„BABEŞ-BOLYAI” UNIVERSITY

PhD School
Faculty of Biology and Geology
Department of Genetics

Allelic frequency of STR loci in some populations from Transylvania

Summary of PhD Thesis

PhD student:
József-Szabolcs DEMETER
Scientific adviser:
Prof. Dr. Octavian POPESCU
Cluj-Napoca, 2010

Content

Abstract

2010 marks the 25th anniversary of the first paper describing the genetic identification of human individuals by DNA fingerprint analysis. Since then, DNA analysis has become a major tool in relating disease diagnostics to therapy, forensic identification, taxonomy, phylogenesis, and others. Microsatellites, also called short tandem repeats (STRs) are among the most polymorphic DNA markers suitable to a series of genetic, forensic and medical applications.

The aim of this study. According to the 2002 Romanian census, in Transylvania there is a considerable Hungarian ethnic community (19.6%), mostly in the Szekler counties of Covasna and Harghita. In the last decade, genetic parameters for Szekler and Csángó populations from Harghita have been published, but no population study has been conducted on Szekler communities from Covasna and other (non-Szekler) Hungarians from Transylvania. Given this situation, the present study was conducted to compare Hungarian populations (Csángó, Szekler and non-Szekler) with each other and with Romanian populations, and to contribute with new data to Transylvanian and Romanian STR databases.

Methods. Allele frequencies for 15 STR autosomal loci (D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818 and FGA), forensic and paternity parameters were determined for Covasna's Szekler Hungarians and non-Szekler Transylvanian Hungarians represented in this study by Hungarians from Cluj county.

Results. 424 DNA profiles and allele frequencies were determined for 146 unrelated individuals from Cluj, and 278 individuals from Covasna counties. Sixteen off-ladder alleles involving three loci (D19S433, D2S1338 and FGA) were identified in these two populations. Transylvanian and Romanian autosomal STR Databases were updated.

Population comparison tests resulted in significantly different values for non-Szekler Hungarians *versus* Szekler Hungarians from Harghita (locus D5S818), Csángós, Transylvanians and Wallachians (locus D3S1358). Population data comparison performed on Szeklers in Covasna county shows fewer differences in comparison to populations from Western Romania and Dobruja region, while the greatest differences were found in the comparison with population data of Wallachians.
D21S11, D13S317 and D3S1358, D5S818 loci seem to be the most polymorphic autosomal STR markers for the Romanian regions, when matched to the entire Romanian database and the FBI databases.

Conclusion. Genetic differences between the two neighbouring Szekler populations are higher than expected. Future investigations and further population comparison tests must be performed in order to clarify these populations' genetic characteristics.
Although additional tests are required to explain the given effect of the Transylvanian Hungarian communities upon the Romanian DNA database, this study indicates that ethnical origin should also be taken into consideration when comprehensive DNA databases are created.

Keywords

15 polymorphic autosomal STR loci; Romania; Updated Transylvanian Database; Moldavia; Dobruja; Wallachia; Population Data; Covasna; Harghita; Cluj; AmpFlSTR Identifiler; Genetic diversity.

Preliminary

Between 1999 and 2003, I attended the courses of the Biology Department of the "Babeş-Bolyai" University. In summer 2001, I enrolled in a summer work exchange program dedicated to students. The well known 9/11 (1991) terrorist attacks found me in New York City. At this moment, I still have too many inexpressible memories about that hopeless situation. Smoke, odour, chaos, isolation, loneliness, unpleasant feelings…

On the day of the attacks, thousands of people went tens of miles on foot: some trying to escape from the inferno, others trying to find their relatives. So many people started instinctively a "trail of tears" to the way to massive grave at Ground 0. Highways were full of these people… I hadn't seen and experienced anything like that before. The terrifying experience of that massive human disaster marked me considerably. Many "Whys?" and "How is this possible?" questions were formulated.

For scientists, the disastrous events of 9/11 represent an unprepared homework for an undesirable lesson called today: human victim identification of mass fatality incidents. For me, it is a life experience. Today I know how it is possible to identify any human remains mingled with debris, but most of the questions formulated there still suffer from lack of answer.

This study is not a sentimental recall of my memories, but it is a scientific review for methods and techniques used in human identification, a re-evaluation of genomics of DNA markers, of electrophoresis based DNA Profiling technique, of DNA Projects, DNA databases, of STR based population studies, and so more.

The scientific literature review, the rigorous allelic designation control, and the detailed procedure of statistical data interpretation are just a few of the outputs of this study which may support any DNA profiling group or laboratory. The new population data, the updated Transylvanian database, and the comprehensive Romanian autosomal STR database represent a significant contribution to the Romanian forensic community.

This thesis is a human population genetic study which involves DNA profiling.

Acknowledgements

I have countless reasons to thank my scientific supervisor Prof. Dr. Octavian POPESCU for His support and continuous advice, which were required for my successful achievement of the tasks proposed 5 years earlier.

I would also like to express my appreciation to Dr. György BÓDIS and Dr. Zsombor VAJNA for their support in finding sample donors and collecting the samples.

I would like to thank Dr. Katalin Ferencz BEATRIX, Dr. Beatrice KELEMEN and PhD. Sergiu CHIRA for their permanent advice during the entire technological process, as well as to Dr. Gyöngyi SZÉKELY, Dr. Simona Maria BEL, Dr. Endre JAKAB and Dr. Szilárd BÜCS for having offered lots of valuables professional suggestions.

Júlia GALACZI (teacher of English language) and Beáta GOMBÁR (teacher of Romanian language) performed the stylistic revision of the present thesis. Róbert BODNÁR (librarian of BCU) ensured my access to scientific bibliography. László KORÓDI (teacher of geography) helped me with editable graphics of the Romanian map. Many thanks for their contribution.

I am also grateful to Dr. László FODORPATAKI (my scientific adviser during the academic years) for his continuous motivation.

Last but not least, I have to mention the generosity of Mr. András FARKAS (IT manager of MOL Romania PP) who made it possible for me to finish this research work.

THANK YOU!

Cluj-Napoca,
20th September, 2010

Summary of Thesis

Part I: Review of scientific literature related to STR loci

I.1 Introduction

Technological and scientific innovations have an impact on human life in many ways. Criminal acts, forensic cases, terror attacks or human mass disasters are just a few of the undesirable, unpredictable, and inevitable situations relying on human identification, which requires continuous technological and scientific development.

2010 is an anniversary from several points of view:

- 1 year since the Council of Europe introduced the new European Standard SET (ESS) loci, which contains 12 STR markers (November 2009);
- 5 years since degraded DNA samples and human mass disaster cases forced the Forensic Science Institutes (ENFSI) and the European DNA Profiling Group (EDNAP) to start development and validation of a next generation STR system including miniSTR's (Glasgow, 2005);
- 15th anniversary of the first DNA database (April 1995), which was set around a second generation multiplex (SGM) system containing markers for TH01, VWA, FGA, D8S1179, D18S51, and D21S11 loci;
- 20th anniversary of the first STR markers developed;
- 25th anniversary of the first paper describing the genetic identification of human individuals by DNA fingerprint analysis (Gill et al. 1985).

DNA analysis has become a major tool in relating biological evidence to the persons involved in criminal situations, mostly thanks to Genomics – the science of genetic material – which has the potential to offer daily updatable tools applicable for disease diagnostics and therapy, forensic identification, taxonomy, phylogenesis, and many more. The ability to explore genetic variation of the autosomal and sex chromosomes and mtDNA are just a few outstanding results in genomics, which now provide sufficient genetic information for the high probability individual identification.

Further genome research will upgrade the existing means and methods and surely provide new tools in forensic identification, especially those genetic factors that will help in reconstructing the physical appearance of an individual or infer the geographical origin of a person.

I.2 DNA Polymorphism, Variability, Genetics and Genomics of STR Markers. Genetic polymorphism and variability provides our biochemical individuality and allows us to use DNA information for human identity purposes. This variability is mostly generated by DNA replication and recombination processes. During the replication (multiplication), numerous “mistakes” may occur: point mutations, deletions or translocations. These variations are collectively known as "polymorphisms" (literally, multiple shapes) and they are exhibited in the form of different alleles (various possibilities) at a particular locus *(Figure 1.1).* Two forms of polymorphism are possible at DNA level: sequence polymorphism and length polymorphism *(Figure 1.2).*

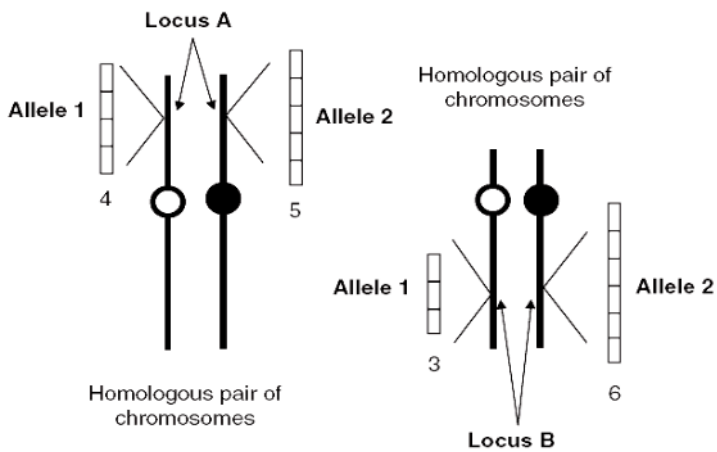

Figure 1.1: Schematic representation of two different loci (A and B) on different pairs of homologous chromosomes. The chromosomes with the open circle centromeres are paternally inherited while the solid centromere chromosomes are maternally inherited. Thus, this individual received the four-repeat allele at locus A and the three-repeat allele at locus B from their father, and the five-repeat allele at locus A and the six-repeat allele at locus B from their mother. (fig. modified according to Butler 2005)

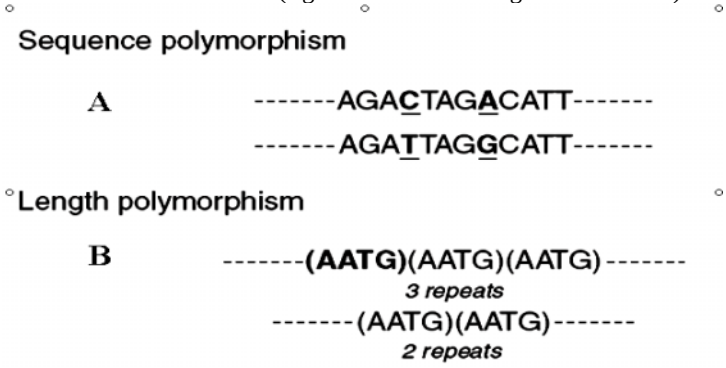

Figure 1.2: Two forms of variation: A = sequence polymorphisms; B = length polymorphisms (fig. modified according to Butler 2005).

Short Tandem Repeats (STRs) also called **Microsatellites or Simple Sequence Repeats (SSRs)** are among the most polymorphic markers and are made up of the tandem repeats of sequences ranging from 2-6 base pairs (Tautz 1989).

Mononucleotide repeats

A C

Dinucleotide repeats

AC AG AT CG

Trinucleotide repeats

AAC AAG AAT ACC ACG ACT AGC AGG ATC CCG

Tetranucleotide repeats

AAAC	AAAG	AAAT	AACC	AACG	AACT	AAGC	AAGG	AAGT	AATC
AATG	AATT	ACAG	ACAT	ACCC	ACCG	ACCT	ACGC	ACGG	ACGT
ACTC	ACTG	AGAT	AGCC	AGCG	AGCT	AGGC	AGGG	ATCC	ATCG
ATGC	CCCG	CCGG							

Figure 1.3: STR markers categorized by the number of the repeating nucleotides; AGAT-motif is mostly used in forensic studies (fig. modified according to Jin L. et al. 1994)

STRs show wide and uniform distribution throughout the genome, PCR feasible short sequence length and high level of relatively stable polymorphism. These varied properties of STR loci decide their suitability to several genetic applications. The number of repeats in STR markers can be highly variable among individuals, which makes these STRs effective in human identification. Among the various types of STR systems, tetranucleotide repeats (*Figure 1.4*) have become more popular than di- or trinucleotides because they are easier to identify with size-based electrophoretic separations.

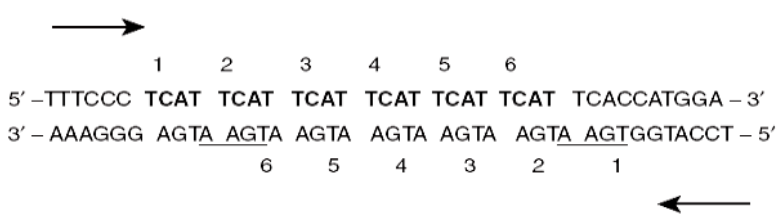

Figure 1.4: Example of DNA sequence in a STR repeat region. (fig. modified according to Butler 2005)

MiniSTRs. The current commercially available multiplex STR kits used in forensic DNA typing can generate amplicons in the size range of 100 bp to 450 bp (Krenke et al. 2002; Holt et al. 2002). When dealing with degraded DNA samples, the best way to recover information is the usage of miniSTRs due to the reduced size of the PCR products (<150 bp) achieved by moving primers as close as possible to the repeat region (Krenke et al. 2002; Wiegand et al. 2001).

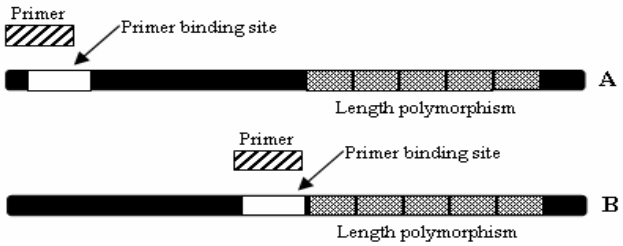

Figure 1.5: STR (A) and miniSTR (B) comparison: The primer binding region of miniSTRs is closer to the repeat region, which results in <150 bp amplicon size.

Unfortunately, at the moment, miniSTRs have an important disadvantage: the closer the primer binding site is to the repeat region, the higher the allele dropout frequency gets.

I.3 Core STR Loci Used in Human Identity Testing. Common sets of short tandem repeat (STR) markers or "core loci" are required for entry of DNA genotype data into national or international databases used for linking serial crimes and offenders. Core loci permits equivalent genetic information to be shared and compared. The most important core STR markers are listed in *Table 1.1*.

Standard set / STR marker	EU (ESS-2009)	US (CODIS)	German	Interpol
FGA	**Required**	**Required**	**Required**	**Required**
TH01	**Required**	**Required**	**Required**	**Required**
VWA	**Required**	**Required**	**Required**	**Required**
D3S1358	**Required**	**Required**	**Required**	**Required**
D8S1179	**Required**	**Required**	**Required**	**Required**
D18S51	**Required**	**Required**	**Required**	**Required**
D21S11	**Required**	**Required**	**Required**	**Required**
Amelogenin	Optional	Optional	**Required**	Optional
D16S539	Optional	**Required**	-	-
TPOX	Optional	**Required**	-	-
D1S1656*	**Required**	-	-	-
D12S391*	**Required**	-	-	-
*D2S441**	**Required**	-	-	-
*D10S1248**	**Required**	-	-	-
*D22S1045**	**Required**	-	-	-
CSF1PO	-	**Required**	-	-
D5S818	-	**Required**	-	-
D7S820	-	**Required**	-	-
D13S317	-	**Required**	-	-
D2S1338	Optional	Optional	-	-
D19S433	Optional	Optional	-	-
SE33	-	-	Required	-
STR markers	**12**	**13**	**8**	**7**

Table 1.1: A summary of forensic STR Core loci; EU= European Standard Set loci 2009, US= United States of America CODIS STR Core Set; STRs marked with "*" symbol represent recently added markers according to EU Council Resolution 2009/C296. Markers in italics are miniSTRs.

I.4 The Short Tandem Repeat DNA Internet database (STRBase) is a website maintained by the U.S. National Institute of Standards and Technology (NIST) and is located at http://www.cstl.nist.gov/biotech/strbase/. The purpose of STRBase has been and

continues to be an attempt to bring together the abundant literature and information of STRs in the field of forensic genetics in a cohesive fashion in order to make current and future work easier. New materials are regularly added to expand the valuable information contained on the STRBase website (Butler 2008).

I.5 The Combined DNA Index (CODIS) and National DNA Index (NDIS). The FBI started CODIS as a pilot project in 1990. The official launch of the nation-wide DNA database was on 13th October 1998. In June 2010, the number of samples in CODIS grew to several hundred thousands: over 8,483,906 offender profiles and 324,318 forensic profiles. At the STR Project meeting on 13–14 November 1997, 13 core STR loci were chosen to be the basis of the CODIS national DNA database (Budowle 1998). The 13 CODIS core loci are listed in *Table 1.1*. All 13 CODIS core loci were tested for random match probability, which is about one in a 94 trillion among unrelated individuals (Chakraborty et al. 1999). The three most polymorphic markers are FGA, D18S51, and D21S11, while TPOX shows the least variation between individuals.

I.6 Mass Fatality Incident (MFI) Databases. DNA typing proved to be a robust, consistent and reliable technique, ideally suited for disaster victim identification (DVI) of mass fatality incidents (MFIs). Today laboratories accessing these databases can support DVI in several MFI cases: accidental catastrophes, natural occurrences and intentional acts, such as terrorist attacks, wars, or political crises (Leclair 2004; Bradford 2010; Budowle et al. 2005).

I.7 Beginnings of STR Marker Related Studies on Romanian Populations. In the last decades, 8 population studies have been performed in Romania, two of them with non-CODIS STR core set. The first population data were obtained from a population sample of 105–122 unrelated individuals born in Transylvania and Banat (Anghel 2003). The second study contains 13 tetra- and pentameric STR loci population data, and it was performed on Romanians from the Bucharest area (Barbarii, 2004).

I.8 CODIS (13 core STR) Related Population Studies in Romania. The first article containing data of two Hungarian speaking populations from Transylvania (the Szekler and the Csángó populations from Harghita county) was published in 2005 (Egyed et al. 2005). Allele frequencies for 219 unrelated western Romanian individuals were determined in 2006 (Marian at al. 2007). In 2008 Allele frequencies for 15 STR loci included in

AmpFlSTR Identifiler kit were determined in the region of Transylvania, Wallachia, Dobruja, and Moldavia (Sanciu et al. 2009).

I.9 Legal Implication of STR Studies. As a result of the overwhelming success of DNA typing achieved during the last decade, a requirement has emerged to establish the legal basis and the standards of quality for DNA related activities. The 2009/C 296/01 EU Council Resolution brought the latest changes to forensic DNA typing by introducing the new European Standard Set containing 12 DNA STR markers (*Table 1.1*). In the US and Canada, The Scientific Working Group on DNA Analysis Methods (SWGDAM) was issued in January 2010, a new guideline for the interpretation of DNA typing results from short tandem repeats (STRs). Laboratories that analyze DNA samples for forensic casework purposes are required by the Quality Assurance Standards to be tested in Forensic DNA Testing Laboratories (effective July 1, 2009). As a member country of the EU, Romania must conform to the ENFSI recommendations. Also, since 2008 there has been a strict legislation (76/2008) regarding DNA typing and DNA databases (*Legea nr. 76/2008 privind organizarea si functionarea Sistemului National de Date Genetice Judiciare, Publicat in Monitorul Oficial, Partea I nr. 289 din 14/04/2008*)

Part II: Methods and Techniques Used in STR DNA Typing

II.1 Extraction of Genomic DNA from 300µl whole blood was performed using SV Wizard genomic DNA Purification Kit A1620 following the recommendations of the manufacturer (SV Wizard Genomic DNA Purification Kit; Promega 1999).

II.2 Multiplex DNA Amplification. AmpFLSTR Identifiler PCR Amplification Kit was used to amplify 15 STR loci located on 13 autosomes and the Amelogenin, the gender determination locus (Butler 2010). The Identifiler kit includes: AmpF*l*STR PCR Reaction Mix, AmpF*l*STR Identifiler Primer Set, AmpliTaq Gold DNA Polymerase, GeneScan 500 LIZ, AmpF*l*STR Control DNA 9947A, AmpF*l*STR Identifiler Alleleic Ladder. The manufacturer's protocol regarding reactive quantity required for PCR Master mix was modified to values included in *Table 2.1.*

Reactive		**Manufacturer's quantities /sample (µl)**	**Modified quantities /sample (µl)**
PCR MM	AmpFlSTR PCR Reaction Mix	10.5	5.25
	AmpliTaq Gold DNA Polymerase	0.5	0.25
	AmpFlSTR Identifiler Primer Set	5.5	2.75
	H_2O (UPUV, filtered)	-	4
PCR-MM loaded into tube		**15**	**12**
Genomic DNA		**10** (0.05–0,125 ng/µl)	**1** (~0.5 ng/µl)
Reactive loss		*-1*	*-0.25*
Total amount (PCR MM + DNA)		**25**	**13**

Table 2.1: PCR Master Mix components: the original (ABI provided) and the modified PCR protocol.

The amplification program included in *Table 2.2* was used according to AmpF*l*STR Identifiler PCR Amplification Kit manual (AmpFLSTR Identifiler PCR Amplification Kit User's Manual 2006). Following ABI's storage recommendations in case of short-term storage (max. 48 hours), the amplified DNAs were stored at 2-6 ºC, and in case of long-term storage (more than 48 hours) PCR products were stored at -15 to -25 °C (AmpFLSTR Identifiler PCR Amplification Kit User's Manual 2006).

Initial incubation	**Thermal cycling**	**Final Extension**
95 ºC – 11 minutes	94 ºC – 1 minute (denaturize) 59 ºC – 1 minute (anneal) 72 ºC – 1 minute (extend)	60 ºC – 60 minutes
-	Repeat the cycling steps 28 times	-

Table 2.2: AmpFLSTR Identifiler PCR Amplification program provided by Applied Biosystems

II.3 Labelling DNA Fragments required Hi-Di formamide, AmpFlSTR LIZ-500 (fluorescent internal marker) and AmpFlSTR Identifiler Alleleic Ladder (external marker). The reaction tubes containing the reactives *(Table 2.3)* were heated in a heat block for 5 minutes at 95°C (to denature the DNA) and immediately chilled for 5 minutes at -15ºC prior to loading in the ABI 310 (Jakovski et al. 2010).

Reactive	Internal marker quantities /sample (µl)	External marker quantities /sample (µl)
Hi-Di™ formamid	12	12
GeneScan 500 LIZ	0.5	0.5
PCR product	1.5	-
AmpF*l*STR Identifiler Allelic Ladder	-	1.5
TOTAL amount of reactives/tube	**14**	**14**

Table 2.3: Reactives used to mark PCR amplicons. Preparing internal marker: 1.5µl of each PCR product is mixed with 12µl formamide and 0.5µl LIZ-500 size standard. Preparing external marker: 1.5 µl AmpFlSTR Identifiler Alleleic Ladder is mixed with 12µl formamide and 0.5µl LIZ-500 size standard.

II.4 DNA Typing (fragment analyzing, genotyping, profiling) was carried out with ABI PRISM® 310 Genetic Analyzer, product of Applied Biosystems (ABI), at the Molecular Biology Center of Interdisciplinary Research Institute on Bio-Nano-Sciences (Cluj-Napoca). This capillary electrophoresis equipment provides high resolution for short fragments and uses a minimal amount (1.5µl) of PCR product (AmpFLSTR Identifiler PCR Amplification Kit User's Manual 2006).

II.5 Statistical Analysis. Allele frequencies and forensic efficiency parameters (Matching Probability, Power of Discrimination, Polymorphism Information Content, Power of Exclusion, Typical Paternity Index) can be calculated using the Power Stats Microsoft Excel workbook template provided by Promega Corporation (Madison, Wis., *ttp://www.promega.com/geneticidtools/*). Possible departures from the Hardy-Weinberg expectations, population comparison and differentiation tests were carried out with Arlequin v3.5 software.

Part III: Results

III.1 Sample and Population Information. According to the Romanian 2002 census, in Transylvania there is a considerable Hungarian community (19.6%), mostly in the Szekler counties of Covasna and Harghita. In the last decade, genetic parameters for Szekler (HR-Sze) and Csángó (HR-Csn) populations from Harghita have been published, but no population study has been performed on Szekler communities from Covasna and other (non-Szekler) Hungarians from Transylvania. In consequence, a total number of 733 blood samples were collected from the two counties: 206 samples from Cluj county (CJ-Hu) collected until 31.07.2008 and 527samples from Covasna county (CV-Sze) (*Fig 3.1*)

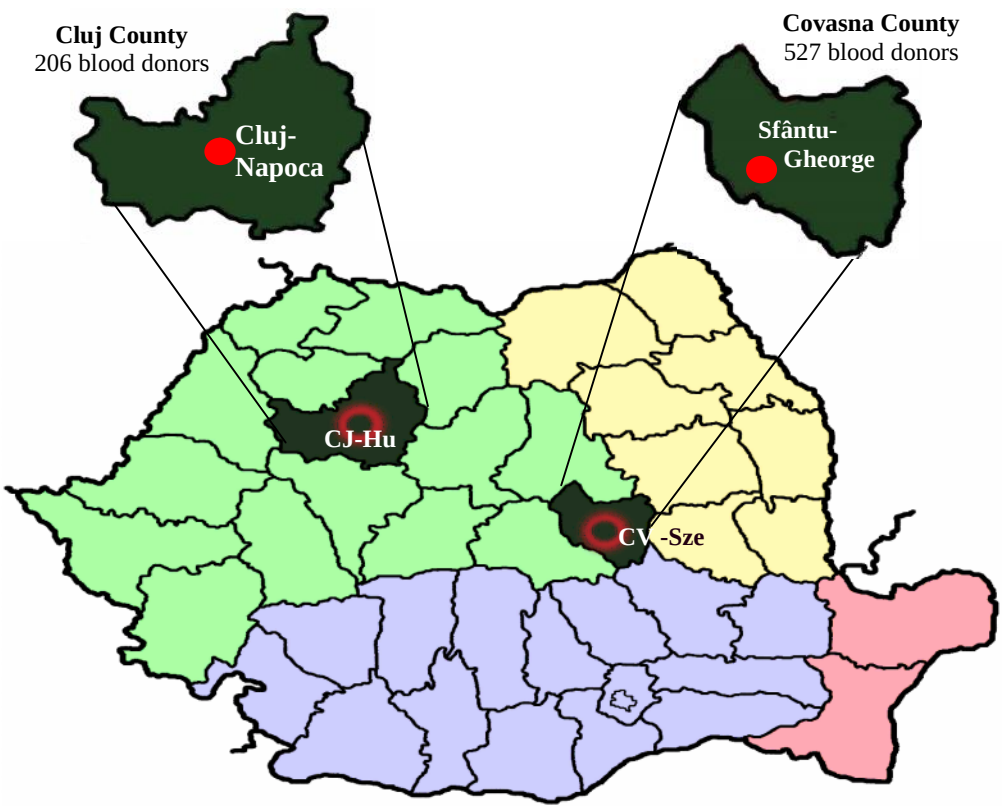

Fig 3.1: Geographic area of the populations included in this study, CJ-Hu = non Szekler Hungarians from Cluj county; CV-Sze = Szeklers from Covasna county

The blood sample donors were people belonging to the Caucasian race of Hungarian nationality, with ages between 0.5 and 79, randomly selected by employees of clinical hospitals in the above mentioned counties. Samples were collected in ethylenediamine tetra-acetic acid (EDTA) vacuum tubes with EDTA K2/EDTA K3 as anticoagulant. The average storage time of samples (time between collection and time of isolation) was 48 hours. All the subjects gave their consent to being taken into this study.

III.2 DNA Quantification and Purity Check. In order to obtain quantity and purity information about our extracted genomic DNA, we performed agarose gel electrophoresis and spectrophotometer measurements.

Agarose gel electrophoresis was performed prior to spectrophotometer check. The amount of DNA was checked by UV-photometry using Ultraviolet Transluminator provided by BioImaging Systems (GelDoc-It Imaging System) hardware and LabWorks Image Acquisition and Analysis (ver. 4.5.00.0) software.

Spectrophotometer measurements were performed with NanoDrop spectrophotometer ND-1000. Our interested data were DNA quantity (ng/µl), 260/280 ratio (check for protein contamination; ~1.8 value is accepted) and 260/230 ratio (check for other contamination; must be greater than 260/280 value; Wilfinger et al. 1997).

As a result for DNA quantification, 594 (81.04%) samples of 733 isolated DNA had over 2ng/µl concentration.

III.3 Quality Control. Introducing controls at each step of the DNA process allows the analyst to identify and troubleshoot possible issues and ensure that the methods used produce accurate and reliable results. *Table 3.1* summarizes the control methods applied in each step of the DNA typing process.

Control step	Control method	Reagents of control method
DNA extraction	Blank extraction	SV Wizard genomic DNA Purification Kit
DNA fragment amplification	Positive control Negative control	AmpF*l*STR Control DNA 9947A PCR Master Mix
DNA fragment labeling	Negative control	Hi-Di formamide
Fragment analysis	Internal sizing standard	GeneScan500 LIZ
Allelic designation	Allelic ladder	AmpF*l*STR Identifiler Allelic Ladder

Table 3.1 A summary of control methods for DNA typing process

As a result of rigorous selection and filter procedures described above, genetic parameters for 424 DNA profiles were finally evaluated and processed for population data analysis and forensic purposes (*Table 3.2*).

Number of isolated genomic DNA samples	Number of samples with ≥ 2 ng/µl DNA concentration	Genetic profiles*
206 (Cluj County)	179 (86.9%)	146 (81.6%)
527 (Covasna County)	415 (78.8%)	278 (67%)
Total 733	**594 (81.04%)**	**424 (71.4%)**

Table 3.2: The effects of quality control procedures on sample numbers (fractional summary); * percentages were calculated from the values of the second column.

III.4 Allelic Designation was done according to recommendations from ENFSI with the aid of AmpFlSTR Identifiler Alleleic Ladder, carried out with Genotyper v3.7 software for Windows NT. Size precision of allelic ladder was compared with the values for positive control DNA sample. If there was any mismatch identified between the typed-designated values and the positive control values, it was required to proceed to Size Precision Correction process.

Difficulties in data interpretation, the genotype assignment for all 424 samples, and the Allele Size Precision Correction processes are not detailed in this summary.

Allelic designation and genotype assignment for all 424 samples were obtained as detailed in *Figure 3.2*.

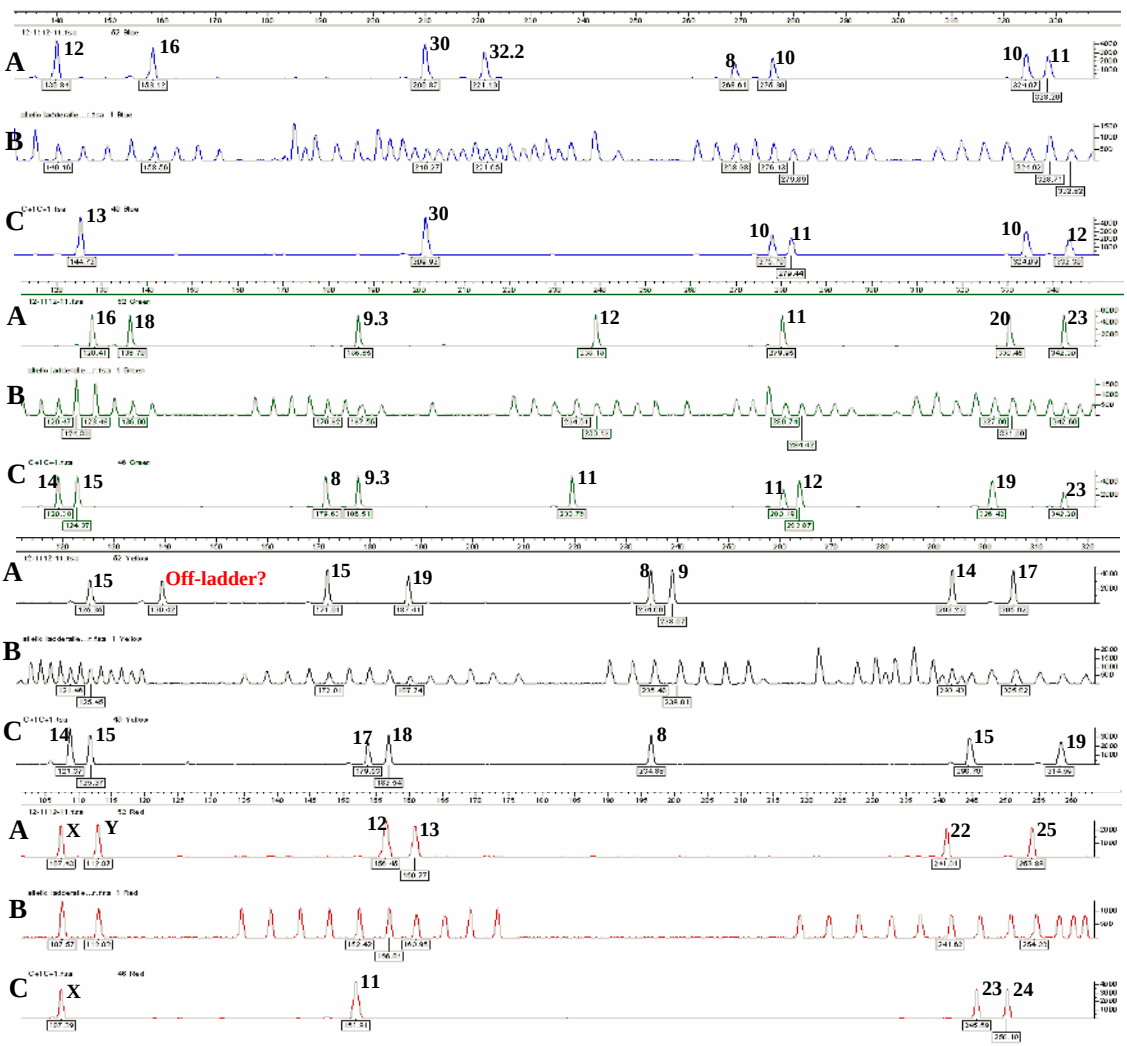

Figure 3.2: Allelic designation process; **(A)** DNA sample 12/11 (CJ135), **(B)** allelic ladder, **(C)** positive control sample (DNA9947A).

Off-ladder alleles were determined by being compared to allelic ladder and positive control (*Figure 3.3*) or allelic ladder and a nearby allele identified in another sample (*Figure 3.4*)

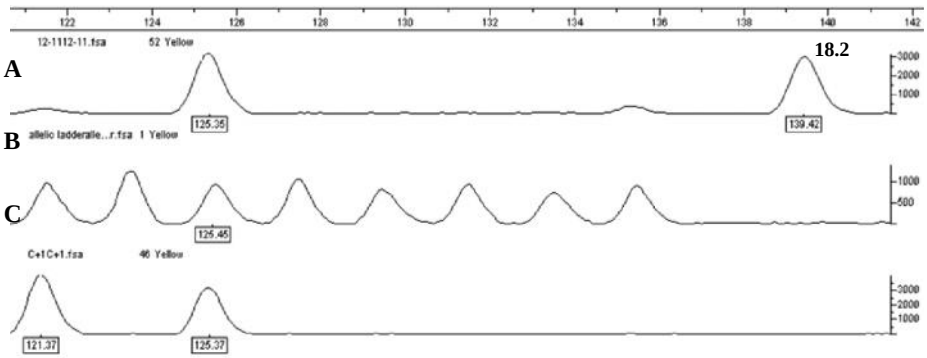

Figure 3.3: Off-ladder allele 18.2 for D19S433 locus obtained with Genotyper v3.7 software. **(A)** DNA sample 12/11 (CJ135), **(B)** allelic ladder, **(C)** positive control sample (DNA9947A).

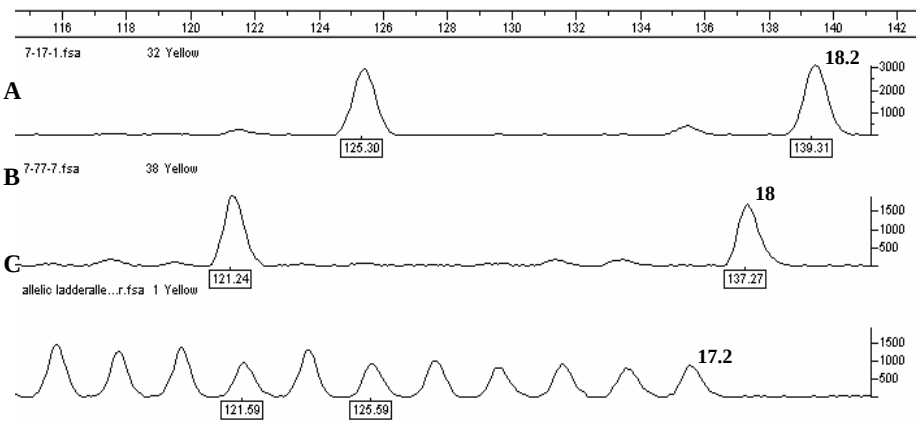

Figure 3.4: Off-ladder allele 18 and 18.2 for D19S433 locus obtained with Genotyper v3.7 software. **(A)** DNA sample 7/1 (CJ049), **(B)** DNA sample 7/7 (CJ054), **(C)** allelic ladder.

A total number of 16 Off-ladder alleles were identified involving 3 loci: D19S433, D2S1338 and FGA. They were reported to STRBase. The summary of these alleles is included in *Table 3.3.*

Locus	Off-ladder allele	Allele size (bp)	Frequency	Verification method	Reported in Romania
D19S433	18	137.27	1 of 424	Re-extract and re-amplify	Csángó, Transylvanian and Moldavia populations
D19S433	18.2	139.31	3 of 424	Re-extract and re-amplify	Dobruja
D2S1338	14	303.85	1 of 424	Re-extract and re-amplify	Dobruja
FGA	32.2	244.17	7 of 424	Re-extract and re-amplify	all four regions
FGA	33.2	248.01	1 of 424	Re-extract and re-amplify	all four regions

Table 3.3: Off-ladder alleles identified in 424 individuals representing the population of Cluj and Covasna counties; bp = bais pair.

III.5 Forensic Related Data.

Statistical values for Matching Probability (MP), Power of Discrimination (PD), Combined Power of Discrimination (PDcomb), and Polymorphism Information Content (PIC) were obtained with PowerStats v1.2 software.

Matching Probability (MP). The average MP values for Cluj and Covasna counties were 0.0864 and 0.0784. As summarized in *Table 3.4*, Cluj county's MP values for 12 of 15 loci were greater than for values obtained in Covasna. The highest MP values were obtained for TPOX locus in boat counties. Values above the average were also found for CSF1PO and D5S818 loci. MP for 4 loci (D21S11, D2S1338, D18S51 and FGA) shows values below the averages.

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ	0.0739	0.0494	0.0810	0.1287	0.0905	0.0855	0.0864	0.1014	0.0377	0.0680	0.0669	0.2198	0.0417	0.1207	0.0448
CV	0.0640	0.0386	0.0693	0.1196	0.0998	0.0882	0.0858	0.0850	0.0281	0.0669	0.0677	0.1779	0.0290	0.1192	0.0369

Table 3.4: Matching probability (MP) values for populations of Cluj (CJ) and Covasna (CV) counties, obtained with PowerStats v1.2 software.

Power of Discrimination (PD) and Combined Power of Discrimination (PDcomb) The average PD values for Cluj and Covasna counties were 0.9135 and 0.9215. PD values obtained with PowerStats v1.2 software are detailed in *Table 3.5.*

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ	0.9261	0.9506	0.9190	0.8713	0.9095	0.9145	0.9136	0.8986	0.9623	0.9320	0.9331	0.7802	0.9583	0.8793	0.9552
CV	0.9360	0.9614	0.9307	0.8804	0.9002	0.9118	0.9142	0.9150	0.9719	0.9331	0.9323	0.8221	0.9710	0.8808	0.9631

Table 3.5: Power of Discrimination (PD) values for populations of Cluj (CJ) and Covasna (CV) counties, obtained with PowerStats v1.2 software.

The PD values of all STR markers are close to the averages, except for TPOX locus, which was found to be below these values. The **PD*comb*** values obtained for the 15 loci were 0.99999999999999992 in the case of Cluj, and 0.999999999999999997 in the case of Covasna counties, which are significantly high to allow us to use these loci for individual identification purposes.

Polymorphism Information Content (PIC). The average PIC values for Cluj and Covasna counties were 0.7545 and 0.7673. PIC values were obtained with PowerStats v1.2 software and are detailed in *Table 3.6.*

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ	0.7627	0.8371	0.7513	0.6686	0.7461	0.7468	0.7339	0.7397	0.8655	0.7732	0.7956	0.5362	0.8579	0.6687	0.8351
CV	0.7884	0.8407	0.7778	0.7032	0.7380	0.7384	0.7419	0.7545	0.8750	0.7762	0.7841	0.5935	0.8672	0.6816	0.8495

Table 3.6: Polymorphism Information Content (PIC) values for populations of Cluj (CJ) and Covasna (CV) counties, obtained with PowerStats v1.2 software.

Values for 12 loci are greater for Covasna when compared to Cluj county. Three loci (CSF1PO, TPOX and D5S818) show values below the averages. Except for TPOX locus, there were no significant differences between the values obtained for these two populations.

III.6 Paternity Related Data.

Statistical values for Power of Exclusion (PE), Combined Power of Exclusion (PEcomb), and Typical Paternity Index (PI) were obtained with PowerStats v1.2 software.

Power of Exclusion (PE) and Combined Power of Exclusion (PEcomb). The average PE values for Cluj and Covasna counties were 0.6011 and 0.5838. PE values obtained with PowerStats v1.2 software are detailed in *Table 3.7.*

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ	0.5275	0.8179	0.5889	0.3657	0.6535	0.5639	0.5639	0.6273	0.8742	0.5367	0.6935	0.2623	0.8601	0.3753	0.7070
CV	0.6577	0.6716	0.5700	0.5190	0.6098	0.5253	0.5128	0.5765	0.7501	0.5442	0.5897	0.3617	0.7501	0.4194	0.6998

Table 3.7: Power of Exclusion (PE) values for populations of Cluj and Covasna counties obtained with PowerStats v1.2 software

CSF1PO, TPOX and D5S818 loci show values below averages (*Figure 3.12*). The higher values were obtained for D21S11, D2S1338 and D18S51 loci. PE values for 10 of 15 loci were greater for Cluj county, when compared to Covasna's values. **The Combined Power of Exclusion (PEcomb) Values** obtained for the 15 loci were 0.9999998 in case of Cluj, and 0.999998 in case of Covasna counties, which are high enough to allow using these loci for individual identification purposes.

Typical Paternity Index (PI). The average PI values for Cluj and Covasna counties were 3.2498 and 2.5791. PI values obtained with PowerStats v1.2 software are detailed in *Table 3.8.*

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ	2.09	5.58	2.43	1.46	2.92	2.28	2.28	2.70	8.11	2.13	3.32	1.18	7.30	1.49	3.48
CV	2.96	3.09	2.32	2.04	2.57	2.07	2.01	2.36	4.09	2.17	2.44	1.45	4.09	1.64	3.39

Table 3.8: Typical Paternity Index (PI) values for populations of Cluj (CJ) and Covasna (CV) counties, obtained with PowerStats v1.2 software.

Values of 8 (Covasna) and 10 loci (Cluj county) were below the averages. Very high values were obtained in the case of Cluj county for D21S11, D2S1338, and D18S51. All values were situated above 1.

III.7 Population Genetic Data

Statistical values for Heterozygosity (H), Hardy-Weinberg Equilibrium (HWE) exact test were obtained with Arlequin V3.5 software. Allele frequencies were calculated using allele count method, and were computed with Microsoft Office Excel 2003 Workbook.

Observed (Hobs) and Expected (Hexp) Heterozygosity. The average heterozygosity level for both observed and expected values were above 78%. The required level of 70% was reached, so these loci are suitable for use in human identification. The average Hobs value was greater for Cluj (0.794) than for Covasna (0.7889) county. The opposite was observed with the Hexp values (CJ=0.7857; CV=0.7964). The observed values were higher than expected for 8 (Cluj) and 4 (Covasna) of 15 loci (*Table 3.9*).

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ Hobs	0.7603	0.9110	0.7945	0.6575	0.8288	0.7808	0.7808	0.8151	0.9384	0.7671	0.8493	0.5753	0.9315	0.6644	0.8562
CJ Hexp	0.7907	0.8532	0.7820	0.7208	0.7812	0.7811	0.7671	0.7745	0.8778	0.7978	0.8202	0.5987	0.8714	0.7167	0.8525
CV Hobs	0.8309	0.8381	0.7842	0.7554	0.8058	0.7590	0.7518	0.7878	0.8777	0.7698	0.7950	0.6547	0.8777	0.6942	0.8525
CV Hexp	0.8127	0.8563	0.8061	0.7466	0.7744	0.7738	0.7748	0.7866	0.8859	0.8006	0.8111	0.6482	0.8795	0.7262	0.8646

Table 3.9: Table with the observed **(**Hobs) and expected (Hexp) heterozygosity values for 15 loci studied in the populations of Cluj (CJ) and Covasna (CV) counties.

In the case of Cluj county, observed heterozygosity values were situated between 0.5753 (TPOX) and 0.9384 (D2S1338). The expected heterozygosity values were situated between 0.5987 (TPOX) and 0.8778 (D2S1338). Hobs values for 7 of 15 loci were below the average and 3 of 15 loci were under the recommended 70%. Hexp values for 5 of 15 loci were found to be below the average and only one locus (TPOX) under the recommended 70%.

In the case of Covasna county, observed heterozygosity values were situated between 0.5753 (TPOX) and 0.9384 (D2S1338). The expected heterozygosity values were situated between 0.6547 (TPOX) and 0.8777 (D2S1338 and D18S51). Hobs values for 8 of 15 loci were below the average and 2 of 15 loci were under the recommended 70%. Hexp values for 7 of 15 loci were found to be below the average and only one locus (TPOX) under the recommended 70%.

Hardy-Weinberg Equilibrium (HWE) Exact Test

In the case of the population of Cluj county, HWE test value for 2 of the 15 tested loci were found to be below the 0.05 significance level: CSF1PO P=0.021 and D18S51 P=0.0036 (*Table 3.10*). No deviations from the Hardy–Weinberg equilibrium were observed after applying a Bonferroni-type correction: 0.05/15 = 0.0033.

In the case of the population of Covasna county, deviations from the Hardy–Weinberg equilibrium (P < 0.05) were observed at CSF1PO, D16S539 and D19S433 loci (*Table 3.10*). Significant P-values for D16S539 and D19S433 were eliminated when a Bonferroni-type correction was performed.

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
CJ	0.2471	0.1511	0.0996	*0.0210*	0.2993	0.0519	0.4589	0.1653	0.0627	0.0720	0.2215	0.5769	*0.0036*	0.2844	0.2631
CV	0.2342	0.4394	0.1145	*0.0014*	0.1019	0.3677	0.2346	*0.0049*	0.0738	*0.0397*	0.1182	0.1952	0.6350	0.5244	0.0637

Table 3.10: Table with HWE exact test P values for 15 autosomal STR loci studied in the populations of Cluj (CJ) and Covasna (CV) counties; values in italics show deviations from HWE at 0.05 significance level.

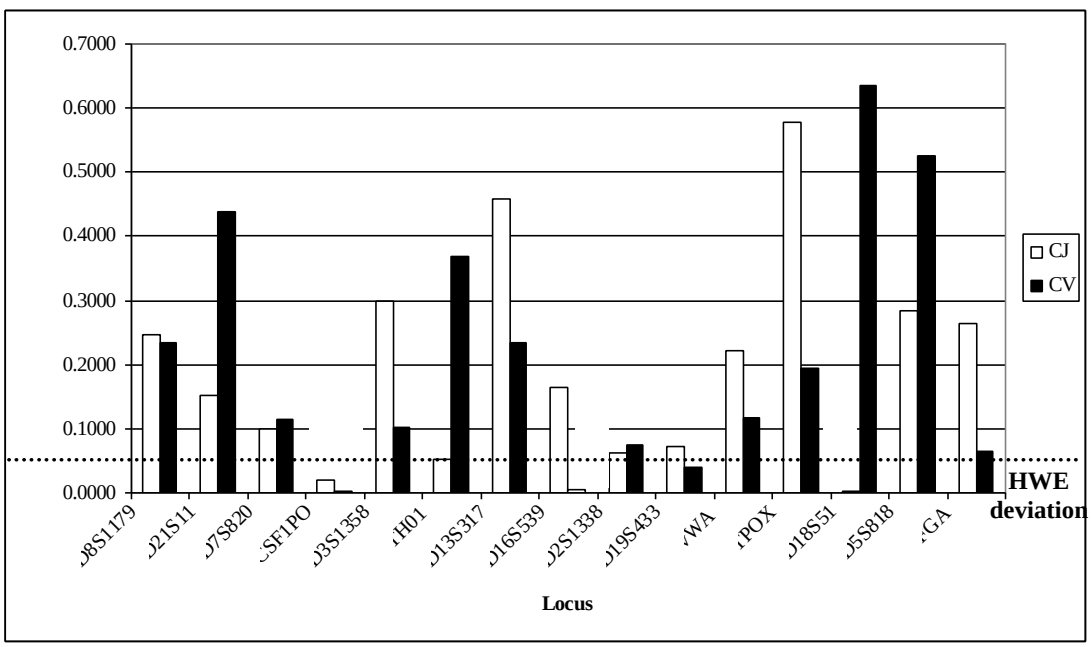

Figure 3.5: Column chart of HWE exact test P values for 15 autosomal STR loci studied in the populations of Cluj (CJ) and Covasna (CV) counties.

Since only one test (Covasna's CSF1PO locus) registered departures from HWE, there is no cause for HWE rejection.

Allele Frequencies

The observed allele frequencies for the 15 STR loci tested in Cluj and Covasna counties were calculated using allele count method and were computed with Microsoft Office Excel 2003 Workbook. Allele frequencies calculated for Hungarians from Cluj county are included in *Table 3.11.* Allele frequencies calculated for the Szekler population from Covasna county are included in *Table 3.12.*

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
6						0.2466									
7			0.0171			0.1301									
8	0.0205		0.1541	0.0068		0.1267	0.1267	0.0103				0.5480		0.0034	
9	0.0068		0.1027	0.0274		0.1849	0.0856	0.0993		0.0034		0.0891	0.0034	0.0377	
9.3						0.3014									
10	0.0616		0.3356	0.2774		0.0103	0.0513	0.0719				0.0445	0.0137	0.0788	
10.2													0.0034		
11	0.0616		0.2398	0.3253			0.3391	0.2945		0.0172		0.3013	0.0103	0.3322	
12	0.1747		0.1062	0.3049			0.2946	0.2877		0.0651	0.0034	0.0171	0.0822	0.3733	
13	0.3356		0.0445	0.0514			0.0651	0.1986		0.2637	0.0034		0.1507	0.1610	
13.2										0.0342					
14	0.2090			0.0068	0.1164		0.0308	0.0377		0.3253	0.1096		0.1953	0.0068	
14.2										0.0342					
15	0.1200				0.2568		0.0068			0.1165	0.1267		0.1541	0.0068	
15.2										0.0548					
16	0.0068				0.2329				0.0753	0.0548	0.2260		0.1267		
16.2										0.0172					
17	0.0034				0.2671				0.1883	0.0034	0.2364		0.1233		
18					0.1164				0.0925	0.0034	0.2021		0.0548		0.0068
18.2										0.0068					
19					0.0104				0.0788		0.0616		0.0480		0.0719
20									0.1575		0.0274		0.0102		0.1233
21									0.0240				0.0171		0.1507
22									0.0240		0.0034		0.0034		0.2193
22.2															0.0068
23									0.1370						0.1473
24									0.0925				0.0034		0.1678
24.2		0.0034													
25									0.1130						0.0753
26		0.0034							0.0137						0.0274
27		0.0274							0.0034						
28		0.1267													
29		0.1815													0.0034
29.2		0.0034													
30		0.2467													
30.2		0.0377													
31		0.0856													
31.2		0.0993													
32		0.0240													
32.2		0.1267													
33.2		0.0308													
35		0.0034													

Table 3.11: Allele frequencies for the 15 STR loci, tested in Hungarians from Cluj county.

Allele	D8S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D19S433	vWA	TPOX	D18S51	D5S818	FGA
6						0.2086						0.0018			
7			0.0144	0.0018		0.1619						0.0036	0.0018	0.0018	
8	0.0126		0.1529	0.0126		0.1007	0.1511	0.0198				0.4982		0.0018	
9	0.0144		0.1763	0.0701		0.1979	0.0791	0.1565				0.1205		0.0431	
9.3						0.3273									
10	0.0827		0.2716	0.3004		0.0036	0.0576	0.0594				0.0594	0.0180	0.0863	
10.2													0.0036		
11	0.0755		0.2104	0.3040			0.3147	0.2481		0.0018		0.2913	0.0126	0.2752	
12	0.1440		0.1421	0.2517	0.0018		0.2932	0.2985		0.0791		0.0252	0.1529	0.3904	
12.2										0.0036					
13	0.2931		0.0323	0.0468			0.0863	0.1817		0.2301	0.0018		0.1259	0.1906	
13.2										0.0270			0.0018		
14	0.2249			0.0090	0.0791		0.0180	0.0360	0.0018	0.3308	0.1169		0.1565	0.0036	
14.2										0.0360					
15	0.1295			0.0036	0.2464				0.0018	0.1511	0.1205		0.1241	0.0072	
15.2										0.0450					
16	0.0198				0.2985				0.0630	0.0504	0.2338		0.1474		
16.2										0.0343					
17	0.0035				0.2195				0.1834	0.0054	0.2428		0.0989		0.0054
17.2										0.0036					
18					0.1457				0.1295		0.2050		0.0809		0.0108
18.2										0.0018					
19					0.0090				0.0953		0.0702		0.0450		0.0486
20									0.1169		0.0090		0.0288		0.1546
21									0.0306						0.1601
22									0.0486						0.1690
22.2															0.0090
23									0.1295				0.0018		0.1367
23.2															0.0018
24									0.0845						0.1601
25									0.0971						0.0899
26		0.0072							0.0180						0.0468
27		0.0288													0.0072
28		0.1331													
29		0.2266													
29.2		0.0054													
30		0.2068													
30.2		0.0575													
31		0.0630													
31.2		0.0845													
32		0.0162													
32.2		0.1205													
33.2		0.0414													
34.2		0.0054													
35.2		0.0018													
38		0.0018													

Table 3.12: Allele frequencies for the 15 STR loci, tested in the Szekler population from Covasna county.

III.8 Population Comparison and Exact Population Differentiation Tests Results

A comparison of genetic parameters for these two counties was performed. Significant values for non-Szekler Hungarians (CJ-Hu) were obtained for locus D5S818 versus Szeklers from Harghita (HR-Sze; Fst and exact test P <0.001), and for locus D3S1358 versus Csángós (HR-Csn; Fst P=0.031; exact P=0.024), Transylvanian population (Trs; Fst P=0.014) and Wallachian population (Fst P=0.016). When comparing Covasna's Szekler population data with the same populations as in the case of CJ-Hu, more significant values were obtained. Significant values for one locus were found in the comparison with the population of Szeklers from Harghita (D5S818), Western Romania and Dobruja (TH01). Significant differences were found for two loci when compared to HR-Csn (D3S1358, D16S539), Transylvanian population, and Moldavia (CSF1PO, D2S1338). For 6 loci (D21S11, CSF1PO, TH01, D16S539, D2S1338 and D18S51), population data of Wallachia shows the greatest distance from CV-Sze. Detailed comparison results, including additional significant values obtained versus Romanians from Bucharest area, are summarized in *Table 3.13.*

Marker	Test	Population data	CV-Sze n=278	HR-Sze n=257	HR-Csn n=220	B-Ro n=243	Trs n=1977	West n=219	Moldavia n=1321	Dobruja n=569	Wallachia n=1910
D21S11	Fst	CV-Sze vs		0.612	0.104	0.730	0.863	0.455	0.747	0.833	0.634
	Exact test	CV-Sze vs		0.934	0.313	0.601	0.308	0.756	0.344	0.817	**0.018**
D7S820	Fst	CJ-Hu vs	0.217	0.200	0.102	nd	0.055	0.451	0.180	0.064	0.076
	Exact test	CJ-Hu vs	0.387	0.271	**0.047**	nd	0.159	0.562	0.442	0.232	0.287
CSF1PO	Fst	CV-Sze vs		0.366	0.412	nd	**0.017**	0.294	**0.035**	0.202	**0.040**
	Exact test	CV-Sze vs		0.326	0.533	nd	**<0.001**	0.323	**<0.001**	0.047	**0.003**
D3S1358	Fst	CV-Sze vs		0.886	**0.024**	0.160	0.067	0.870	0.474	0.763	0.312
	Fst	CJ-Hu vs	0.368	0.585	**0.031**	0.220	**0.014**	0.260	0.152	0.193	**0.016**
	Exact test	CV-Sze vs		0.870	**0.006**	0.167	0.214	0.897	0.627	0.876	0.402
	Exact test	CJ-Hu vs	0.496	0.765	**0.024**	0.285	0.058	0.348	0.231	0.472	0.065
TH01	Fst	CV-Sze vs		0.711	0.648	0.315	0.137	**0.049**	0.250	**0.047**	**0.034**
	Exact test	CV-Sze vs		0.463	0.549	0.398	0.321	**0.020**	0.553	0.172	0.220
D16S539	Fst	CV-Sze vs		0.612	**0.038**	0.335	0.387	0.519	0.312	0.355	**0.011**
	Exact test	CV-Sze vs		0.629	**0.010**	0.532	0.270	0.542	0.364	0.373	**0.010**
D2S1338	Fst	CV-Sze vs		0.347	0.660	0.127	0.088	0.573	0.163	0.404	0.146
	Exact test	CV-Sze vs		0.350	0.656	0.065	**0.001**	0.554	**0.016**	0.245	**0.016**
D18S51	Fst	CV-Sze vs		0.403	0.494	0.321	0.242	0.839	0.389	0.346	**0.070**
	Exact test	CV-Sze vs		0.208	0.604	0.069	0.052	0.877	0.014	0.198	**0.001**
D5S818	Fst	CV-Sze vs		**<0.001**	0.959	nd	0.805	0.814	0.982	0.835	0.958
	Fst	CJ-Hu vs	0.610	**<0.001**	0.877	nd	0.262	0.830	0.696	0.271	0.353
	Exact test	CV-Sze vs		**<0.001**	0.828	nd	0.126	0.586	0.221	0.849	0.327
	Exact test	CJ-Hu vs	0.970	**<0.001**	0.895	nd	0.330	0.693	0.559	0.602	0.531

Table 3.13: Pairwise population comparison (Fst) and exact population differentiation test (Exact test) results for the populations of Cluj (CJ) and Covasna (CV) counties. **CV-Sze** = Szeklers from Covasna county; **HR-Sze** = Szeklers from Harghita county; **HR-Csn** = Csángós from Harghita county; **B-Ro** = Romanians from Bucharest area; **Trs** = Transylvanian database; **West** = population of Western Romania; values marked with bold represent significant values at P<0.05 significance level.

III.9 Updated Transylvanian Database (TRS-6)

Since genomic DNA typing projects were started in Romania, five studies representing six populations were carried out in Transylvania, and the initial Transylvanian STR database contained 1977 individuals. Although for one (Zalău) out of the sixteen Transylvanian counties there are no population data available, we consider that by adding data characterizing the population of Western Romania, Szeklers from Harghita, Csángós from Harghita, Szeklers from Covasna and Hungarians from Cluj to the initial Transylvanian database (Trs) we contribute to the completeness of the Transylvanian database (*Table 3.14*). All these results were cumulated and an updated Transylvanian database (Trs-**6**; contains data for **six** populations) was created (*Figure 3.6*).

No.	Population	Number of individuals	Fractions represented in the Transylvanian STR database
1	TRS	1977	64
2	Cluj	146	5
3	Covasna	278	9
4	Harghita Szekler	257	8
5	Harghita Csángó	220	7
6	Western Romania	219	7
Total TRS-6		**3097**	**100**

Table 3.14: The composition of the updated Transylvanian database (Trs-6): population name, number of individuals per each population, and the fraction of these populations in the Transylvanian STR database.

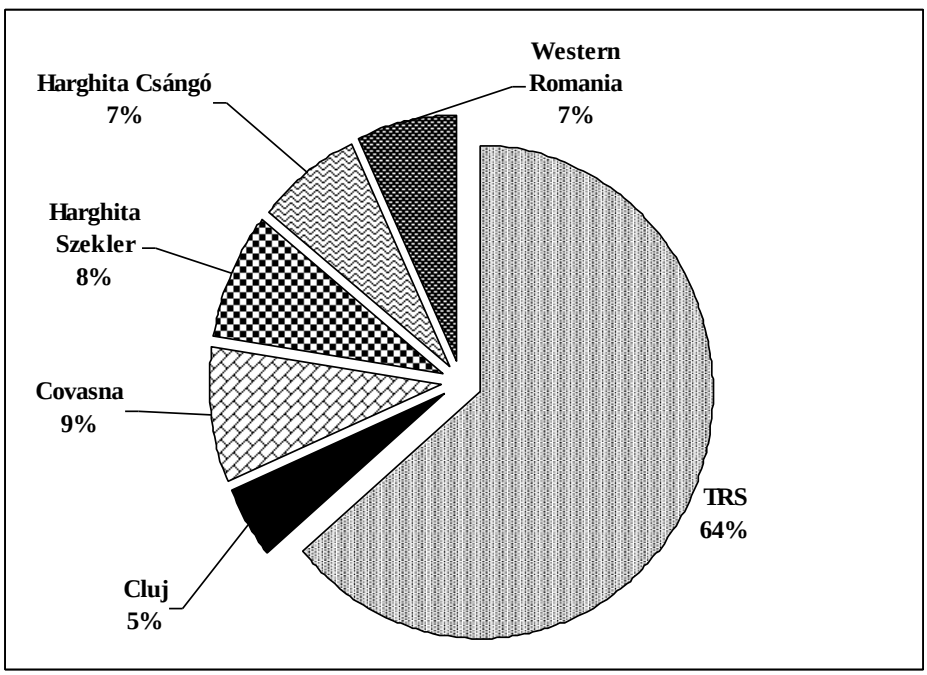

Figure 3.6: The updated Transylvanian database containing six populations and their fractions represented in this STR database.

Genetic Variances in Transylvanian Populations. The updated Transylvanian database (Trs-6) contains genetic data of six populations. Genetic variances expressed in statistical data and allele frequencies are not detailed in this summary.

Biostatistical Comparison of Transylvanian Populations

A comparison of the initial (Trs, n=1977) and updated Transylvanian databases (Trs-6, n=3097) was set up in order to check the consequences of the update.

Performing the pairwise population comparison test (Fst), a significant difference was identified at locus D21S11. As a result of exact population differentiation test, significant differences were observed at loci D21S11, D13S317, D2S1338, and vWA. When the two databases were compared to other population data like that of Moldavia (n=1321), Dobruja (n=569), Wallachia (1910), as well as the databases of Romania and the FBI, three cases concerning significant differences were observed.

The first situation identified refers to a given significance level, which in both databases is $P<0.05$: exact differentiation test P value for D21S11 and D13S317 loci, when compared to the entire Romanian database; Fst and exact test P values for D8S818 and D16S539 loci, when compared to databases of the FBI and Wallachia.

In the second situation, significant values in the comparison of the initial database lost their significance after the population data update: Fst P value for the D19S433 locus, when compared to Moldavia; exact test P value for the FGA locus, when compared to the FBI database; Fst and exact test P value for the D3S1358 locus, when compared to the FBI database.

The third and the most frequent situation identified refers to changes of initially insignificant values to significant ones: Fst and exact test P value for the D21S11 locus, when compared to the databases of Moldavia and Wallachia; exact test P value for the vWA locus, when compared to the databases of Moldavia, Dobruja, Wallachia, the FBI and Romania; exact test P value for the D13S317 locus, when compared to Moldavia, Dobruja, Wallachia.

A total number of 39 significant values involving 11 STR markers were identified, 29 related to the updated database.

Therefore, the use of the updated database is recommended any time when population data comparison with Transylvanian database has to be done.

Table 3.15: Comparison of the initial and updated Transylvanian databases with both pairwise population comparison (Fst) and exact population differentiation tests P values.

Marker	Test	Population	Trs n=1977	Moldavia n=1321	Dobruja n=569	Wallachia n=1910	Romania n=6897	FBI n=195-203
D8S1179	Fst	Trs-6 vs	0.496	0.801	0.597	0.450	0.917	0.423
	Fst	Trs vs		0.434	0.954	0.960	0.753	0.199
	Exact test	Trs-6 vs	0.768	0.593	0.594	0.647	0.989	0.380
	Exact test	Trs vs		0.295	0.927	0.901	0.864	0.148
D21S11	Fst	Trs-6 vs	0.002*	**0.007**	0.115	**<0.001**	0.055	0.245
	Fst	Trs vs		0.579	0.945	0.228	0.170	0.203
	Exact test	Trs-6 vs	<0.001*	**<0.001**	**<0.001**	**<0.001**	*<0.001*	**0.013**
	Exact test	Trs vs		0.251	0.726	0.631	*<0.001*	0.079
D7S820	Fst	Trs-6 vs	0.592	0.869	0.930	0.055	0.708	0.792
	Fst	Trs vs		0.796	0.994	0.533	0.867	0.400
	Exact test	Trs-6 vs	0.935	0.978	0.972	0.137	0.970	0.194
	Exact test	Trs vs		0.894	0.995	0.387	0.868	0.156
CSF1PO	Fst	Trs-6 vs	0.596	0.531	0.499	0.627	0.993	0.922
	Fst	Trs vs		0.423	0.329	0.686	0.585	0.922
	Exact test	Trs-6 vs	0.807	0.541	0.641	0.457	0.931	0.118
	Exact test	Trs vs		0.693	0.438	0.737	0.646	0.161
D3S1358	Fst	Trs-6 vs	0.297	0.640	0.938	0.256	0.959	0.161
	Fst	Trs vs		0.163	0.304	0.210	0.140	0.043†
	Exact test	Trs-6 vs	0.649	0.775	0.790	0.331	0.997	0.144
	Exact test	Trs vs		0.410	0.523	0.586	0.364	0.041†
TH01	Fst	Trs-6 vs	0.789	0.896	0.485	0.146	0.733	0.477
	Fst	Trs vs		0.680	0.865	0.449	0.850	0.383
	Exact test	Trs-6 vs	0.925	0.886	0.204	0.290	0.941	0.177
	Exact test	Trs vs		0.734	0.425	0.564	0.895	0.201
D13S317	Fst	Trs-6 vs	0.074	0.111	0.449	**0.043**	0.379	0.208
	Fst	Trs vs		0.628	0.760	0.767	0.501	0.208
	Exact test	Trs-6 vs	<0.001*	**<0.001**	**<0.001**	**<0.001**	*<0.001*	0.187
	Exact test	Trs vs		0.753	0.677	0.904	*<0.001*	0.342
D16S539	Fst	Trs-6 vs	0.757	0.462	0.878	*<0.001*	0.394	0.505
	Fst	Trs vs		0.472	0.983	*0.034*	0.816	0.417
	Exact test	Trs-6 vs	0.945	0.655	0.926	*0.005*	0.795	0.404
	Exact test	Trs vs		0.694	0.985	*0.012*	0.958	0.368
D2S1338	Fst	Trs-6 vs	0.478	0.720	0.865	0.658	0.974	n/a
	Fst	Trs vs		0.938	0.706	0.960	0.767	n/a
	Exact test	Trs-6 vs	<0.001*	**0.005**	0.334	**<0.001**	0.265	n/a
	Exact test	Trs vs		0.972	0.586	0.971	0.151	n/a

Marker	Test	Population	Trs n=1977	Moldavia n=1321	Dobruja n=569	Wallachia n=1910	Romania n=6897	FBI n=195-203
D19S433	Fst	Trs-6 vs	0.319	0.313	0.347	0.797	0.985	n/a
	Fst	Trs vs		0.016†	0.515	0.915	0.192	n/a
	Exact test	Trs-6 vs	0.949	0.835	0.410	0.865	0.999	n/a
	Exact test	Trs vs		0.290	0.788	0.780	0.873	n/a
vWA	Fst	Trs-6 vs	0.332	0.306	0.782	0.107	0.319	0.851
	Fst	Trs vs		0.682	0.982	0.883	0.959	0.574
	Exact test	Trs-6 vs	<0.001*	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.035**
	Exact test	Trs vs		0.691	0.948	0.616	0.869	0.694
TPOX	Fst	Trs-6 vs	0.909	0.433	0.498	0.546	0.952	0.290
	Fst	Trs vs		0.255	0.733	0.901	0.918	0.247
	Exact test	Trs-6 vs	0.986	0.527	0.453	0.967	0.983	0.405
	Exact test	Trs vs		0.451	0.333	0.940	0.888	0.214
D18S51	Fst	Trs-6 vs	0.846	0.952	0.207	0.080	0.584	0.442
	Fst	Trs vs		0.938	0.604	0.582	0.917	0.312
	Exact test	Trs-6 vs	0.994	0.606	0.560	0.139	**<0.001**	0.715
	Exact test	Trs vs		0.757	0.846	0.729	0.064	0.635
D5S818	Fst	Trs-6 vs	0.662	0.457	0.753	0.589	0.894	*<0.001*
	Fst	Trs vs		0.134	0.943	0.522	0.410	*0.003*
	Exact test	Trs-6 vs	0.977	0.663	0.816	0.926	0.996	*0.016*
	Exact test	Trs vs		0.610	0.758	0.925	0.970	*0.002*
FGA	Fst	Trs-6 vs	0.746	0.992	0.828	0.682	0.987	0.554
	Fst	Trs vs		0.976	0.964	0.631	0.799	0.218
	Exact test	Trs-6 vs	0.943	0.989	0.735	0.680	0.998	0.193
	Exact test	Trs vs		0.956	0.925	0.627	0.814	0.027†

n = number of samples; vs = *versus*; Trs-6 = updated Transylvanian database; Trs = initial Transylvanian database; FBI = US Caucasian population (Budowle et al); n/a = not available; values of Trs column marked with "*" symbol show significant differences between Trs-6 and Trs; values in italics show no changes of significant values when Trs and Trs-6 were compared to other populations; values marked with "†" symbol show how significant values of the initial database lose their significance after the population update; values in bold prove how initially insignificant values change to significant ones after update; significance level P<0.05

III.10 The Romanian STR Database

The updated Transylvanian data was cumulated with data from the other historical Romanian regions (Moldavia, Dobruja and Wallachia) and resulted in a comprehensive Romanian autosomal STR database which covers the entire country (*Table 3.16*).

No.	Region	Number of individuals	Fractions represented in the Romanian STR database
1	Trs-6	3097	45
2	Moldavia	1321	19
3	Wallachia	1910	28
4	Dobruja	569	8
Total		**6897**	**100**

Table 3.16: The composition of the Romanian database: number of individuals per each region, and the fractions of these in the Romanian STR database.

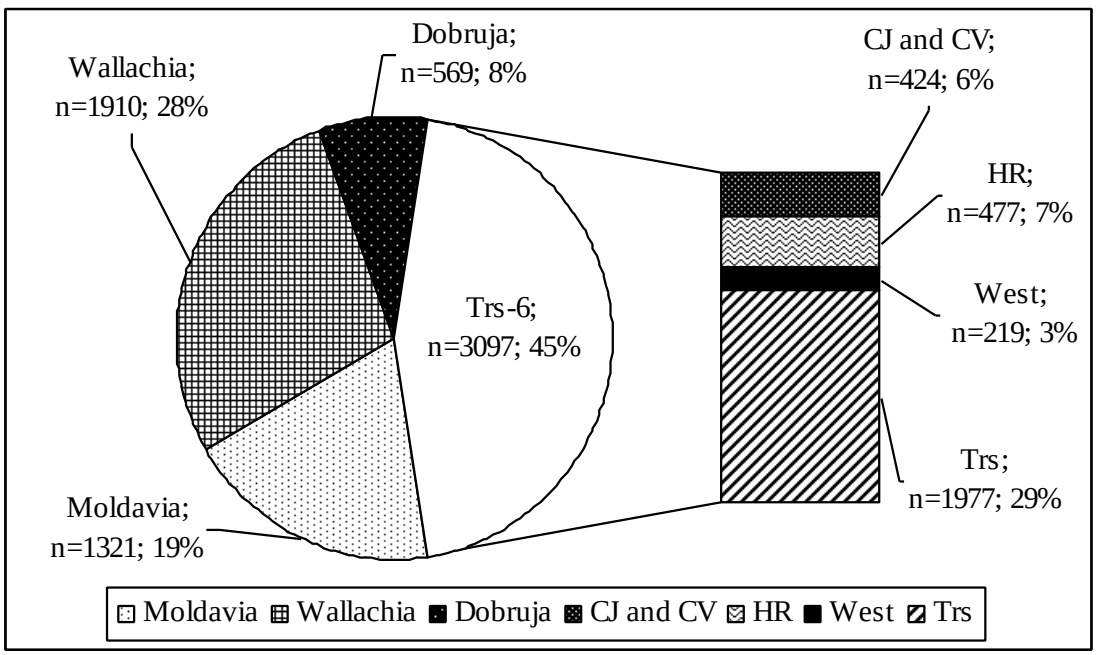

Figure 3.7: The Romanian STR database containing the four historical regions and their proportion represented in this database.

Genetic Variances for the Four Historical Romanian Regions. The Romanian STR database contains genetic data for the four historical regions: Transylvania, Wallachia, Moldavia and Dobruja. Genetic variances expressed in statistical data and allele frequencies are not detailed in this summary. The thesis also includes information about the American FBI databases.

Comparison of the Databases of the Romanian Regions. The results of this comparison defined D21S11, D13S317, D3S1358, and D5S818 loci as the most representative polymorphic autosomal STR markers of the Romanian population. When the entire Romanian population database was matched to the four regions, significant differences were found at loci D21S11, D13S317and vWA versus Trs-6 (n=3097) and at D21S11, D13S317 and D16S539 versus Wallachia (n=1910). On the other hand, comparison to the database of Dobruja (n=569) shows no departure from data in the Romanian database. As a result of comparison to the FBI database, the entire Romanian population shows the greatest genetic deviation at the D5S818 locus for both the pairwise population comparison and the exact population differentiation tests. Romanian regions were also compared to the FBI database. Significant differences at five loci were identified: D21S11 (vs. Trs-6 and Wallachia), D3S1358 (vs. Trs and Wallachia) vWA (vs. Trs-6), D5S818 (vs. all four regions) and FGA (vs. Trs).

In conclusion, D21S11, D13S317 and D3S1358, D5S818 loci seem to be the most polymorphic autosomal STR markers for the Romanian regions, when matched to the entire Romanian database and the FBI databases *(Table 3.17)*.

Marker	Population	Test	FBI n=195-203	TRS-6 n=3097	Trs n=1977	Moldavia n=1321	Dobruja n=569	Wallachia n=1910
D21S11	**Romania *vs.***	Fst	0.1914		0.2373	0.2920	0.7559	0.0840
	Romania* *vs.*	Fst	0.1475		0.9297	0.5391	0.9785	0.4619
	Romania *vs.*	Exact test	0.0519	**<0.001**	**<0.001**	**0.0004**	0.1302	**<0.001**
	Romania* *vs.*	Exact test	0.0744	**<0.001**	0.9948	0.4741	0.5488	0.9145
	FBI *vs.*	Fst		0.2637	0.1865	0.3428	0.2305	**0.0420**
		Exact test		**0.0160**	0.0682	0.4616	0.2086	**0.0177**
D3S1358	**FBI *vs.***	Fst		0.1611	**0.0430**	0.1719	0.4160	**0.0391**
		Exact test		0.1437	**0.0407**	0.0895	0.5541	**0.0417**
D13S317	**Romania *vs.***	Fst	0.1748		0.5010	0.5400	0.8643	0.3916
	Romania* *vs.*	Fst	0.1719		0.9404	0.7041	0.8848	0.8555
	Romania *vs.*	Exact test	0.4446	**<0.001**	**<0.001**	**0.0018**	0.2099	**<0.001**
	Romania* *vs.*	Exact test	0.2493	**<0.001**	0.9294	0.8546	0.6545	0.9731
D2S1338	**Romania *vs.***	Exact test	n/a		0.1111	0.6535	0.6936	0.2490
	Romania* *vs.*	Exact test	n/a		0.9997	0.9917	0.6191	0.9972
D16S539	**Romania *vs.***	Fst	0.4199	0.3936	0.8164	0.4346	0.9922	**0.0147**
	Romania* vs.	Fst	0.3320	0.0850	0.7129	0.2637	0.9815	0.1025
vWA	**Romania *vs.***	Exact test	0.7692	**<0.001**	0.8692	0.7465	0.9472	0.5346
	Romania* *vs.*	Exact test	0.6506	**<0.001**	0.9803	0.6449	0.9426	0.7279
	FBI *vs.*	Exact test		**0.0351**	0.6944	0.8750	0.8595	0.4253
D5S818	**Romania *vs.***	Fst	**<0.001**	0.8936	0.4102	0.5361	0.7305	0.8848
	Romania* *vs.*	Fst	**0.0029**	0.8057	0.6309	0.3447	0.8711	0.9297
	Romania *vs.*	Exact test	**0.0106**	0.9963	0.9703	0.6677	0.7798	0.9946
	Romania* *vs.*	Exact test	**0.0037**	0.9243	0.9943	0.7005	0.7138	0.9930
	FBI *vs.*	Fst		**<0.001**	**0.0029**	**0.0068**	**0.0029**	**0.0020**
		Exact test		**0.0164**	**0.0017**	**0.0115**	**0.0157**	**0.0164**
FGA	**FBI *vs.***	Exact test		0.1933	**0.0275**	0.2079	0.1219	0.1620

Table 3.17: Pairwise population comparison test (Fst) and exact population differentiation test (Exact test) results for the Romanian Caucasian population

Part IV: Discussions

The average expected heterosygosity (Hexp) value of CJ-Hu was lower than the observed heteorsygosity (Hexp=0.7857; Hobs=0.7941). The contrary was observed in the CV-Sze population (Hexp=0.7965; Hobs=0.7890). If we consider this value as a signal for potential inbreeding in the case of the CV-Sze population, we may give this as cause for why CV-Sze shows more significant differences than CJ-Hu, when compared to several Romanian populations.

If we consider that there should be a positive correlation between genetic and geographical distances, unexpected comparison results were obtained when CJ-Hu and CV-Sze allele frequencies were compared to each other, and when both populations were compared to HR-Sze, B-Ro, Wallachia and Dobruja. We expected to find more significant differences for the CV-Sze population, when compared to CJ-Hu and B-Ro, than to HR-Sze. Our results do not show any significant differences for CV-Sze versus CJ-Hu and B-Ro, but when compared to the nearby HR-Sze population, 1 STR loci reflected genetic differences. While the results of this may be influenced by the number of loci compared (only 10 loci were available for B-Ro population), future investigations including all 15 loci for B-Ro and further population comparison tests involving other historical Hungarian populations (from Hungary, Serbia, Slovakia, Ukraine and Austria) must be done in order to clarify their genetic characteristics.

As what refers to the comparison of the databases of Romania, we expected to have a negative correlation between the number of individuals per region and population distance. We expected Transylvania (n=3097; 44.9%) and Wallachia (n=1910; 27.7%) to show less significant differences than Dobruja (n=569; 8.25%), when compared to the entire Romanian population. Given the contrary results, we performed additional comparison tests with a modified Romanian database (Hungarian populations were excluded). As a result, the number of significantly different loci decreased from three to zero versus Wallachia and from two to zero versus Moldavia. The change raised the number of

significantly different loci from three to four versus Trs-6, but it did not significantly influence the match to Dobruja.

Although no tests have been performed to explain the given effect of the Transylvanian Hungarian communities upon the Romanian DNA database, this study indicates that ethnical origin should also be taken into consideration when comprehensive DNA databases are created.

Part V: Conclusions

The tested populations are suitable for use in human identification and micro-population differentiation studies because:

- the PDcomb (0.99999999999999992 - 0.999999999999999997) and PEcomb (0.9999998 - 0.999998) values obtained for the 15 loci are significantly high;
- the average heterozygosity level for both observed and expected values were above 78% and the required level of 70% was reached
- with the exception of Covasna's CSF1PO locus, all tested loci were in agreement with HWE

No significant differences were found between non-Szekler Hungarians from Cluj county (CJ-Hu) and Szeklers of Covasna county (CV-Sze).

Both populations (CJ-Hu and CV-Sze) show one significant difference when compared to Szeklers from Harghita county (HR-Sze).

None of these two populations (CJ-Hu and CV-Sze) show departures from Romanians living Bucharest (B-Ro); Statistical comparison for 5 loci was not determined.

Significant values involving 4 loci were found for CV-Sze when matched to Romanians from region of Wallachia

Population comparison test defined D21S11, D13S317 and D3S1358, D5S818 loci being the most polymorphic autosomal STR markers for the Romanian regions, when matched to the entire Romanian database and the FBI databases.

This study also indicates that ethnical origin should also be taken into consideration when comprehensive DNA databases are created.

Bibliography

(SELECTIVE)

AmpFLSTR Identifiler PCR Amplification Kit User's Manual. Protocol for 310 Genetic Analyzer. Applied Biosystems. 2006.

Anghel A, Marian C, Pitulescu M, Daba A, Sirbu IO, Rusu V, Budowle B. Population genetic study of eight short tandem repeat loci CSF1PO, TPOX, TH01, F13A01, FESFPS, vWA, F13B and LPL in the Western Romanian population. Forensic Sci Int. 2003;131:218-9.

Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Lincoln P, Mayr W, Olaisen B, DNA recommendations. Further report of the DNA Commission of the ISFG regarding the use of short tandem repeat systems; Forensic Sci Int. 1997; 87:179-4.

Bär W, Brinkmann B, Lincoln P, Mayr W, Rossi U, Budowle B, Eisenberg A, Fourney R, Gill P, Rand S, Editorial: Recommendations of the DNA Commission of the International Society for Forensic Haemogenetics relating to the use of PCR-based polymorphisms; Forensic Sci. Int. 1992; 55:1-3.

Barbarii LE, Burkhard R, Constantinescu C, Hohoff C, Petre C, Dermengiu D, Allele frequencies of 13 short tandem repeat (STR) loci in the Romanian population; Forensic Sci. Int. 2004; 141:171–4.

Bradford L. Disaster victim investigation recommendations from two simulated mass disaster scenarios utilized for user acceptance testing CODIS 6.0. Forensic Sci Int Genetics. 2010. doi:10.1016/j.fsigen.2010.05.005.

Brinkmann B, Klintschar M, Neuhuber F, Hoehne J, Burkhard R, Mutation Rate in Human Microsatellites: Influence of the Structure and Length of the Tandem Repeat; Am. J. Hum. Genet. 1998; 62:1408-15.

Budowle B, Allen RC, Analysis of amplified fragment-length polymorphisms (VNTR/STR loci) for human identity testing; Methods Mol. Biol. 1998; 98:155-71.

Budowle B, Bieber F, Eisenberg A. Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. Leg. Med. 2005;7:230-43.

Budowle B, Moretti TR, Keys KM, Koons BW, Smerick JB. Validation studies of the CTT STR multiplex system. J Forensic Sci. 1998;42(4):701-7.

Butler JM. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. Second Ed. Academic Press. 2005.

Butler JM. New resources for the forensic genetics community available on the NIST STRBase website. Forensic Sci Int Genetics Supplement Series. 2008;1:97-9.

Butler, John M., and Dennis J. Reeder. NIST standard reference database SRD 130: Chromosomal Locations for DNA Typing Markers. http://www.cstl.nist.gov/biotech/strbase/chrom.htm (accessed: 05 July 2010).

Carracedo A, John M. Butler, Gusma L, Parson W, Roewer L, Schneider PM, Editorial/Publication of population data for forensic purposes; Forensic Sci. Int. Genetics 2010; 4:145–147.

Chakraborty R, Sample size requirements for addressing the population genetic issues of forensic use of DNA typing; Human Biology 1992; 64:141-59;

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B. The utility of STR loci beyond human identification: Implications for the development of new DNA typing systems. Electrophoresis. 1999;20:1682-96.

Council of The European Union, No.doc: 15870/09 ENFOPOL 287 CRIMORG 170.

Egyed B, Füredi S, Pádár Z, Population genetic study in two Transylvanian populations using forensically informative autosomal and Y-chromosomal STR markers, Forensic Sci. Int. 2005; November 25, PMID: 16314060 [Epub ahead of print].

Egyed B, Füredi S, Pádár Z. Population genetic study in two Transylvanian populations using forensically informative autosomal and Y-chromosomal STR markers. Forensic Sci Int. 2005;PMID:16314060.

Excoffier L, Laval G, Schneider S, Arlequin ver. 3.5: an integrated software package for population genetics data analysis, Evolutionary Bioinformatics Online 2005; 47–50; http://cmpg.unibe.ch/software/arlequin35/Arl35Downloads.html

EU Council Resolution on the exchange of DNA analysis results; Official Journal of the European Union. 2009;C296.

Gill P, Jeffreys AJ, Werett DJ. Forensic application of DNA "fingerprints". Nature. 1985;318:577-79.

Holt CL, Buoncristiani M, Wallin JM, Nguyen T, Lazaruk KD, Walsh PS. TWGDAM validation of AmpFlSTR PCR amplification kits for forensic DNA casework. J Forensic Sci. 2002;47(1):66-96.

Jakovski Z, Ksenija N, Biljana J, Zdravko C, Aleksandar S, Verica P, Goran P, Aleksej D. Forensic DNA analysis in the identification of human remains in mass graves. J Clin Path & Forensic Med. 2010;1(1):001-4.

Jin L, Zhong Y, Chakraborty R. The exact numbers of possible microsatellite motifs. American Journal of Human Genetics. 1994;55(11):582-83.

Krenke BE, Tereba A, Anderson SJ, Buel E, Culhane S, Finis CJ. Validation of a 16-locus fluorescent multiplex system. J Forensic Sci. 2002;47(4):773-85.

Leclair B. Large-scale comparative genotyping and kinship analysis: evolution in its use for human identification in mass fatality incidents and missing persons databasing. International Congress Series. 2004;261:42-4.

Marian C, Anghel A, Bel SM, Ferencz BK, Ursoniu S, Dressler M, Popescu O, Budowle B. STR data for the 15 AmpFlSTR identifiler loci in the Western Romanian population. Forensic Sci Int. 2007;170:73-5.

NanoDrop Spectrophotometers Technical Bulletin T042. Thermo Scientific.

Raymond M, Rousset F, GENEPOP (v1.2): Population genetics software for exact tests and ecumenicism; J. Heredity, 1995; 86:248-9.

Romanian census 2002 official site: http://www.recensamant.ro

Rousset F, Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux; Mol. Ecol. Resources 2008; 8:103-106.

Sprecher CJ, Puers C, Lins AM, Schumm JW, General approach to analysis of polymorphic short tandem repeat loci, BioTechniques 1996; 20:266-76.

Stanciu F, (A) Popescu OR, Stoian IM, Allele frequencies of 15 STR loci in Moldavia region (NE Romania); Forensic Sci Int.(Genetics) 2009; 4:39–40.

Stanciu F, (B) Stoian IM, Popescu OR, Comprehensive STR data for the AmpFlSTR Identifiler from Transylvania (NW Romania); Legal Medicine 2009; 11:48–50.

Stanciu F, (C) Stoian IM, Popescu OR, Population Data for 15 Short Tandem Repeat Loci from Wallachia Region, South Romania; Croat Med J. 2009; 50:321-5.

Stanciu F, (D) Stoian IM, Popescu OR, STR data for the AmpFlSTR Identifiler from Dobruja region (SE Romania); Forensic Sci Int.(Genetics) 2009; 3:146–7.

SV Wizard Genomic DNA Purification Kit (Promega), 1999; Technical Manual No.050:4-5.

Tautz D. Hypervariability of simple sequences as a general source of polymorphic DNA markers. Nucleic Acids Res. 1989;17:6463-71.

Transylvania definition is available on online wersion of Oxford Dictionary:

Tsukada K, Takayanagi K, Asamura H, Ota M, Fukushima H. Multiplex short tandem repeat typing in degraded samples using newly designed primers for the TH01, TPOX, CSF1PO, and vWA loci. Legal Med. 2002;4:239-45.

Wiegand P, Kleiber M. Less is more (length reduction of STR amplicons using redesigned primers). Int J Legal Med. 2001;114(4–5):285-7.

Wilfinger WV, Mackey K, Chomczynski P, Effect of pH and Ionic Strength on the Spectrophotometric Assessment of Nucleic Acid Purity. BioTechniques. 1997;22:474-81.

Zhivotowsky LA, Feldman WM, Microsatellite variability and genetic distances; Proc. Natl. Acad. Sci. 1995; 92:11549-11552.