times.[5].

4.3.6. In the investigated Romani sample groups could be found other haplogroups which are not characteristic in Europe with low frequencies (C3-M217, N1c-Tat, R2-M124 és P*-M45). It can be supposed that these haplogroups were incorporated into the gene pool of the Romanies before their arrival to Europe. In the Malaysian Indian population sample the F*-M89, L-M11 and R2-M124 haplogroups reach high frequency, these haplogroups are characteristic for South-Asia. Haplogroup F (and its parahaplogroup F*-M89) in India have common history with the H, R2 and M124 haplogroups. Haplogroup R2-M124 can also be found among the Romanies, this also supports their Indian origin [5].

5. CONCLUSIONS, SUMMARY

It is impossible to apply the X-chromosomal STR loci in forensic investigations without establishing allele frequency databases and calculating the population statistical values. I increased the former Hungarian X-STR database with the data of 4 loci (DXS10079, DXS10101, DXS10134, DXS10135) to help the effective solution of deficiency paternity cases. I dtermined the allele and haplotype frequencies for the tightly linked locus duos in the Hungarian population. Then I calculated the population statistical values and compared the allele frequency data of the Hungarian population with the published allele frequency data of the other populations. The 4 loci investigated in the present study and further 8 loci (12 X-STR loci together) are available to solve relationships and forensic cases in Hungary.

I compared the Y-SNP data of the investigated Hungarian male population with other European population data in order to determine the genetic relationships between the Hungarian and neigbouring or other Finno-Ugric speaking populations. According to the genetic distances and the phylogenetic tree the neighbouring (Slovenian, Romanian, Yugoslavian and Bulgarian) and the Norwegian populations are genetically closest to the Hungarian population. This finding is in consistence wtih the fact that the Hungarian

population could have been admixed with the neighbouring ones in the past centuries. The populations belonging to the Finno-Ugric language family (Finnish, Saami, Estonian and Mari) were genetically the furthest from the Hungarian population data. The phylogenetic tree also reflected this result which contradicts the linguistic studies.

I compared the Y-STR and Y-SNP data of the investigated Romani populations to other Indian (possible place of origin) and formerly investigated European Romani and non-Romani populations as well. The haplogroups detected in the Romani populations can be classified into 4 main geographic areas: Indian (H1a-M82), Near Eastern/West-Asian (J2a2-M67, J2*-M172 and E1b1b1a-M78) European (I1-M253 and I2a-P37.2) and Central-Asian/West-Eurasian (R1a1-M198 and R1b1-P25). The presence of these haplogroups in the Romani population samples can be interpreted in the mirror of their migration route from the Indian subcontinent to the Carpathian Basin and their admixture with other populations. In conclusion, it can be stated that these observations shed light on a coherent version of the genetic history of the Roma people which is consistent with written sources.

6. RELATED PUBLICATIONS

[1] **Zalán A**, Völgyi A, Jung M, Peterman O, Pamjav H (2007) Hungarian population data of four X-linked markers: DXS8378, DXS7132, HPRTB, and DXS7423. Int J Legal Med 121:74-77

[2] **Zalán A**, Völgyi A, Brabetz W, Schleinitz D, Pamjav H (2008) Hungarian population data of eight X-linked markers in four linkage groups. Forensic Sci Int 175(1):73-78

[3] Völgyi A, **Zalán A**, Szvetnik E, Pamjav H (2009) Hungarian population data for 11 Y-STR and 49 Y-SNP markers. Forensic Sci Int Genet 3:27-28

[4] **Zalán A**, Béres J, Pamjav H (2011) Paternal genetic history of the Vlax Roma. Forensic Sci Int Genet 5(2):109-113

[5] Pamjav H, **Zalán A**, Béres J, Nagy M, Chang YM (2011) Genetic structure of the paternal lineage of the Roma people. Am J Phys Anthropol 145(1):21-29

[6] Horváth G, **Zalán A**, Kis Z, Pamzsav H (2012) A genetic study of 12 X-STR loci in the Hungarian population. Forensic Sci Int Genet 6:e46-e47