VILNIUS UNIVERSITY

INGRIDA UKTVERYTĖ

ANALYSIS OF GENETIC STRUCTURE OF LITHUANIAN ETHNO-LINGUISTIC GROUPS USING INFORMATIVE GENOMIC MARKERS

Summary of doctoral dissertation

Biomedical Sciences, Medicine (06 B)

Vilnius, 2014

Doctoral dissertation was prepared at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University in 2010–2014.

Principal supervisor:

Prof. Habil. Dr. Vaidutis Kučinskas (Vilnius University, biomedical sciences, medicine – 06 B)

The dissertation will be defended at Vilnius University, Academic Board of Examiners in Medicine:

Chairperson:

Prof. Dr. Algirdas Utkus (Vilnius University, biomedical sciences, medicine − 06 B)

Members:

Prof. Dr. Rimantas Jankauskas (Vilnius University, biomedical sciences, medicine − 06 B)

Prof. Habil. Dr. Limas Kupčinskas (Lithuanian University of Health Sciences, biomedical sciences, medicine − 06 B)

Prof. Dr. Andres Metspalu (Tartu University, biomedical sciences, biology − 01 B)

Prof. Dr. Gražina Slapšytė (Vilnius University, biomedical sciences, biology − 01 B)

The defence of the dissertation will be public at the meeting of the Academic Board of Examiners in Medicine on the 21st of November at 14:30 p.m. in the 1st Auditorium of the VUH Santariškių Klinikos (Audiences housing).

Address: Santariškių Str. 2, LT−08661 Vilnius, Lithuania.

The summary of the doctoral dissertation was distributed in 21st October, 2014.

A copy of the doctoral dissertation is available for review at the Library of Vilnius University and in Vilnius University website: http://www.vu.lt/lt/naujienos/ivykiu-kalendorius.

VILNIAUS UNIVERSITETAS

INGRIDA UKTVERYTĖ

LIETUVOS ETNOLINGVISTINIŲ GRUPIŲ GENETINĖS STRUKTŪROS ANALIZĖ REMIANTIS INFORMATYVIAIS GENOMO ŽYMENIMIS

Daktaro disertacijos santrauka

Biomedicinos mokslai, Medicina (06 B)

Vilnius, 2014 metai

Disertacija rengta 2010–2014 metais Vilniaus universiteto Medicinos fakulteto Žmogaus ir medicininės genetikos katedroje.

Mokslinis vadovas:

Prof. habil. dr. Vaidutis Kučinskas (Vilniaus universitetas, biomedicinos mokslai, medicina − 06 B)

Disertacija ginama Vilniaus universiteto Medicinos mokslo krypties taryboje:

Pirmininkas:

Prof. dr. Algirdas Utkus (Vilniaus universitetas, biomedicinos mokslai, medicina − 06 B)

Nariai:

Prof. dr. Rimantas Jankauskas (Vilniaus universitetas, biomedicinos mokslai, medicina − 06 B)

Prof. habil. dr. Limas Kupčinskas (Lietuvos sveikatos mokslų universitetas, biomedicinos mokslai, medicina − 06 B)

Prof. dr. Andres Metspalu (Tartu universitetas, biomedicinos mokslai, biologija − 01 B)

Prof. dr. Gražina Slapšytė (Vilniaus universitetas, biomedicinos mokslai, biologija − 01 B)

Disertacija bus ginama viešame Medicinos mokslo krypties tarybos posėdyje, kuris vyks 2014 m. lapkričio 21 d. 14 val. 30 min. VšĮ Vilniaus universiteto ligoninės Santariškių klinikų pirmoje auditorijoje (Auditorijų korpusas).

Adresas: Santariškių g. 2, LT–08661 Vilnius, Lietuva.

Disertacijos santrauka išsiųsta 2014 m. spalio 21 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir http://www.vu.lt/lt/naujienos/ivykiu-kalendorius.

CONTENT

LIST OF ABBREVIATIONS

AMOVA – Analysis of Molecular Variance

ASW – African Ancestry in Southwest US

bp – base pair

CEU – Utah Residents with Northern and Western European Ancestry

CHB − Han Chinese in Beijing, China

CHD − Chinese in Metropolitan Denver, Colorado

CHS – Southern Han Chinese, China

CLM – Colombian in Medellin, Colombia

CV – common variant, MAF >5%

DNA – deoxyribonucleic acid

FIN – Finnish in Finland

GBR – British in England and Scotland

GIH – Gujarati Indians in Houston, Texas

HVR – hypervariable region

IBS – Iberian populations in Spain

YRI – Yoruba in Ibadan, Nigeria

JPT – Japanese in Tokyo, Japan

kb – kilobase

LD – linkage disequilibrium

LFV – low-frequency variant, MAF 0,5−5%

LWK – Luhya in Webuye, Kenya

MAF – minor allele frequency

Mb – megabase

MDS – multidimensional scaling

MKK – Maasai in Kinyawa, Kenya

mtDNA – mitochondrial DNA

MXL/MEX – Mexican Ancestry in Los Angeles, California

NJ – neighbor joining

np – nucleotide pair

NRY – non-recombining region of the Y chromosome nt – nucleotide

PAR – pseudoautosomal region

CI – confidence intervals

PC – principal component

PCA – principal component analysis

LGM – last glacial maximum

PUR − Puerto Rican in Puerto Rico

rCRS – Revised Cambridge Reference Sequence NC_012920.1, GI:251831106

RSRS – Reconstructed Sapiens Reference Sequence

RV – rare variant, MAF <0,5%

TMRCA – time to most recent common ancestor

TSI – Toscani in Italy

STRs – short tandem repeats

UEPs – unique event polymorphisms

SNPs – single nucleotide polymorphisms

VU MF DHMG – Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University

EA – East Aukštaitija

SA – South Aukštaitija

WA – West Aukštaitija

NŽ – North Žemaitija

SŽ – South Žemaitija

WŽ – West Žemaitija

A – Aukštaitija

Ž – Žemaitija

1. INTRODUCTION

Differences in genetic variation are seen among populations from different continents (Henn *et al.* 2010), among populations within the same continent (Nelis *et al.* 2009) and even within the same populations (Jakkula *et al.* 2008). The major reasons for intrapopulation and interpopulation genetic variation are differences in language, culture and geographical location. The highest correlation is seen in the geographic and genetic distances among populations. Restricted migration due to geographic obstacles, i.e. distances, can be one of the fundamental forces which form and modulates the genetic diversity of a population. Fundamental human population studies are based on using different genetic markers: (1) Y chromosome SNPs and STRs; (2) complete mtDNA; (3) genome-wide autosomal SNPs; (4) exome genetic variants. The minor extent of genetic diversity among or within populations can be determined using genome-wide genetic markers (high density SNP arrays) and applying appropriate quality parameters for generated data. Hidden genetic stratification within a population or genetic differences among populations can lead to the false-positive results and misinterpretations of the association studies. This issue becomes more important when small sample sizes are available for the study or the aim of the study is to identify genetic variants which only slightly increase the risk of common diseases or phenotypes. Identification of the genetic stratification of a population before performing an association study increases the number of successfully identified risk alleles.

Previous studies of the Lithuanian population have showed statistically significant differences in the frequencies of P1 and LW^b allele (P and LW blood group systems) between South Aukštaitija and the rest of the ethno-linguistic groups of Lithuania. Moreover, statistically significant differences between North Žemaitija and South Aukštaitija were identified based on the analysis of *Alu* TPA25 distribution (Kučinskas 2001). However, the genotyping of Y chromosome SNPs, STRs and the sequencing of mtDNA HVRI resulted in no significant differences among ethno-linguistic groups of Lithuania (Kasperavičiūtė *et al.* 2004).

Novelty. This study is based on a complete mtDNA sequencing, an increased number of genotyped Y chromosome SNPs as and STRs compared with the previous studies of Lithuanian population. Furthermore, for the first time genome-wide data and exome genetic variants are used to analyze the genetic structure of the Lithuanian population and the genetic distances among the other populations. Finally, for the first time a comparison and an evaluation of different genetic markers used to study the Lithuanian population were performed.

Relevance. The results of this study supplement previous studies and will serve as the background for the future genetic structure analysis of the Lithuanian population. Analysis of the Lithuanian population using genomic markers is important for the subsequent studies and can help to avoid the results and conclusions of such studies to be influenced by population genetic stratification. Genome-wide, exome data generated using next-generation technologies give the possibility to compare the genetic structure of the Lithuanian population with the recent study results of other populations. The analysis of comprehensive genomic data helped to detect minor genetic differences within the Lithuanian population and among the neighboring populations. Moreover, the evidences that the gene pool of the Lithuanian population was influenced by other populations were detected in this study. Finally, the generated next-generation sequencing and genotyping data encourage collaboration with other scientific groups.

Follow-on. The future analysis of different genome-wide markers of extended sample sets could approve or supplement the results of this study. The increased number of analyzed Y chromosome SNPs of two most frequent haplogroups (N1c1 and R1a1a) in the Lithuanian population together with the known historical events could help to solve the current puzzle of patrilineal migration. The first results of next-generation sequencing and genome-wide data not only helped to gain knowledge but also raised more questions. The increased number of samples and the usage of improved analysis methods could help to solve some of these questions. Implementation of next-generation sequencing and genome-wide genotyping methods at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University can be the background for the future Lithuanian genome database establishment.

Aim of the study

To evaluate the genetic structure of the Lithuanian population and to compare it with that of the neighboring populations using informative genomic markers.

Main tasks of the research study

(1) To determine and evaluate intrapopulation genetic structure and interpopulation genetic distances of the Lithuanian population using Y chromosome genetic markers.

(2) To analyze the genetic structure of the Lithuanian population and to evaluate the genetic distances among the neighbor populations using complete mtDNA sequences.

(3) To determine and evaluate genome-wide patterns of genetic variation within the Lithuanian population and to compare them with those of the populations of different origin.

(4) To evaluate the distribution of exome variation in the Lithuanian population.

(5) To analyze, evaluate, and compare the different genomic markers used in this study.

Statements to be defended

(1) The Lithuanian population is homogeneous as the differences among groups based on the geographic location of ethno-linguistic groups on the territory of Lithuania account for <2% of all genetic variation.

(2) The distribution of Y chromosome haplogroups and haplotypes analysis showed the North–South gradient while the study of complete mtDNA revealed the West–East gradient of genetic variation within the Lithuanian population.

(3) The study of genome-wide data showed the Northwest–Southeast gradient of genetic variation within the Lithuanian population.

(4) Genetic distances among populations depend mostly on the geographic distances among the populations studied.

(5) The distribution of different genomic markers within the Lithuanian population is unequal.

2. SUBJECTS AND METHODS

2.1. Subjects

Unrelated individuals who indicated at least three generations of Lithuanian nationality formed the sample set of the general Lithuanian population. The general Lithuanian population group was divided into two main groups Aukštaitija and Žemaitija – based on dialect. Each of the two main groups consists of three subpopulations based on the subdialect (six ethno-linguistic groups of Lithuania): South, West, East Aukštaitija and North, South and West Žemaitija (Girdenis, Zinkevičius 1966) (Fig. 2.1, Fig. 2.2).

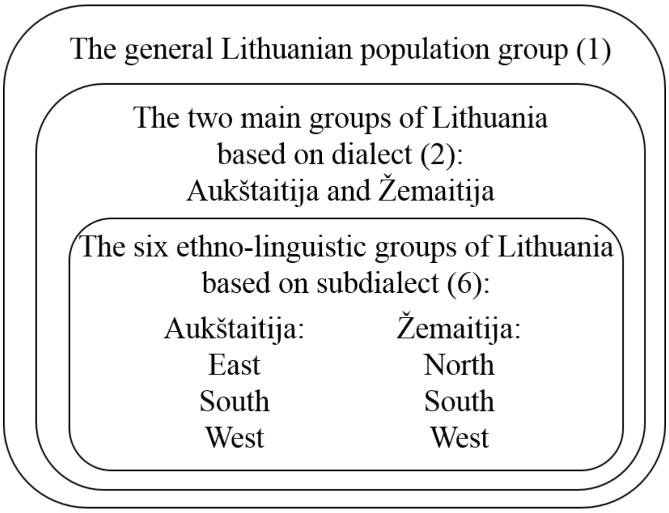

The general Lithuanian population group (1)
The two main groups of Lithuania based on dialect (2): Aukštaitija and Žemaitija
The six ethno-linguistic groups of Lithuania based on subdialect (6):
Aukštaitija:
East
South
West
Žemaitija:
North
South
West

Fig. 2.1. Groups of samples in this study.

Venous blood samples were collected in 1994 – 1995 and during the LITGEN project (VP1-3.1-ŠMM-07-K-01-013) in 2011– 2013. Genomic DNA was extracted by specialists of the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University.

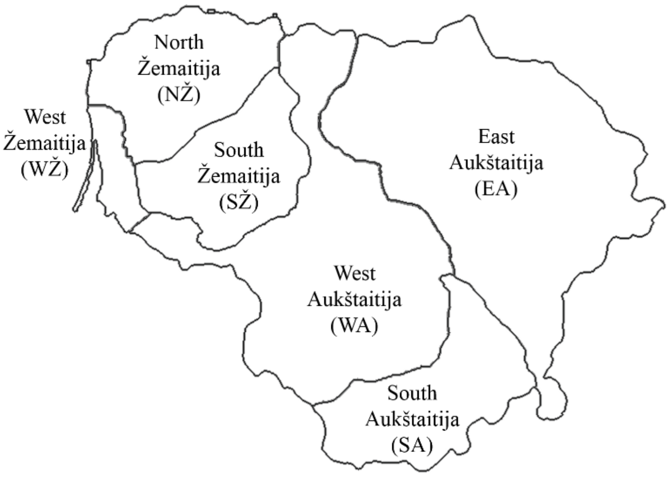


Fig. 2.2. Territory of six ethno-linguistic groups of Lithuania.

Sample sets and sample sizes of this study are shown in Fig. 2.3, Fig. 2.4, Fig. 2.5, Fig. 2.6.

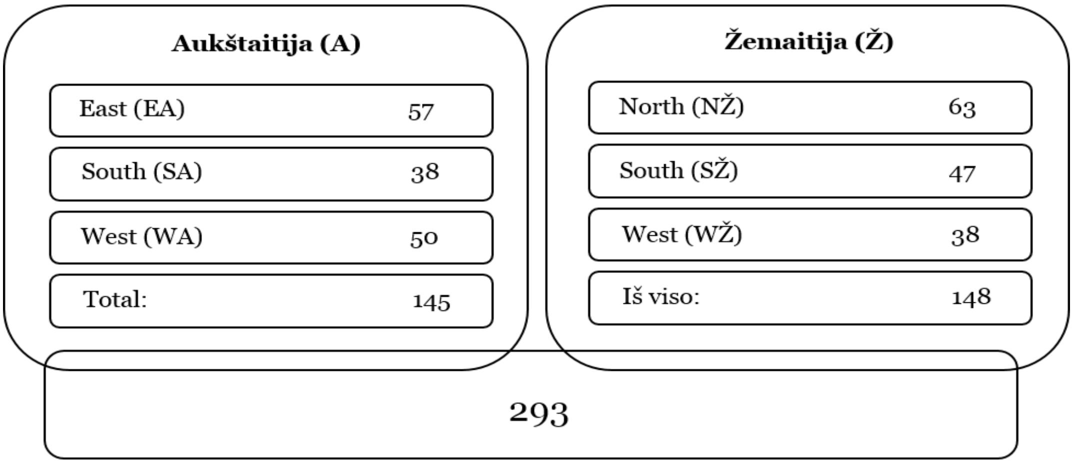


Fig. 2.3. Distribution of sample sets included in the Y chromosome SNPs and STRs analyses.

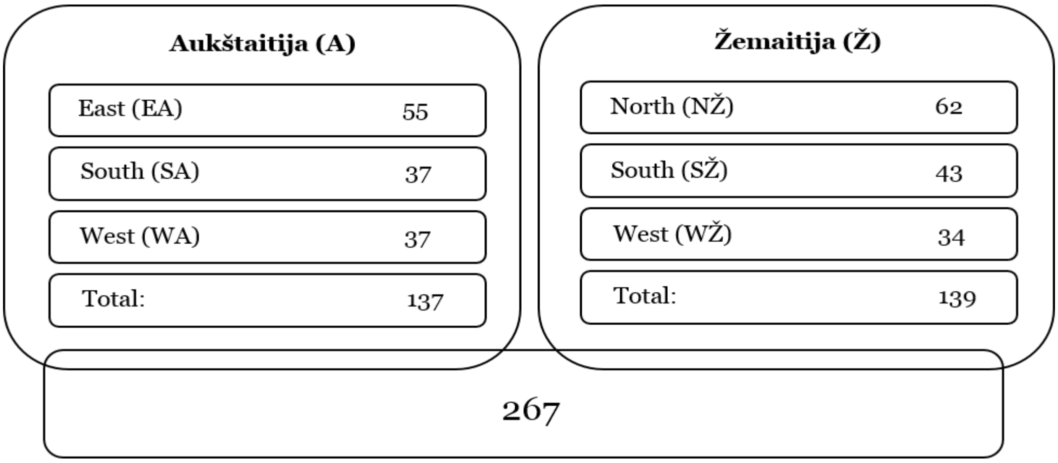


Fig. 2.4. Distribution of sample sets included in the complete mtDNA analyses.

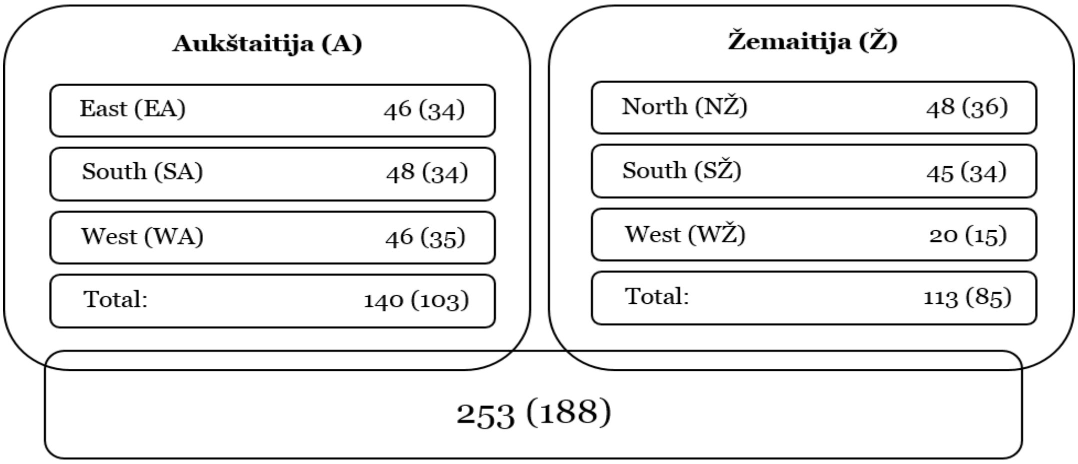
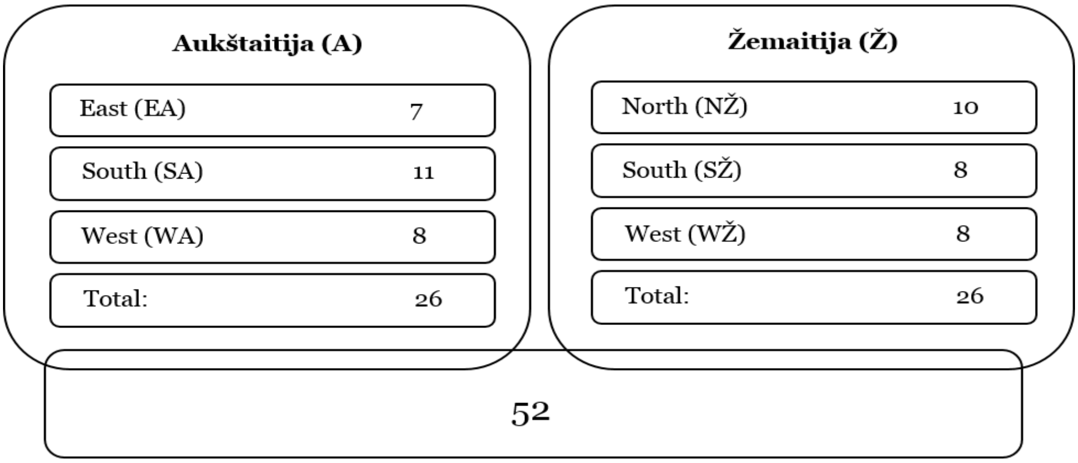


Fig. 2.5. Distribution of sample sets included in the genome-wide (719,666 SNPs) data analyses. Number in the brackets indicates the distribution of sample sets included in the genome-wide data analyses after performing quality control.



Aukštaitija (A)	
East (EA)	7
South (SA)	11
West (WA)	8
Total:	26

Žemaitija (Ž)	
North (NŽ)	10
South (SŽ)	8
West (WŽ)	8
Total:	26

52

Fig. 2.6. Distribution of sample sets included in the exome variant analyses.

The approval to conduct genetic and genomic research projects was provided by the Vilnius Regional Research Ethics Committee: (1) “Genetic diversity of the population of Lithuania and changes of its genetic structure related with evolution and common diseases”, acronym - “LITGEN“, No. 158200-05-329-79, date: 2011-05-03; (2) the local approval to conduct genetic researches at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University.

2.2. Methods

Genomic DNA was extracted from blood using the phenol–chloroform extraction method or automated nucleic acid purification using paramagnetic particles (Freedom EVO® Nucleic Acid Purification Workstation). The quality and quantity of purified genomic DNA was evaluated with a spectrophotometer. The Y chromosome SNPs genotyping was performed using the TaqMan® probe assays. The commercial kit Applied Biosystems® AmpFlSTR® Yfiler™ was used and the fragment length analysis by ABI PRISM 3130xl Genetic Analyzer was performed to genotype Y chromosome STRs. The Y chromosome analysis was performed by the author at the Population Genetics Laboratory of the Research Centre for Medical Genetics of the Russian Academy of Medical Sciences (Moscow, Russian Federation) under the supervision of PhD O. Balanovsky and prof. E.V. Balanovska. Multiplex sequencing on the *Illumina GAII* platform after in-solution capture enrichment was used to obtain complete mtDNA genome sequences with an average of 352-fold coverage depth. The complete mtDNA analysis was performed by the author at the Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany) under the supervision of Prof. M. Stoneking. Genome-wide genotyping was performed using high density *Illumina HumanOmniExpress-12v1.1* arrays (719,666 SNPs) at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University in collaboration with colleagues. Next-generation exome sequencing after in-solution capture enrichment with an average of a 40-fold coverage depth was performed using *5500 SOLiD™ Sequencer* at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University in collaboration with colleagues. Detailed protocols of molecular genetic methods used in this study can be found in "1 PRIEDAS. TYRIMO METODAI" (in Lithuanian).

2.3. Data analyses

Haplogroups of the Y chromosome or mtDNA were pooled if it was necessary for further analyses. Haplogroups of the Y chromosome or the mtDNA phylogenetic tree were grouped in the hierarchical manner, e.g. Y chromosome haplogroups G2a and G2ab3 were pooled into one haplogroup named G2 or G2a. The frequency of a new haplogroup was a sum of the pooled haplogroups (G2a and G2ab3) frequencies.

The Y chromosome STRs DYS385a and DYS385b were removed from further analyses as there was one pair of primers in the commercial kit Applied Biosystems® AmpFlSTR® Yfiler™ used to amplify DYS385a and DYS385b genomic markers in the duplicated genomic region. The genetic markers DYS389I and DYS389b were used for the further analyses instead of the genetic markers DYS389I and DYS389II which were identified using the same commercial kit. The genetic marker DYS389II is a compound marker which includes the DYS389I and DYS389b markers.

2.3.1. Analyses of Y chromosome haplogroups and haplotypes

The primary data analysis was performed by specialists of the Population Genetics Laboratory of the Research Centre for Medical Genetics of the Russian Academy of Medical Sciences (Moscow, Russian Federation).

Table 2.1. List of software used for the primary and secondary analyses of Y chromosome haplogroups and haplotypes

Data analysis	**Software**	**Description**
Primary	***GeneMapper® Software v4.0*** (*GMSv4.0*)	STRs fragment length analysis
Secondary	***Haplogroup Predictor***[1] (Athey 2005)	Haplogroup prediction
	Arlequin v3.5.1.2[1] (Excoffier, Lischer 2010)	Haplogroup and haplotype molecular diversity indices
		F_{ST} distance matrices based on haplogroup distribution and R_{ST} distance matrices based on haplotype distribution (Belle *et al.* 2010)
		Analysis of Molecular Variance (AMOVA)
	Phylip v3.69[1] (Felsenstein 2004)	Nei (D) genetic distance (Nei 1972) (Zoossmann-Diskin 2010)
		Unweighted Pair Group Method with Arithmetic Mean (UPGMA) or Neighbor Joining (NJ) clustering
	SPSS Statistics Desktop v22.0.0 (desktop version)[1]	Multidimensional scaling (MDS)
	Microsoft® Office Excel	Graphics, diagrams
		Data analysis and management
	R v3.0.3 **(*Ade4, Rcmdr*)**[1]	Correlation of two distance matrices (Mantel test)
		Principal component analysis (PCA)
	Barrier v2.2[1] (Manni *et al.* 2004)	Barriers of genetic variation
	Network v4.6.1.2[1] (www.fluxus-engineering.com)	Phylogenetic tree using haplotype data
	BATWING[1] (Wilson *et al.* 2003)	Time to Most Recent Common Ancestor (TMRCA)

[1] Freely available.

2.3.2. Analyses of mtDNR haplogroups and haplotypes

The primary data analysis was performed by specialists of the Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology (Leipzig, Germany): (1) mtDNA sequence mapping; (2) alignment and reconstruction of complete mtDNA sequence with allowance of two mismatches and one gap; (3) quality control: position with coverage depth of one and position with minor allele frequency (MAF) >30% were tagged as N, i.e. not detected.

Table 2.2. List of software used for the primary and secondary analyses of complete mtDNA

Data analysis	Software	Description
Primary	***Haplogrep***[1] (Kloss-Brandstatter *et al.* 2011).	Haplogroup prediction
Secondary	***Muscle***[1] (Edgar 2004)	Multiple alignment
	BioEdit v7.1.3.0[1] (Hall 1999)	DNA sequence edit
	DnaSP v5.10.01 (Librado, Rozas 2009)	DNA sequence variant analysis
	Arlequin v3.5.1.2[1] (Excoffier, Lischer 2010)	Haplogroup and haplotype molecular diversity indices, Tajima D
		F_{ST} distance matrices based on haplogroup distribution and Φ_{ST} distance matrices based on haplotype distribution (Barbieri *et al.* 2012; Kasperavičiūtė *et al.* 2004)
		AMOVA
	SPSS Statistics Desktop v22.0.0 (Desktop version)[1]	MDS
	Microsoft® Office Excel	Graphics, diagrams
		Data analysis and management
	R v3.0.3 **(*Ade4, Rcmdr*)**[1]	Correlation of two distance matrices (Mantel test)
		PCA
	Barrier v2.2[1] (Manni *et al.* 2004)	Barriers of genetic variation
	Network v4.6.1.2[1] (www.fluxus-engineering.com)	Phylogenetic tree based on mtDNA sequence (577−16 023 np)
		ρ statistic based on mtDNA sequence (16 564 np)
	Age calculator (Costa *et al.* 2013)	Age estimate on mtDNA sequence (16 564 np)
	Python v2.7.4	Data analysis and management using inhouse script[2]

[1] Freely available.

[2] Performed by author in collaboration with a specialist.

2.3.3. Analyses of genome-wide autosomal data

The primary data analysis was performed by PhD student I. Domarkienė at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University.

The first 100 principal components (PC) were used to identify the principal components' ancestral informative markers (PCAIMS). The most variable 200 SNPs with the highest weighting coefficients were selected from each of the first 100 PC

(*EIGENSOFT v5.0.1*). The relative weighting coefficients were calculated for all unique SNPs. Contributions of each SNP to each PC were normalized to the highest weight so that the SNPs that contributed most to a PC was given a weight of 1. Weights of each SNP were summed across all PCs to get rank of all SNPs. The further analyses included all SNPs that had a relative weighting coefficient >1 (Huckins *et al.* 2014).

Table 2.3. List of software used for the primary and secondary analysis of genome-wide data

Data analysis	Software	Description
Primary	***GenomeStudio v2011.1*** (*Illumina® GenomeStudio 2011*, producer *Illumina, Inc. 2003−2011*)	Quality control (Annex1, Table 13, Table 14)
Secondary	***PLINK v1.9***[1] (Purcell *et al.* 2007)	Autosomal SNPs filtering
		Identical by Descent (IBD) test, remove PI HAT < 0.4
		Inbreeding coefficient (F)
		Hardy−Weinberg test, remove $p < 10^{-5}$
		SNPs with missing genotypes, remove >1%
		MAF test, remove < 0.01
		Linkage disequilibrium (LD) test, remove $r^2 < 0.2$
		MDS ($r^2 < 0.2$)
		LD
	EIGENSOFT v5.0.1[1] (Patterson *et al.* 2006)	PCA ($r^2 < 0.2$)
		PCA Tracy−Widom test
		Most variable SNPs detection
		F_{ST} distance matrices
	Microsoft® Office Excel	Graphics, diagrams
		Data analysis and management
		PCAIMS
	R v3.0.3 (Ade4, Rcmdr)[1]	Correlation of two distance matrices (Mantel test)
		Independent *t* test
	Barrier v2.2[1] (Manni *et al.* 2004)	Barriers of genetic variation
	Phylip v3.69[1] (Felsenstein 2004)	UPGMA or NJ clustering
	Structure v.2.3.4[1] (Pritchard *et al.* 2000)	Population structure
	Python v2.7.4[1]	Data analysis and management using inhouse script[2]

[1]Freely available.

[2]Performed by author in collaboration with a specialist.

2.3.4. Analyses of exome variants

The primary, secondary, tertiary analyses of exome data were performed by PhD L. Ambrozaityte in collaboration with the author at the Department of Human and Medical Genetics, Faculty of Medicine, Vilnius University.

Table 2.4. List of software used for the primary, secondary, tertiary and quarterly analyses of exome data

Data analysis	**Software**	**Description**
Primary **Secondary** **Tertiary**	***LifeScope™ Genomic Analysis Software v2.5***.	Parameters (Annex 1, Table 16)
Quarterly	***Arlequin v3.5.1.2***[1] (Excoffier, Lischer 2010)	Pairwise differences based on 6,939 autosomal SNPs
	Microsoft® Office Excel	Graphics, diagrams
		Data analysis and management
	***R v3.0.*3 (*Ade4, Rcmdr*)**[1]	Venn diagram
		Linear regression model
		PCA
	Python v2.7.4[1]	Data analysis and management using inhouse script[2]
	PostgreSQL v.9.2.3[1]	Relational database of exome annotated variants[2]
		SQL queries for the analysis of exome annotated variants[2]

[1] Freely available.

[2] Performed by author in collaboration with a specialist.

3. RESULTS

3.1. Genetic structure and genetic distances of Lithuanian population based on the distribution of Y chromosome haplogroups and haplotypes

Sixteen different Y chromosome haplogroups (Fig. 3.1) were identified in the studied Lithuanian population (Fig. 2.2). All identified haplogroups belonged to eight different Y chromosome phylogenetic lineages (E, G, H, I, J, N, R, T). Six phylogenetic lineages (E, G, I, J, N, R) which were present in more than 99% of the studied Lithuanian population are the most frequent Y chromosome phylogenetic lineages in European populations (Wiik 2008).

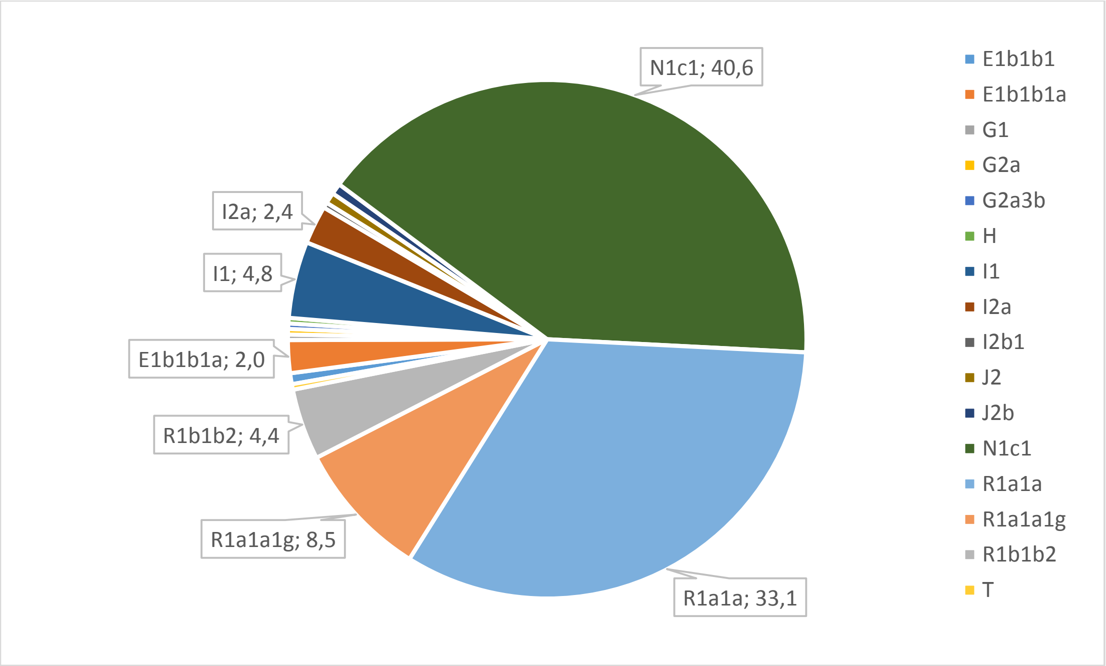


Fig. 3.1. Distribution of identified Y chromosome haplogroups (frequency,%) in the studied Lithuanian population.

The most frequent Y chromosome haplogroups in the studied Lithuanian population were N1c1(M178) – 40.6%, R1a1a(M198) – 33.1% which together with R1a1a1g(M458) comprised 41.6%. Three most frequent haplogroups comprised approximately ~82% and the remaining13 haplogroups ~18% of the studied Lithuanian population.

Three most frequent haplogroups showed the North−South gradient of distribution. Haplogroup N1c1 was more frequent in the North than in the South Lithuanian population (Fig. 3.2), whereas the haplogroups R1a1a and R1a1a1g were detected more frequently in the South and less frequently in the North Lithuanian population (Fig. 3.3, Fig. 3.4).

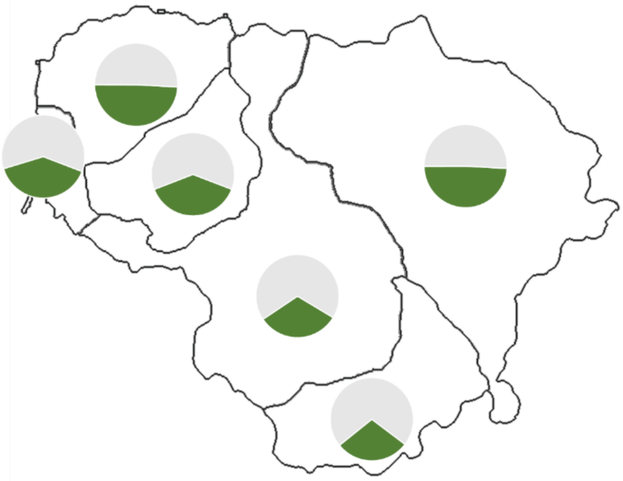


Fig. 3.2. Distribution of Y chromosome haplogroup N1c1 (green label) within six ethno-linguistic groups of Lithuania.

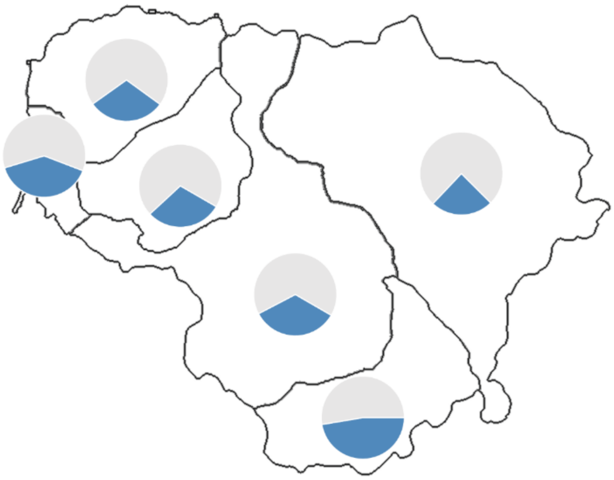


Fig. 3.3. Distribution of Y chromosome haplogroup R1a1a (blue label) within six ethno-linguistic groups of Lithuania.

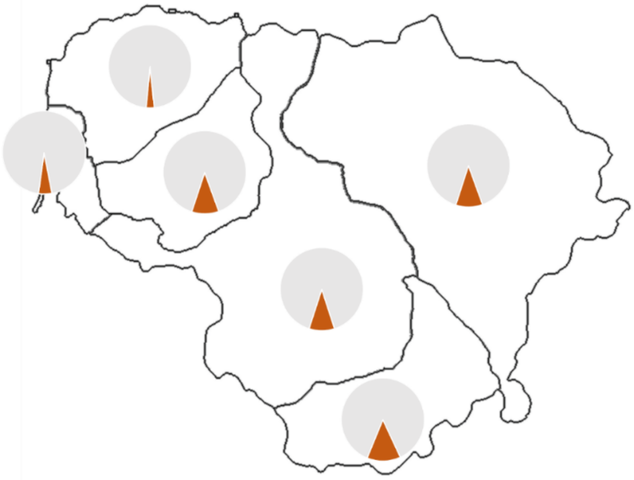


Fig. 3.4. Distribution of Y chromosome haplogroup R1a1a1g (red label) within six ethno-linguistic groups of Lithuania.

Three STRs (DYS458, DYS456, DYS635) showed the highest variation and one (DYS437) the lowest variation within the studied Lithuanian population (Fig. 3.5).

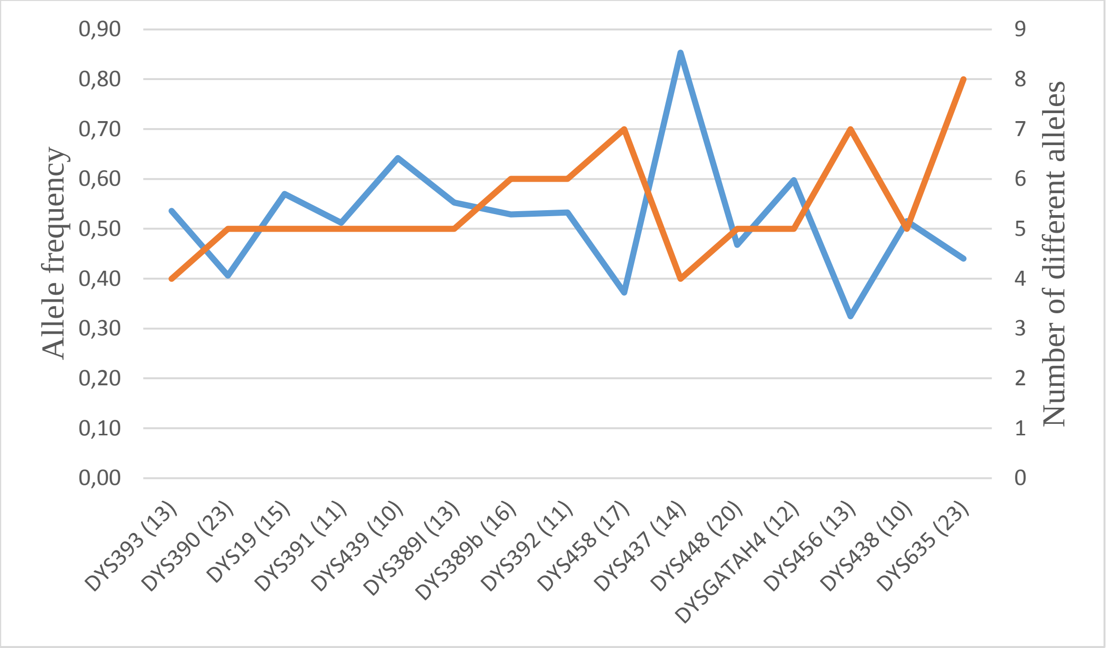


Fig. 3.5. Distribution of 15 Y chromosome STRs in the studied Lithuanian population. The number in brackets indicates the most frequent allele of a particular STR, and the frequency is shown by a blue line (axis on the left). The number of different alleles of a particular STR is marked by an orange line (axis on the right).

Seven different alleles with the most frequent allele (17) which comprised 37% of the studied population were identified when analyzing DYS458. Analysis of the DYS456 distribution showed seven different alleles with the most frequent allele (13) which comprised 32% of the studied population. Eight different alleles with the most frequent allele (23), which comprised 44% of studied population were identified analyzing DYS635. Analysis of DYS437 showed four different alleles with the most frequent allele (14) which comprised 85% of the studied population.

Totally, 250 different haplotypes were identified in the Lithuanian population. The most frequent haplotype was detected within the haplogroup N1c1 in five samples and comprised 1.7% of the Lithuanian population studied. Other haplotypes, which were detected in more than one sample, were identified within the haplogroup N1c1: two different haplotypes each of which found in four samples (1.3%), and five different haplotypes each of which found in three samples (1.0%). Each of 23 different haplotypes were found in two samples (0.7%). Finally, 219 different unique haplotypes were detected in the Lithuanian population studied.

Calculated molecular diversity indices are presented in Table 3.1.

Table 3.1. Molecular diversity indices of Y chromosome haplogroups and haplotypes

Population	N	Haplogroup diversity (SD)	Number of haplotypes	Haplotype diversity (SD)
S. Aukštaitija	38	0.6899 (±0.0537)	38	1.0000 (±0.0060)
E. Aukštaitija	57	0.6898 (±0.0480)	54	0.9981 (±0.0037)
W. Aukštaitija	50	0.7739 (±0.0370)	49	0.9992 (±0.0042)
Aukštaitija	145	0.7300 (±0.0232)	137	0.9991 (±0.0010)
N. Žemaitija	63	0.6667 (±0.0439)	59	0.9980 (±0.0033)
W. Žemaitija	38	0.6970 (±0.0488)	37	0.9986 (±0.0065)
S. Žemaitija	47	0.7595 (±0.0414)	40	0.9926 (±0.0062)
Žemaitija	148	0.7023 (±0.0258)	129	0.9980 (±0.0011)
Lithuania	293	0.7152 (±0.0173)	250	0.9986 (±0.0005)
Haplogroup	**N**	**Haplogroup diversity (SD)**	**Number of haplotypes**	**Haplotype diversity (SD)**
N1c1	119	–	85	0.9927 (±0.0023)
R1a1a	97	–	91	0.9987 (±0.0017)
R1a1a1g	25	–	24	0.9967 (±0.0125)

The mean number of pairwise differences was 8.8 (±4.1), and the average haplotype diversity over loci was 0.5893 (±0.3015) in the Lithuanian population.

Nei (D) and F_{ST} distances based on the Y chromosome haplogroup (E1b1b1 (+E1b1b1a), I1, I2a, N1c1, R1a1a, R1a1a1g, R1b1b2, others (+G1, G2a, G2a3b, H, I2b1, J2, J2b, T)) frequencies between six ethno-linguistic groups of Lithuania were calculated (Table 3.2).

Table 3.2. Nei (D) and F_{ST} distances between six ethno-linguistic groups of Lithuania. Above the diagonal Nei (D) distances, below F_{ST} distances. Statistically significant indices are in bold ($P < 0.05$; 10,000 permutation). Abbreviations: NŽ – North Žemaitija, SŽ – South Žemaitija, WŽ – West Žemaitija, WA – West Aukštaitija, SA – South Aukštaitija, EA – East Aukštaitija

	NŽ	SŽ	WŽ	WA	SA	EA
NŽ	0	0,003416	0,003673	0,005699	0,013052	0,002597
SŽ	−0,00224	0	0,002536	0,001633	0,006654	0,003187
WŽ	−0,00305	−0,01234	0	0,002300	0,004174	0,005575
WA	0,00875	−0,01391	−0,01282	0	0,003824	0,007274
SA	**0,04120**	0,00640	−0,00638	−0,00581	0	0,015460
EA	−0,00394	−0,00469	0,00471	0,01399	**0,04945**	0

The F_{ST} distances ($P < 0.05$; 10,000 permutations) were statistically significant between two pairs of groups: South Aukštaitija and East Aukštaitija, South Aukštaitija and North Žemaitija. Results of the Mantel test indicated a statistically significant correlation between Nei (D) and geographic distances (Annex 1, Table 2) (r = 0.47; $P = 0.0454$; 10,000 permutations).

Table 3.3. Results of AMOVA based on haplogroups and haplotypes distribution. Statistically significant indices are in bold (*P* < 0.05; 10,000 permutation). Grouping of ethno-linguistic groups of Lithuania was based on geographic locations in the territory of Lithuania: (1) North Žemaitija, East Aukštaitija; (2) West Žemaitija, West Aukštaitija, South Žemaitija; (3) South Aukštaitija

Haplogroup (F_{ST} distances)			
Source of variation	**df**[1]	**Fixation index (F)**	**Percentage of diversity (%)**
between groups	2	**0.0189** (*P* = 0.01683)	1.89
between populations within group	3	−0.00915[2]	−0.9
within population	287	0.00992[2]	99.01
Haplotype (R_{ST} distances)			
Source of variation	**df**[1]	**Fixation index (F)**	**Percentage of diversity (%)**
between groups	2	**0.02023** (*P* = 0.01653)	2.02
between populations within group	3	−0.00988[2]	−0.97
within population	287	0.01055[2]	98.95

[1] df – degrees of freedom.

[2] not significant (*P* > 0.05).

Classification of the ethno-linguistic groups of Lithuania for the AMOVA analysis (F_{ST} and R_{ST} distances) was based on geographical location, language and cultural differences. In all classification cases, almost all (~99%) variations fell among samples within the ethno-linguistic groups of Lithuania. Statistically significant AMOVA results were obtained when the groups were classified based on geographic location within the territory of Lithuania (*P* < 0.05; 10,000 permutations) (Table 3.3).

The MDS analysis based on F_{ST} distances was performed. In the scatter plot of MDS, four clusters were seen: (1) North Žemaitija, East Aukštaitija; (2) West Žemaitija, West Aukštaitija; (3) South Žemaitija; (4) South Aukštaitija. Moreover, the MDS analysis based on Nei (D) distances was performed. Visualized results of this analysis showed three clusters: (1) North Žemaitija, East Aukštaitija; (2) West Žemaitija, West Aukštaitija, South Žemaitija; (3) South Aukštaitija. According to the latter analysis, the territory of Lithuania could be divided into three regions: North, Middle, and South (Fig. 3.6).

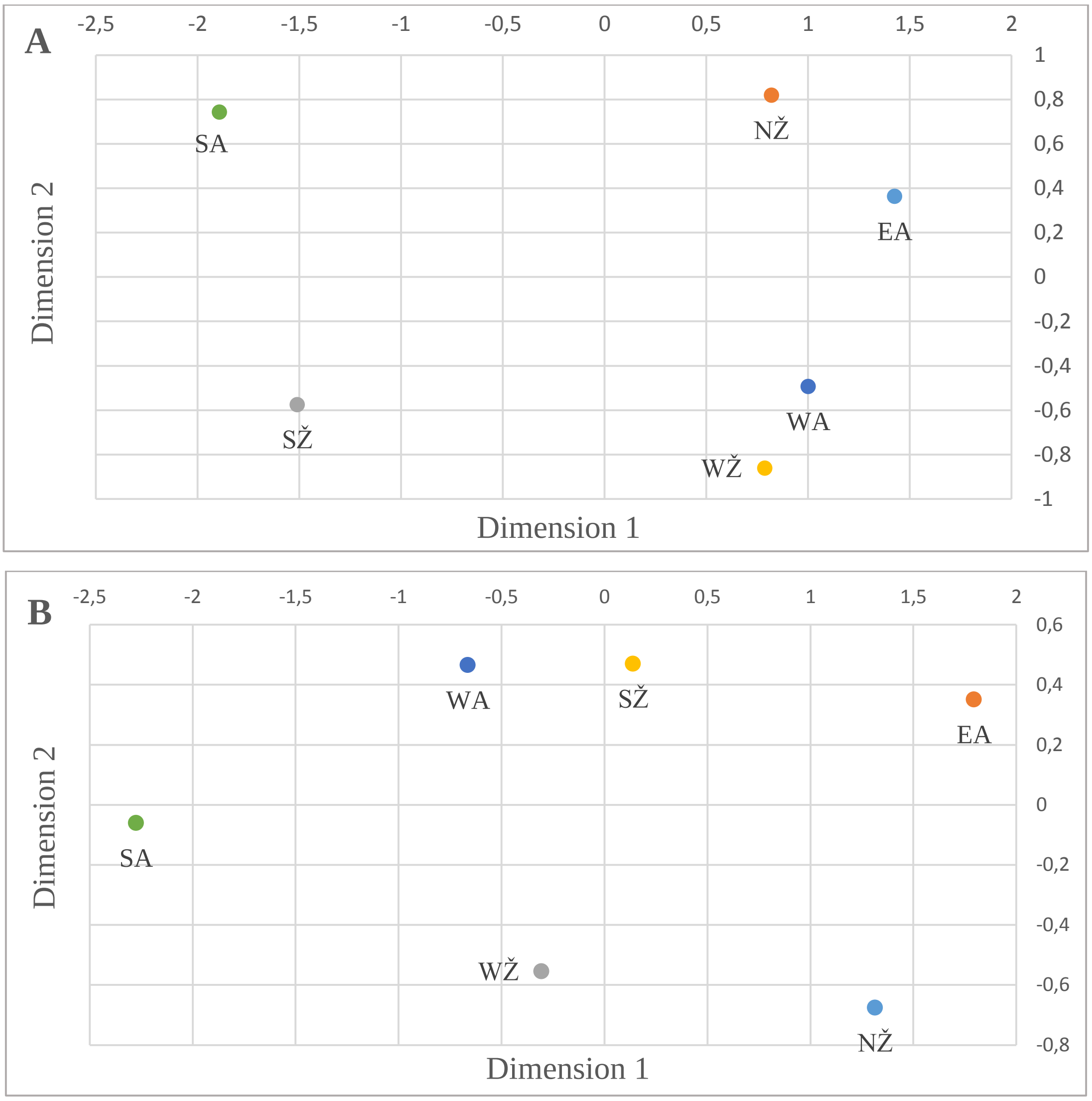


Fig. 3.6. Scatter plot of MDS analysis based on (A) F_{ST} and (B) Nei (D) distances (E1b1b1 (+E1b1b1a), I1, I2a, N1c1, R1a1a, R1a1a1g, R1b1b2, other (+G1, G2a, G2a3b, H, I2b1, J2, J2b, T)). Stress and correlation coefficient (r^2): (1) F_{ST} distance: *stress* = 0.0025; r^2 = 0.99996; (2) Nei (D) distance: *stress* = 0.0031; r^2 = 0.99994.

The PCA based on F_{ST} distances was performed. The first two PC explained 48.47% and 30.71% of variations between six ethno-linguistic groups of Lithuania. The PCA yielded a genetic map that shows three clusters (North–South axis): (1) North Žemaitija, East Aukštaitija; (2) South Žemaitija; (3) West Žemaitija, West, South Aukštaitija. The results of PCA based on Nei (D) genetic distances showed that the first two PCs explained 54.82% and 37.49% of variations among six ethno-linguistic groups of Lithuania (Fig. 3.7).

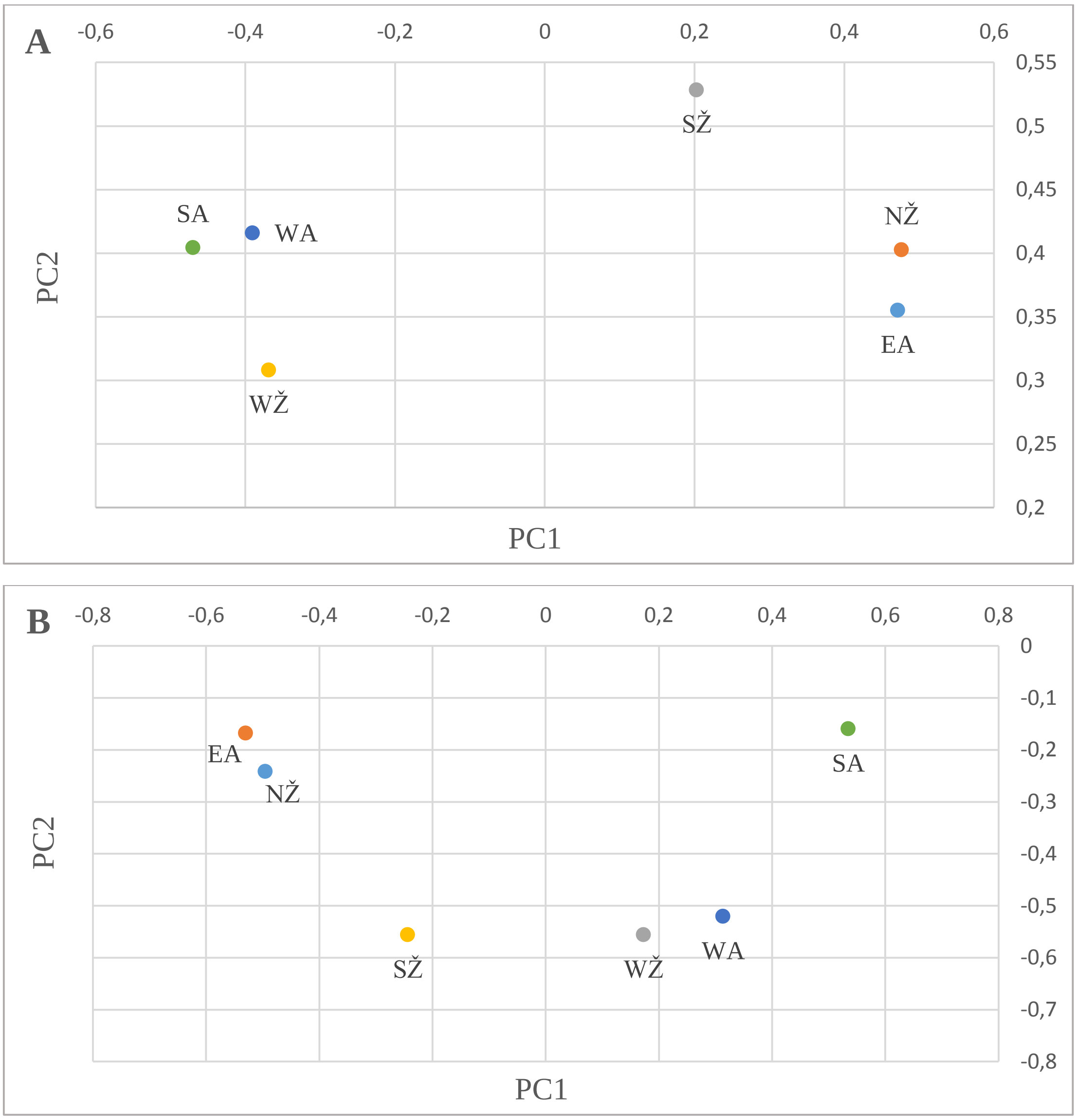


Fig. 3.7. Scatter plot of PCA based on (A) F_{ST} and (B) Nei (D) distances (E1b1b1 (+E1b1b1a), I1, I2a, N1c1, R1a1a, R1a1a1g, R1b1b2, other (+G1, G2a, G2a3b, H, I2b1, J2, J2b, T)).

The UPGMA clustering based on Nei (D) distances was performed. Three clusters of populations could be seen in the constructed UPGMA tree: (1) North Žemaitija, East Aukštaitija; (2) West Žemaitija, West Aukštaitija, South Žemaitija; (3) South Aukštaitija. The branch of the South Aukštaitija is the most farthest one from the other two clusters (Fig. 3.8).

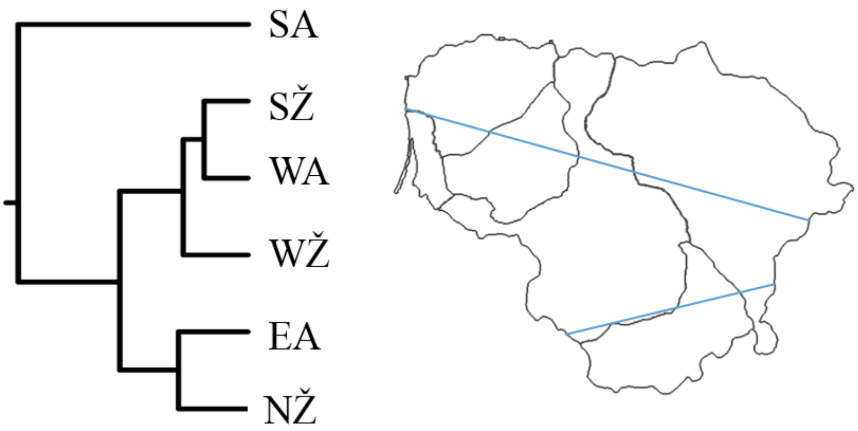


Fig. 3.8. The UPGMA tree based on Nei (D) distances. Right: lines on the map show the boundary of the clusters.

Barriers of genetic variation, i.e. Y chromosome haplogroups, among six ethno-linguistic groups of Lithuania were identified (Fig. 3.9).

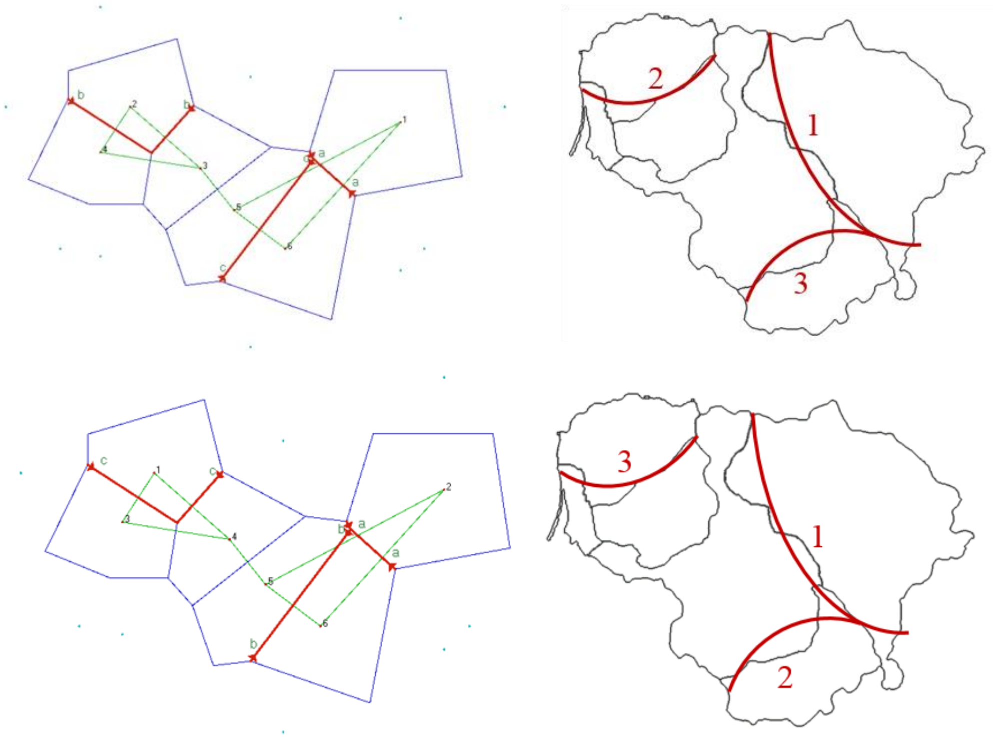


Fig. 3.9. Barriers of genetic variation, i.e. Y chromosome haplogroups, of the six ethno-linguistic groups of Lithuania. Left: output of the software, i.e. a schematic map as of the ethno-linguistic groups of Lithuania according to submitted coordinates (Annex 1,

Table 1). Right: barriers were numbered in the order of priority. Analysis was based on F_{ST} (top) and Nei (D) (bottom) distances.

Phylogenetic networks based on Y chromosome STRs (15) for the most frequent Y chromosome haplogroups (N1c1, R1a1a, R1a1a1g) were created. The phylogenetic network of the haplogroup N1c1 had a star-like structure with the most frequent haplotype in the center (1.7%) (Fig. 3.10). In contrast, the phylogenetic networks of the haplogroups R1a1a and R1a1a1g had no star-like structure (Fig. 3.11, Fig. 3.12).

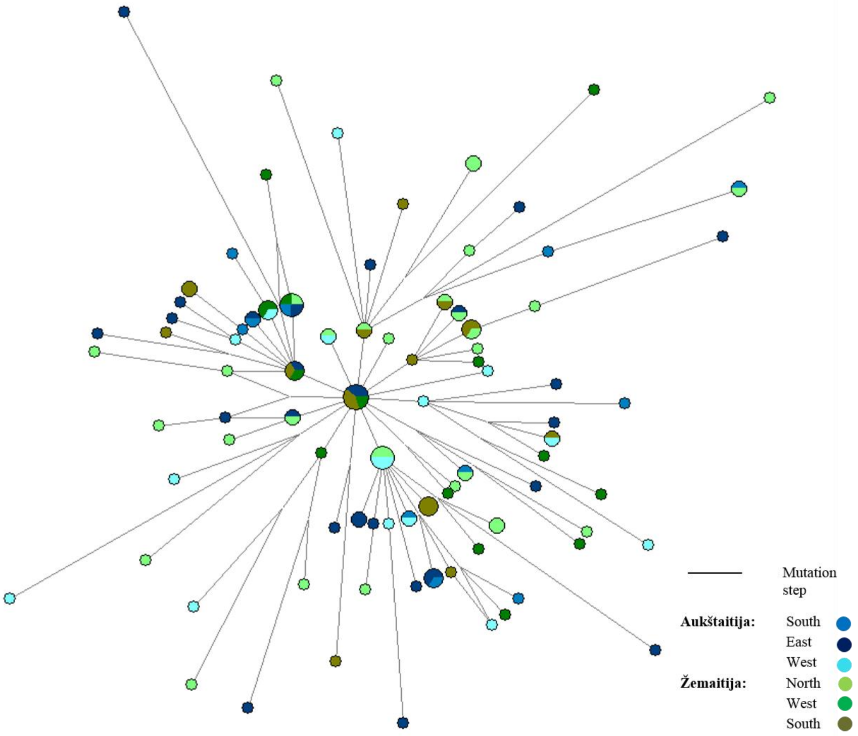


Fig. 3.10. Phylogenetic network of the Y chromosome haplogroup N1c1 based on the distribution of Y chromosome 15 STRs. Reduced median (RM) network algorithm (reduction threshold 1) was used to create the network. Weights for all STRs were given according to the mutation rate of a particular STR (Ballantyne *et al.* 2010) (Annex 1,Table 3).

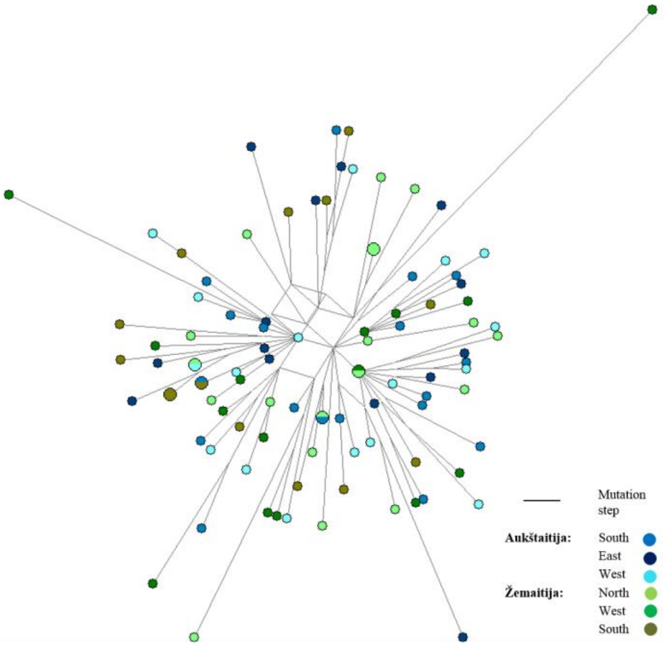


Fig. 3.11. Phylogenetic network of the Y chromosome haplogroup R1a1a based on the distribution of Y chromosome 15 STRs. Reduced median (RM) network algorithm (reduction threshold 1) was used to create the network. Weights for all STRs were given according to the mutation rate of a particular STR (Ballantyne *et al.* 2010).

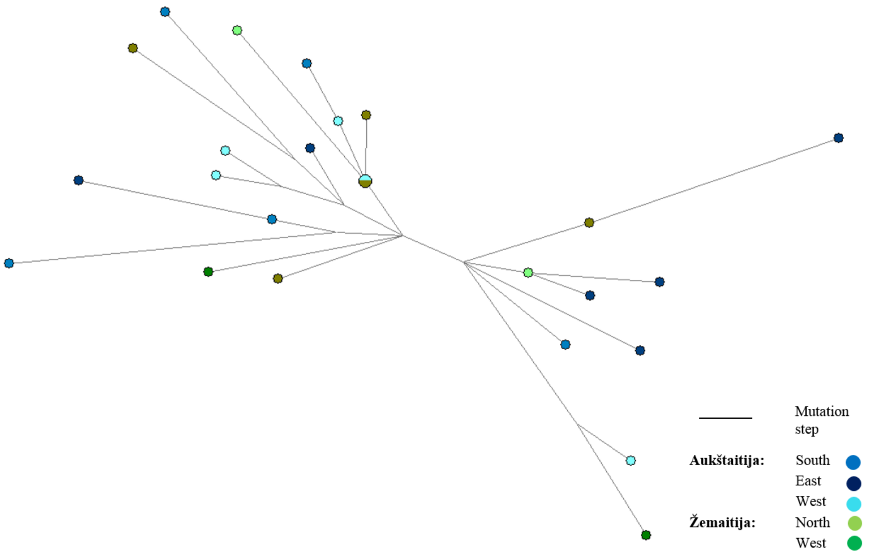


Fig. 3.12. Phylogenetic network of the Y chromosome haplogroup R1a1a1g based on the distribution of Y chromosome 15 STRs. Reduced median (RM) network algorithm (reduction threshold 1) was used to create the network. Weights for all STRs were given according to the mutation rate of a particular STR (Ballantyne *et al.* 2010).

Table 3.4. TMRCA of the most frequent Y chromosome haplogroups based on the distribution of Y chromosome STRs

Mutation rate – 8.5·10^{-3} mutation/marker/generation (Ballantyne *et al.* 2010)			
Haplogroup	**N**	**TMRCA (years)**	**Start of the population growth (years)**
N1c1	119	8,312.3	4,456.1
R1a1a	97	15,512.3	4,580.5
R1a1a1g	25	8,606.6	4,206.7
Mutation rate – 6.9·10^{-4} mutation/marker/generation (Zhivotovsky *et al.* 2004)			
Haplogroup	**N**	**TMRCA (years)**	**Start of the population growth (years)**
N1c1	119	20,382.0	10,531.6
R1a1a	97	49,726.3	11,743.3
R1a1a1g	25	22,666.7	9,799.9
Mutation rate – 2·10^{-3} mutation/marker/generation (Luca *et al.* 2007)			
Haplogroup	**N**	**TMRCA (years)**	**Start of the population growth (years)**
N1c1	119	10,387.0	5,430.1
R1a1a	97	24,638.5	7,027.2
R1a1a1g	25	15,935.1	7,300.8

TMRCA based on the distribution of Y chromosome STRs was calculated for the most frequent Y chromosome haplogroups (N1c1, R1a1a, R1a1a1g). The *a priori* parameters of the model: (1) the first stage of the constant population size with the second stage of population growth; (2) initial population size for the haplogroup R1a1a1g was 500 and for haplogroups N1c1, R1a1a was 1,000; (3) population growth rate was 0–4% per generation; (4) the mutation rate (Annex 1,Table 4) (Ballantyne *et al.* 2010; Kasperavičiūtė *et al.* 2004;

Luca *et al.* 2007). The calculated TMRCA depended mostly on the *a priori* parameter – mutation rate (Table 3.4).

MDS based on Nei (D) distances (Annex 1, Table 6, Table 7) yielded a genetic map of Europe. Genetic distances (Annex 1, Table 5) were calculated based on the distribution of Y chromosome haplogroups (E1, G, I1, J2, N3, R1a, R1b). In the MDS map, clusters of the European population are seen: (1) the Baltic region population (Lithuania, Latvia, Estonia); (2) Scandinavian and North European populations (Norway, Sweden, West and East Finland); (3) Eastern and Central European populations (European part of Russia, Belorussia, Poland, Ukraine); (4) Central and Southern European populations (Moldavia, Czech, Romania, Greece); (5) Western European populations (Germany, Denmark, Britain, Greenland). Positions of populations in the MDS map were according to the geographic locations of populations except for Italy and Turkey (Fig. 3.13).

Results of the PCA based on Nei (D) distances among the European populations showed that PC1 explained 55.45% and PC2 29.36% of genetic variation (Fig. 3.14). The genetic variation explained by PC1 was due to differences between West/East Finland, the Baltic region population and the rest of the studied European populations.

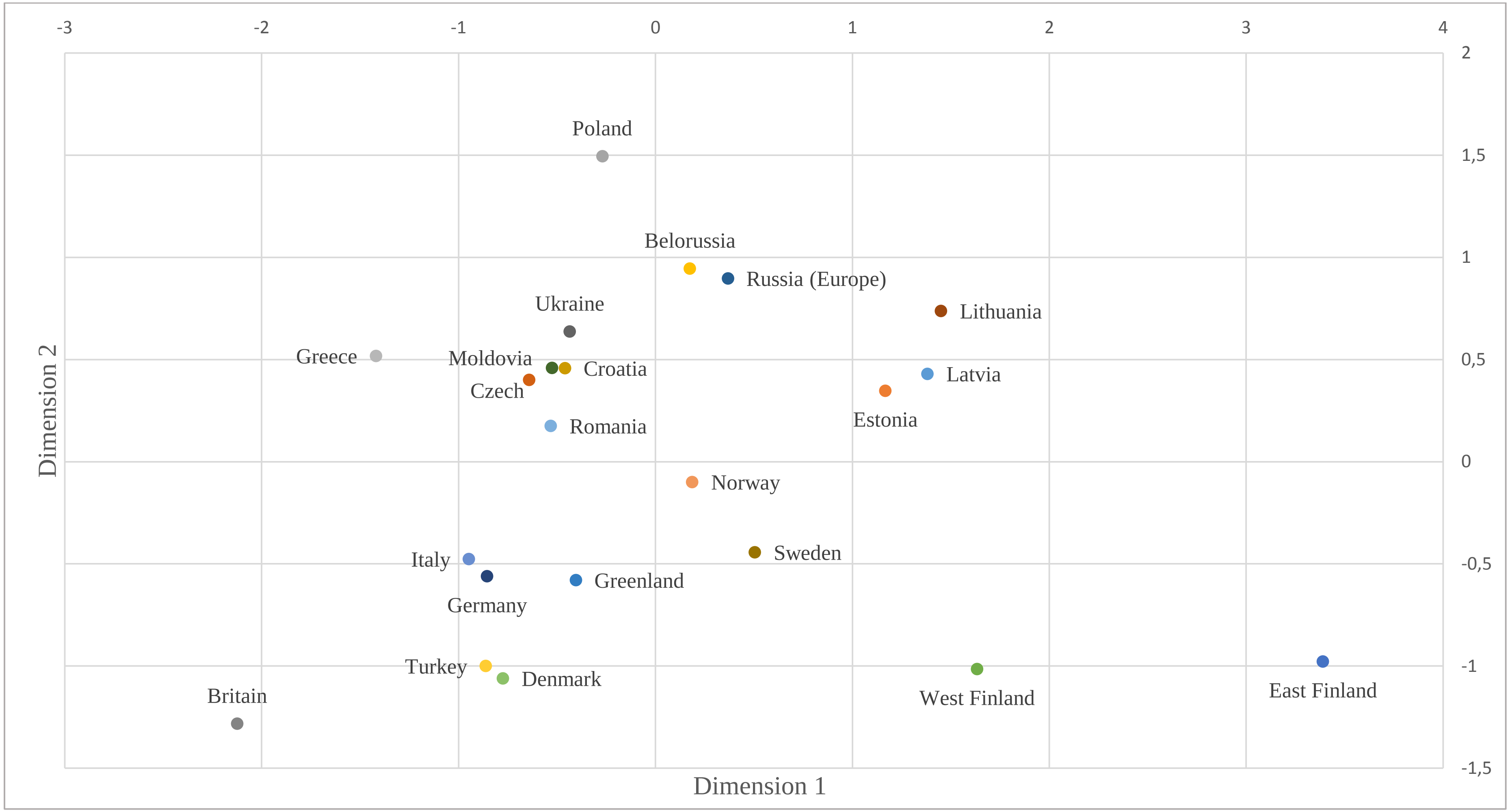


Fig. 3.13. Scatter plot of MDS analysis based on Nei (D) distances (E1, G, I1, J2, N3, R1a, R1b) among European populations. MDS stress and correlation coefficient (r^2): (1) Nei (D) distance, *stress* = 0.16853, r^2 = 0.88982.

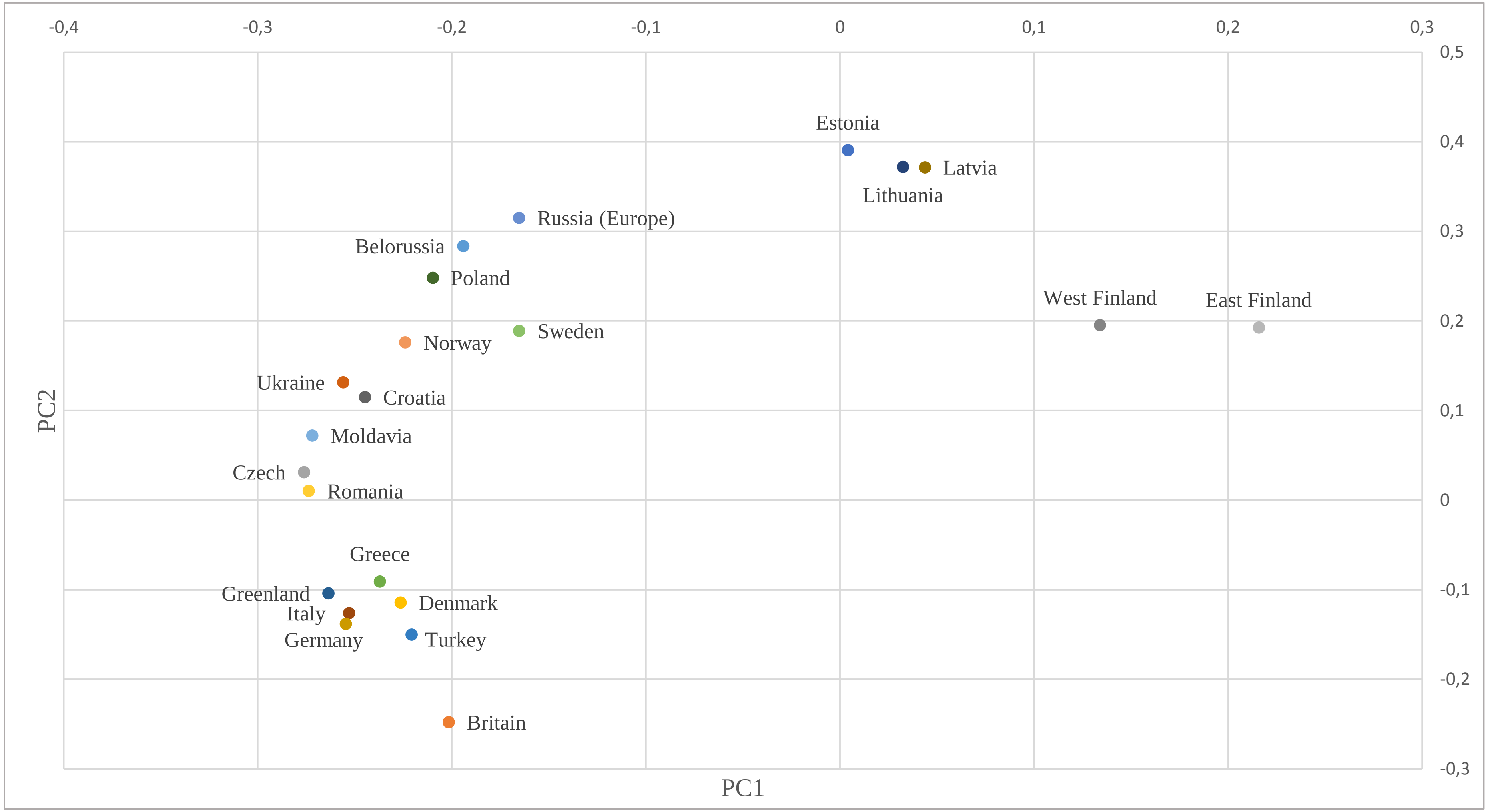


Fig. 3.14. Scatter plot of PCA based on Nei (D) distances (E1, G, I1, J2, N3, R1a, R1b) between European populations. PC1 explains 55.45%, PC2 29.36% of genetic variation among the European populations studied.

The genetic variation barriers, i.e. Y chromosome haplogroups, among the neighboring European populations (Lithuania, Latvia, Estonia, Sweden, East/West Finland, European part of Russia, Belorussia, Poland) were identified. (Fig. 3.15).

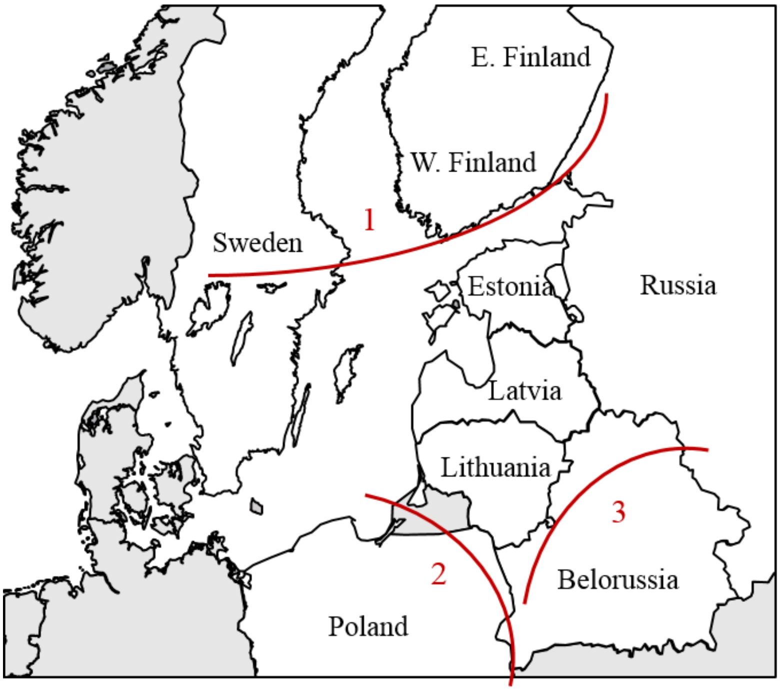


Fig. 3.15. Barriers of genetic variation i.e. Y chromosome haplogroups, among the neighboring European populations. Barriers were numbered in the order of priority. Analysis is based on Nei (D) distances.

SUMMARY

The previous study, performed by D. Kasperavičiūtė and colleagues described the genetic structure of the Lithuanian population based on the distribution of Y chromosome haplogroups and haplotypes (9 STRs). Moreover, the Y chromosome data of Lithuanian population were compared with the data of the other two Baltic region states (Latvia, Estonia) (Kasperavičiūtė *et al.* 2004). Other groups of scientists analyzed North or North East European populations, including the population of Lithuania, based on a lower number of haplogroups and haplotypes (<10 STRs) (Kushniarevich *et al.* 2013; Laitinen *et al.* 2002; Lappalainen *et al.* 2008). The genetic structure of a population can be undetected if samples are included intentionally, i.e. samples of specific ethnics, sample sets that do not represent a population properly.

Data of this study consisted of 16 different Y chromosome haplogroups and 250 haplotypes out of 293 samples. The distribution of Y chromosome haplogroups (25 SNPs) and haplotypes (15 STRs) was determined in the largest sample group as compared with the previous studies. The distribution of the most frequent Y chromosome haplogroups was similar to that in the previous studies, but the calculated frequencies were slightly different. The frequency of the haplogroup N1c1 was 40.3%, i.e. slightly lower as compared with the other most frequent haplogroup R1a1a – 41.6% (including R1a1a1g – 8.5%.). Results of the previous study showed higher differences in the frequency of these two main haplogroups (respectively 37% and 45%) in the Lithuanian population (Kasperavičiūtė *et al.* 2004).

The results of the present study showed that the distribution of three main Y chromosome haplogroups (N1c1, R1a1a, R1a1a1g) was unequal with the gradient from North to South and vice versa. Most of the samples had the haplogroup N1c1 in the Northern part of Lithuania (North Žemaitija and East Aukštaitija, each ~49%) and least in the South (South Aukštaitija ~29%). The haplogroups R1a1a and R1a1a1g were more frequently detected in Southern Lithuania (South Aukštaitija, respectively ~47% and ~13%) as compared with Northern Lithuania (North Žemaitija, respectively ~30% and ~3%, also East Aukštaitija, respectively ~25% and ~11%). No such gradient was detected in previous studies. The observed gradient could be due to the different direction of the migration waves. The haplogroup N1c1 reached the present-day territory of Lithuania from the North as the highest frequency of this haplogroup is detected in the Northern Europe (Lappalainen *et al.* 2008). The haplogroups R1a1a(M198) and R1a1a1g(M458) could reach the territory of the present-day Lithuania from Central and/or Eastern Europe as these parts of Europe hold the highest frequency of these haplogroups. Moreover, there was a discussion that the subhaplogroup R1a1a1g(M458) could originate from the territory of the present-day South or Central Poland (Underhill *et al.* 2010). More detailed analyses of the haplogroup R1a1a(M198) will be needed to find out the migration scenario of this haplogroup in and within the territory of the present-day Lithuania. These hypotheses do no reveal the complexity of the formation and evolution of paternal lineage in the territory of the present-day Lithuania.

The calculated F_{ST} distances between the North (South Žemaitija and East Aukštaitija) and the South (South Aukštaitija) of Lithuania were statistically significant ($P < 0.05$) in this study. The observed significant genetic distances were due to distribution of three main haplogroups (N1c1, R1a1a, R1a1a1g) and were not seen in the previous studies. The possible reasons were discussed by D. Kasperavičiūtė and colleagues: (1) the Baltic tribes from which modern-day Lithuanians originated may have been genetically close; (2) the sample size of 30–40 individuals could be too small to detect genetic differences among the ethno-linguistic groups of Lithuania. A larger sample size (overall ~300 samples) in this study supports the second reason of not identified genetic differences (Kasperavičiūtė *et al.* 2004).

Results of the AMOVA have showed that almost all (~99%) Y chromosome variations (haplogroups and haplotypes) fall among individuals within the ethno-linguistic groups of Lithuania. The previous conclusion that the Lithuanian population is very homogeneous (Kasperavičiūtė *et al.* 2004) is confirmed by the results of this study. Moreover, the cultural and lingual differences are not the main source of the present Y chromosome genetic variation (haplogroups and haplotypes). The geographical location (North, Middle, South) of ethno-linguistic groups on the territory of Lithuania could be the main reason for the observed Y chromosome genetic variation ($P < 0.05$). The observed differences could be due to the certain degree of genetic isolation of the ethno-linguistic groups of Lithuania.

The scatterplot of MDS based on Nei (D) distances showed three clusters of ethno-linguistic groups of Lithuania, and this supported the results of the AMOVA. The results of PCA showed genetic differences among the ethno-linguistic groups of Lithuania which account for about 2% of all genetic variations. PC1 and PC2 explained approximately 79% of genetic variations between the northwestern and southeastern ethno-linguistic groups, and this also supported the results of AMOVA. Moreover, the first three barriers of the Y chromosome haplogroups' variation were identified between East Aukštaitija, South Aukštaitija, North Žemaitija and the rest of the ethno-linguistic groups of Lithuania. All these results confirmed the highest differentiation level between the northwest and the southeast Lithuanian population.

The lower haplotype diversity (0.9927 (±0.0023)) and the star-like phylogenetic tree of the haplogroup N1c1 showed a sudden decrease of the population size (bottle-neck effect) or the small size of settler group (founder effect) with the later population growth. Probably the population growth started after the Last Glacial Maximum (LGM) i.e. 4,456−10,531 years before present (YBP). This period coincides with the start of the Baltic tribe formation, i.e. 5,000 YBP, based on archaeological and linguistic studies. A different phylogenetic tree was seen for the haplogroup R1a1a, based on the distribution of haplotypes. Moreover, haplotype diversity was higher (0.9987 (±0.0017)) as compared with the haplogroup N1c1. The phylogenetic tree was more scattered and its branches were longer showing more mutation steps. This could be due to several reasons. Firstly, more than one genetically close population inhabited the territory of the present-day Lithuania. Secondly, multiple invasions of genetically close populations could occur during the growth of the haplogroup R1a1a which started after LGM, i.e. 4,580−11,743 YBP. Central and Eastern Europe could be the source of repeated migration waves based on the frequency gradient of haplogroups R1a1a and R1a1a1g. The present-day gene pool of parental linage is a result of the interaction of evolutionary forces such as migration, mutation, natural selection, and genetic drift. The results of this study suggest that the evolution of two main Y chromosome haplogroups (N1c1 and R1a1a, R1a1a1g) within the territory of present-day Lithuania may be different.

The scatter plot of MDS showed that the position of the Lithuanian population was near the populations that are geographically close (Latvia and Estonia). The distances were slightly greater to Slavs (European part of Russia, Belorussia, Poland). The results of previous studies were supported by the results of this study (Kasperavičiūtė *et al.* 2004; Kushniarevich *et al.* 2013; Lappalainen *et al.* 2008). The analyzed European and Near East populations in the scatter plot of PCA (PC1 55.45%, PC2 29.36%) were positioned according to their geographical locations. The highest differences of Y chromosome haplogroup variations were seen between West/East Finland, the Baltic region populations, and the rest of the analyzed populations (PC1).

Three barriers of Y chromosome genetic variation were seen among Sweden together with West/East Finland, Poland and Belorussia and the rest of the analyzed neighboring populations. These results suggest that the influence of genetic distances is higher than of the linguistic differences among the analyzed populations. The genetic barrier between the North and the rest of Europe could be due to different frequencies of haplogroups I1 and N1c1. The barrier between Poland and the rest of the analyzed European populations may be due to the higher frequency of the R1a1a haplogroup which could derive from the present-day territory of Central and Southern Poland. Moreover, the genetic pool of Western Europe, i.e. the higher frequency of R1b, was one of the sources of the observed barrier. The gene pool of Belorussia is influenced by Southern Europe, i.e. haplogroups J2 and E1. The results of this analysis could be affected by the different accuracy of published data, i.e. Y chromosome haplogroup distribution and frequencies within certain populations analyzed in this study.

3.2. Genetic structure and genetic distances of Lithuanian population based on the distribution of mtDNA haplogroups and complete mtDNA

Complete mtDNA was sequenced with the average 352.6 (±203.66) coverage depth for all samples from the Lithuanian population (Fig. 2.4). The lowest obtained coverage depth was 23 and the highest 1,753. Totally, 116 different mtDNA haplogroups were identified in the studied Lithuanian population. Grouping of all 116 identified haplogroups according to the hierarchical principles of the phylogenetic tree was used for some analyses based on haplogroup variation: 23 haplogroups or 69 haplogroups. All identified mtDNA haplogroups belong to 11 different phylogenetic lineages (A, H, HV, I, J, K, M, N, T, U, V, W), and seven of them (H, I, J, K, T, U, V) are the most frequent in Europe.

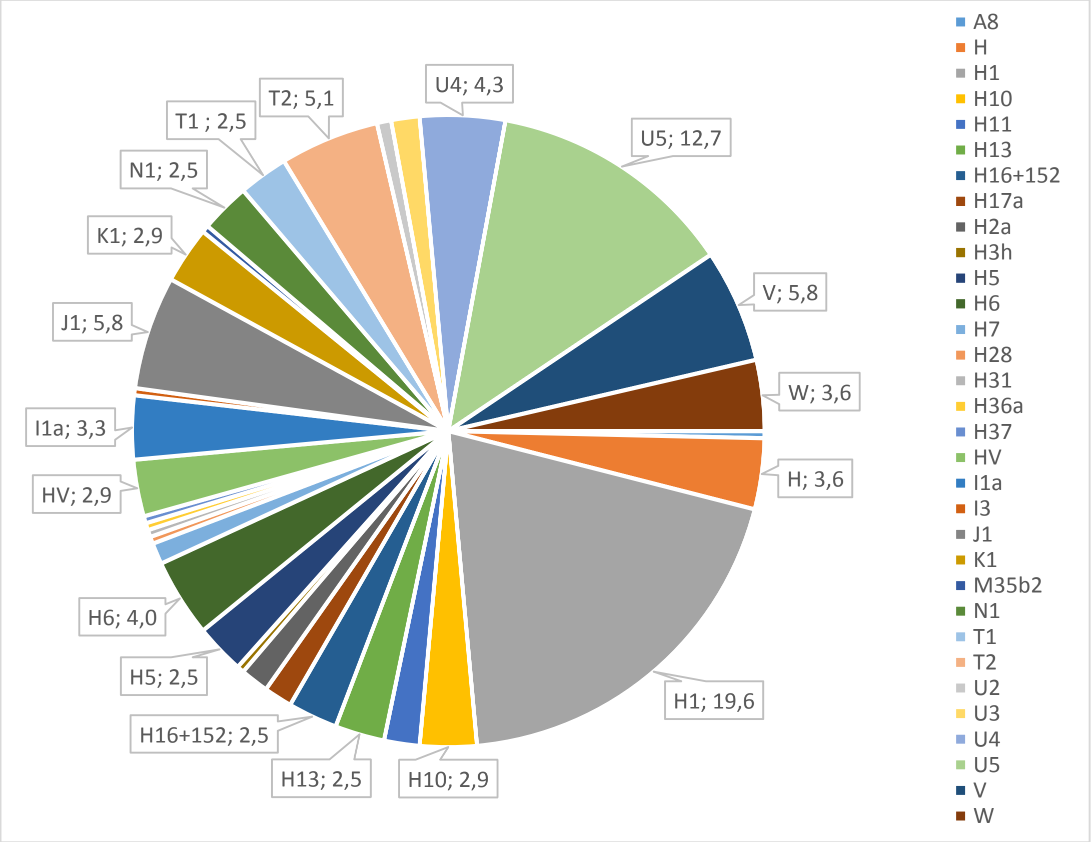


Fig. 3.16. Distribution of mtDNA phylogenetic lineages (frequency,%) in the studied Lithuanian population.

The most frequent phylogenetic lineages in the studied Lithuanian population were the phylogenetic lineage H (45.3%), U (19.2%), T (7.6%), J1 (5.8%), V (5.8%). The phylogenetic lineages A, HV, I, K, M, N, W comprised

<5% of the studied Lithuanian population. Five most frequent phylogenetic lineages (H, J1, T, U, V) comprised ~84% and the other six phylogenetic lineages were present in ~16% of the studied Lithuanian population (Fig. 3.16).

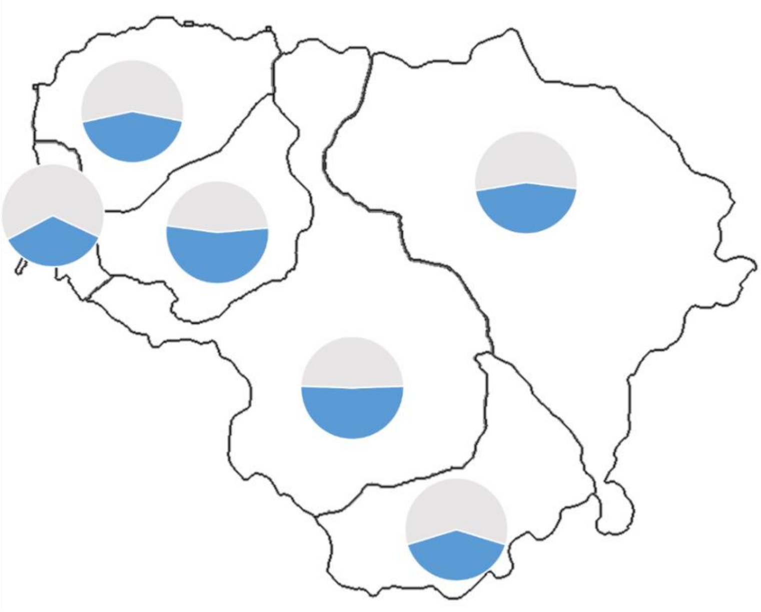


Fig. 3.17. Distribution of mtDNA phylogenetic lineage H (blue label) within six ethno-linguistic groups of Lithuania.

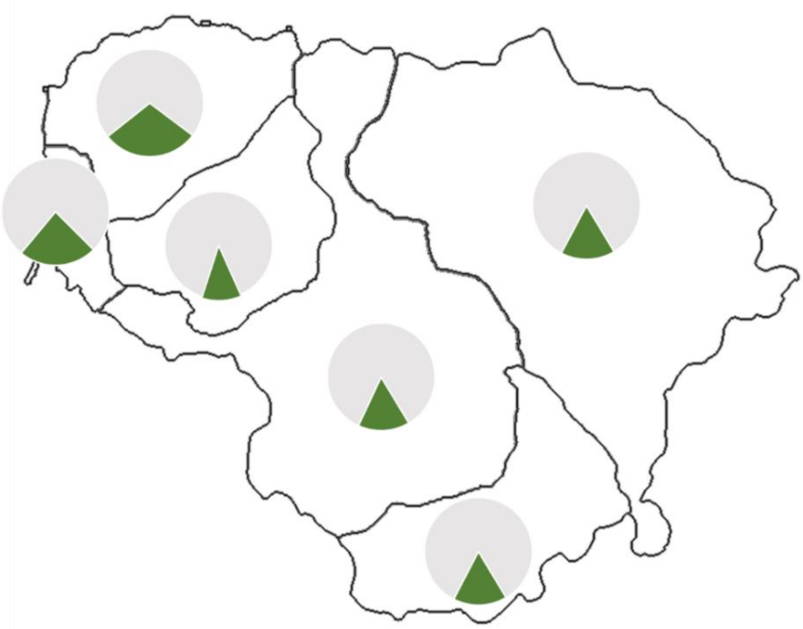


Fig. 3.18. Distribution of mtDNA phylogenetic lineage U (green label) within six ethno-linguistic groups of Lithuania.

No evident patterns of the distribution of mtDNA phylogenetic lineages (H, U) across the territory of Lithuania due to geographic, linguistic or cultural differences were observed (Fig. 3.17, Fig. 3.18).

Table 3.5. Molecular diversity indices of mtDNA* and HVRI region

mtDNA (16,548 bp*)							
Population	**N**[1]	**HD (±SD)**[2]	**S**[3]	**π (±SD)**[4] **n·10^{-3}**	**k (±SD)**[5]	**H**[6]	**Tajima`s D**[7]
E. Aukštaitija	55	0.998 (±0.004)	281	1.645 (±0.81)	27.2 (±12.1)	52	−2.00 (p = 0.005)
S. Aukštaitija	37	0.999 (±0.007)	268	1.706 (±0.85)	28.24 (±12.6)	36	−2.11 (p = 0.003)
W. Aukštaitija	45	0.999 (±0.005)	291	1.677 (±0.83)	27.7 (±12.4)	44	−2.14 (p = 0.003)
Aukštaitija	137	0.999 (±0.001)	519	1.668 (±0.81)	27.61 (±12.2)	129	−2.34 (p < 0.001)
N. Žemaitija	62	0.998 (±0.003)	330	1.624 (±0.80)	26.87 (±11.9)	59	−2.19 (p = 0.002)
S. Žemaitija	43	0.998 (±0.006)	246	1.363 (±0.68)	22.55 (±10.1)	41	−2.23 (p < 0.001)
W. Žemaitija	34	0.996 (±0.008)	279	1.915 (±0.95)	31.69 (±14.2)	32	−2.04 (p = 0.003)
Žemaitija	139	0.999 (±0.001)	494	1.619 (±0.79)	26.79 (±11.8)	128	−2.31 (p < 0.001)
Lithuania	276	0.999 (±0.0004)	730	1.644 (±0.80)	27.21 (±12.0)	252	−2.42 (p < 0.001)
HVRI region (16,001−16,568 bp)							
Population	**N**[1]	**HD (±SD)**[2]	**S**[3]	**π (±SD)**[4] **n·10-3**	**k (±SD)**[5]	**H**[6]	**Tajima`s D**[7]
E. Aukštaitija	55	0.983 (±0.009)	53	10.47 (±5.6)	5.92 (±2.87)	42	−1.67 (p = 0.0016)
S. Aukštaitija	37	0.970 (±0.018)	46	8.92 (±4.93)	5.05 (±2.51)	27	−1.94 (p = 0.01)
W. Aukštaitija	45	0.982 (±0.013)	52	9.52 (±5.19)	5.39 (±2.65)	37	−1.92 (p = 0.008)
Aukštaitija	137	0.981 (±0.007)	84	9.76 (±5.22)	5.52 (±2.67)	94	−2.03 (p = 0.004)
N. Žemaitija	62	0.986 (±0.007)	58	10.51 (±5.63)	5.95 (±2.88)	47	−1.76 (p = 0.009)
S. Žemaitija	43	0.971 (±0.014)	45	8.03 (±4.47)	4.54 (±2.28)	29	−1.97 (p = 0.009)
W. Žemaitija	34	0.991 (±0.009)	51	10.9 (±5.91)	6.17 (±3.01)	29	−1.85 (p = 0.014)
Žemaitija	139	0.985 (±0.004)	76	9.85 (±5.27)	5.57 (±2.69)	84	−1.88 (p = 0.015)
Lithuania	276	0.984 (±0.004)	101	9.81 (±5.23)	5.55 (±2.68)	151	−1.99 (p = 0.001)

[*] All positions with gaps removed.

[1] Sample size.

[2] Haplotype diversity and standard deviation (haplotype – complete mtDNA).

[3] Number of segregating sites.

[4] Nucleotide diversity and standard deviation (n·10^{-3}).

[5] Mean pairwise difference and standard deviation.

[6] Number of haplotypes.

[7] Tajima`s D and p-value.

The mtDNA haplotype diversity (0.999 (±0.0004)) was even across all six ethno-linguistic groups of Lithuania. Totally, 730 segregating sites were identified. The mean pairwise haplotype difference was 27.21 (±12) and the mean nucleotide diversity $1.6 \cdot 10^{-3}$ ($\pm 0.80 \cdot 10^{-3}$) in the studied Lithuanian population. South Žemaitija has the lower and West Žemaitija higher above mentioned estimates as compared with the rest of the ethno-linguistic groups of Lithuania. The calculated Tajima`s D values were statistically significant ($P < 0.05$) for all ethno-linguistic groups of Lithuania (Table 3.5.). Nucleotide diversity (π) of the HVRI region (16,001−16,568 bp) was eight times greater ($9.81 \cdot 10^{-3}$ ($\pm 5.23 \times 10^{-3}$)) as compared with the coding region of mtDNA ($1.206 \cdot 10^{-3}$ ($\pm 0.593 \times 10^{-3}$)).

The control region, the *MT-ATP6* gene had significantly more and the *MT-RNR2* (16S), tRNR, *MT-CO1*, *MT-ND4L*, *MT-RNR1* (12S) genes had significantly (χ^2 test, $P < 0.05$) less variable positions than was expected from the length of each region/gene. The number of variable positions in the control region was three times greater than was expected from the length of this region. The ratio of the transitions to transversions (~16:1) was higher for all regions of mtDNA. The control region, the *MT-ATP6* gene had significantly more and the *MT-RNR1* (12S), *MT-RNR2* (16S), tRNR, *MT-CO1*, *MT-ND4L* genes had significantly less (χ^2 test, $P < 0.05$) transitions than was expected from the length of each region/gene. Moreover, the control region had significantly more and the *MT-RNR2* (16S) gene had significantly less (χ^2 test, $P < 0.05$) transversions than was expected from the length of each region/gene. There was more synonymous changes as compared with nonsynonymous changes (~2:1). The ratio of nonsynonymous polymorphisms per nonsynonymous site to synonymous polymorphisms per synonymous site (p_N/p_S) varied significantly among genes (χ^2 test, $P = 0.0035$). The most extreme values were observed for *MT-ATP6* and *MT-ND6* genes and when these genes were removed from the analysis the p_N/p_S ratio did not vary significantly among the remaining genes (χ^2 test, $P > 0.05$) (Table 3.6).

Table 3.6. The number of mtDNA variable sites, transitions, transversions, nonsynonymous and synonymous polymorphisms, p_N/p_S ratio

Region/gene	**S**[1]	**TS**[2]	**TV**[3]	**Synonymous**	**Nonsynonymous**	**p_N/p_S** [4]
Control r.[5]	141	132	9	−	−	−
MT-RNR1 (12S)	24	22	2	−	−	−
MT-RNR2 (16S)	36	36	0	−	−	−
tRNR	37	35	2	−	−	−
MT-ATP6	42	40	2	15	27	0.640
MT-ATP8	11	11	0	6	5	0.269
MT-CO1	43	40	3	36	7	0.065
MT-CO2	28	25	3	19	9	0.157
MT-CO3	35	35	0	22	13	0.197
MT-CYTB	48	45	3	29	19	0.219
MT-ND1	39	36	3	28	11	0.143
MT-ND2	48	44	4	34	14	0.140
MT-ND3	15	15	0	11	4	0.118
MT-ND4L	5	5	0	3	2	0.255
MT-ND4	52	48	4	45	8	0.063
MT-ND5	91	87	4	66	25	0.128
MT-ND6	18	15	3	16	2	0.042
Total	713	671	42	330	146	−

[1] Number of variable sites.

[2] Number of transitions.

[3] Number of transversions.

[4] Ratio of nonsynonymous polymorphisms per nonsynonymous site to synonymous polymorphisms per synonymous site.

[5] Control region: HVRI region (16,001−16,568 bp) and HVRII region (1−576 bp).

Pairwise differences based on haplotypes (577−16,023 bp) (Annex 1,Table 9) and F_{ST} (Table 3.7), Nei (D) (Annex 1, Table 8) distances based on grouping into 23 haplogroups (the frequency of each haplogroup >1%) and grouping into 69 haplogroups were calculated. No statistically significant correlation between Nei (D) and geographic distances was observed (Mantel test, r = 0.242; $P = 0.238$; 10,000 permutations).

Table 3.7. Pairwise F_{ST} distances based on 23 groups of haplogroups (above diagonal) and based on 69 groups of haplogroups (below diagonal). Significant F_{ST} distances in bold ($P < 0.05$; 10,000 permutations). In Italic $P = 0.0693$

	SA	**EA**	**WA**	**SŽ**	**WŽ**	**NŽ**
SA	0	0.00848	−0.00698	−0.00104	−0.01049	0.00991
EA	0.00449	0	−0.0019	*0.01215*	−0.01344	−0.00294
WA	0.00242	−0.00038	0	−0.00069	−0.01035	0.00714
SŽ	0.00329	**0.00974**	0.00392	0	0.0011	**0.02343**
WŽ	0.00259	−0.00343	−0.00214	0.00392	0	−0.00985
NŽ	0.00388	−0.00488	0.00276	**0.00964**	−0.0067	0

The AMOVA analysis was based on pairwise F_{ST} distances (haplogroups) and pairwise differences (haplotypes). The classification of the ethno-linguistic groups of Lithuania was based on geographical location, language and cultural differences. In all classification cases, almost all variation (~99%) fell among samples within the ethno-linguistic groups of Lithuania. Significant AMOVA results were obtained when the groups were classified based on the geographic West−East location in the territory of Lithuania (P < 0.05; 10,000 permutations) (Table 3.8).

Table 3.8. Results of AMOVA based on mtDNA haplogroups and haplotypes distribution. Significant indices are in bold (P < 0.05; 10,000 permutation). Grouping of ethno-linguistic groups of Lithuania was based on geographic location in the territory of Lithuania: (1) North, West Žemaitija; (2) South Žemaitija; (3) West, South Aukštaitija and/or (4) East Aukštaitija

<table>
<tr><td colspan="4">Haplogroups (23)[1]</td></tr>
<tr><th>Source of variation</th><th>df[1]</th><th>Fixation index (F)</th><th>Percentage of diversity (%)</th></tr>
<tr><td>between groups (4)</td><td>3</td><td><b>0.01127</b> (<i>P</i> = 0.042)</td><td>1.13</td></tr>
<tr><td>between populations within group</td><td>2</td><td>−0.00845[3]</td><td>−0.84</td></tr>
<tr><td>within population</td><td>270</td><td>0.00291[3]</td><td>99.71</td></tr>
<tr><td colspan="4">Haplogroups (69)[1]</td></tr>
<tr><th>Source of variation</th><th>df[1]</th><th>Fixation index (F)</th><th>Percentage of diversity (%)</th></tr>
<tr><td>between groups (4)</td><td>3</td><td>0.0047 (<i>P</i> = 0.071)</td><td>0.47</td></tr>
<tr><td>between populations within group</td><td>2</td><td>−0.00225[3]</td><td>−0.22</td></tr>
<tr><td>within population</td><td>270</td><td>0.00246[3]</td><td>99.75</td></tr>
<tr><td colspan="4">Haplotypes (577−16,023 bp)[2]</td></tr>
<tr><th>Source of variation</th><th>df[1]</th><th>Fixation index (F)</th><th>Percentage of diversity (%)</th></tr>
<tr><td>between groups (3)</td><td>3</td><td><b>0.00555</b> (<i>P</i> = 0.048)</td><td>0.55</td></tr>
<tr><td>between populations within group</td><td>2</td><td>−0.00299[3]</td><td>−0.30</td></tr>
<tr><td>within population</td><td>270</td><td>0.00257[3]</td><td>99.74</td></tr>
<tr><td colspan="4">Haplotypes (16,569 bp)[2]</td></tr>
<tr><th>Source of variation</th><th>df[1]</th><th>Fixation index (F)</th><th>Percentage of diversity (%)</th></tr>
<tr><td>between groups (3)</td><td>3</td><td><b>0.00480</b> (<i>P</i> = 0.031)</td><td>0.48</td></tr>
<tr><td>between populations within group</td><td>2</td><td>−0.00225[3]</td><td>−0.22</td></tr>
<tr><td>within population</td><td>270</td><td>0.00257[3]</td><td>99.74</td></tr>
</table>

[1] F_{ST} distances based on the distribution of mtDNA haplogroups.

[2] Pairwise differences based on the distribution of mtDNA haplotypes (16,569 bp; 577−16,023 bp).

[3] Not significant (P > 0.05).

An MDS analysis based on F_{ST} distances was performed. The analysis of 23 haplogroups yielded the MDS plot with no patterns of ethno-linguistic group locations. In contrast, the MDS plot based on 69 haplogroups showed four clusters: (1) South Žemaitija; (2) West, South Aukštaitija, West Žemaitija; (3) North Žemaitija; (4) East Aukštaitija (Fig. 3.19).

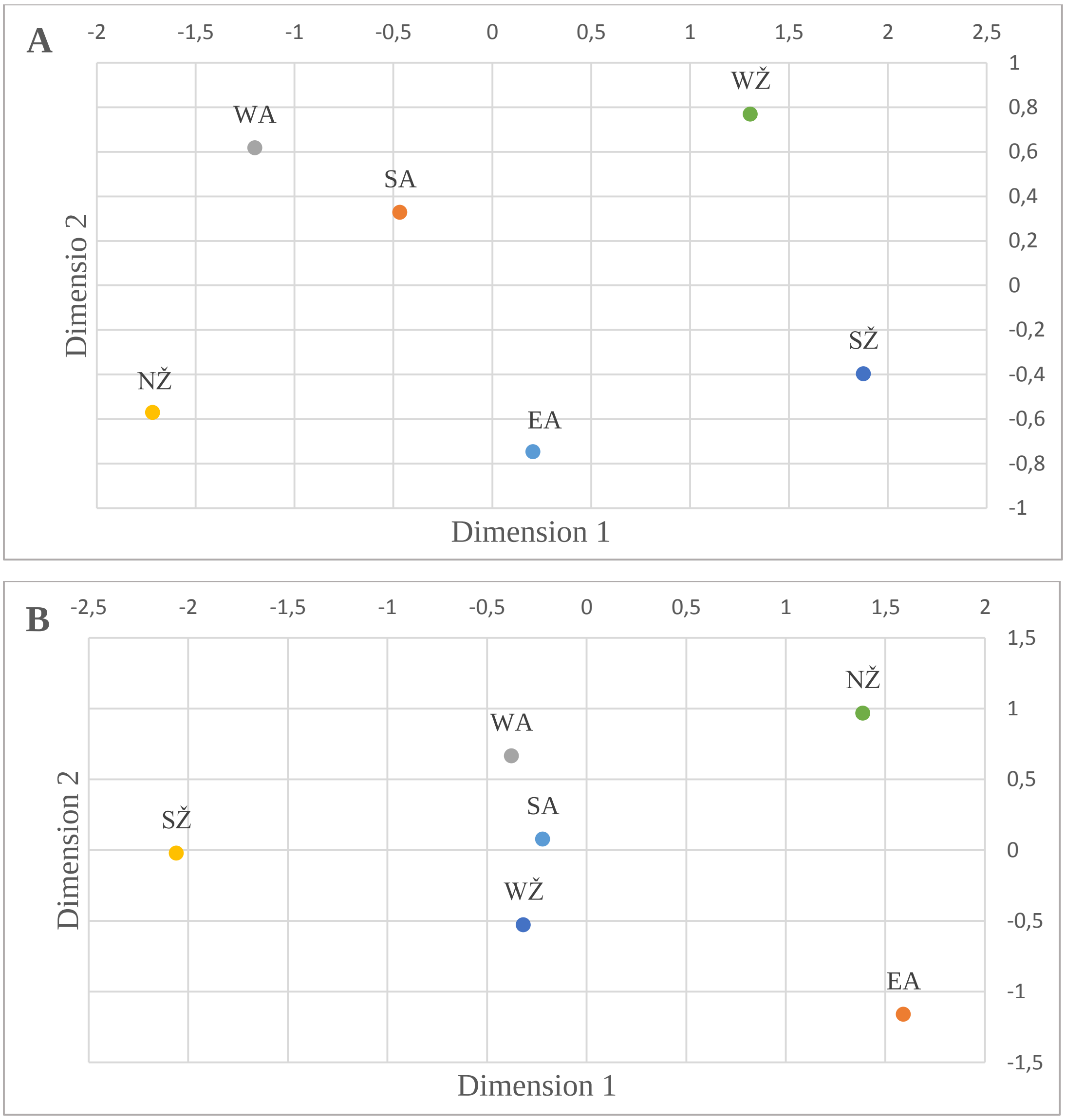


Fig. 3.19. Scatter plot of MDS analysis based on F_{ST} distances ((A) 23 mtDNA haplogroups and (B) 69 mtDNA haplogroups). MDS stress and correlation coefficient (r^2): (1) F_{ST} distance (23 haplogroups), *stress* = 0.00286, r^2 = 0.99994; (2) F_{ST} distance (69 haplogroups), *stress* = 0.00281, r^2 = 0.99995.

A PCA based on F_{ST} distances (69 haplogroups) was performed. The first two PCs explained 63.63% and 21.49% of variation among six ethno-linguistic groups of Lithuania. The PCA based on pairwise differences (577−16,023 bp) showed that PC1 explained 50.0% and PK2 explained 27.96% of among between six ethno-linguistic groups of Lithuania (Fig. 3.20).

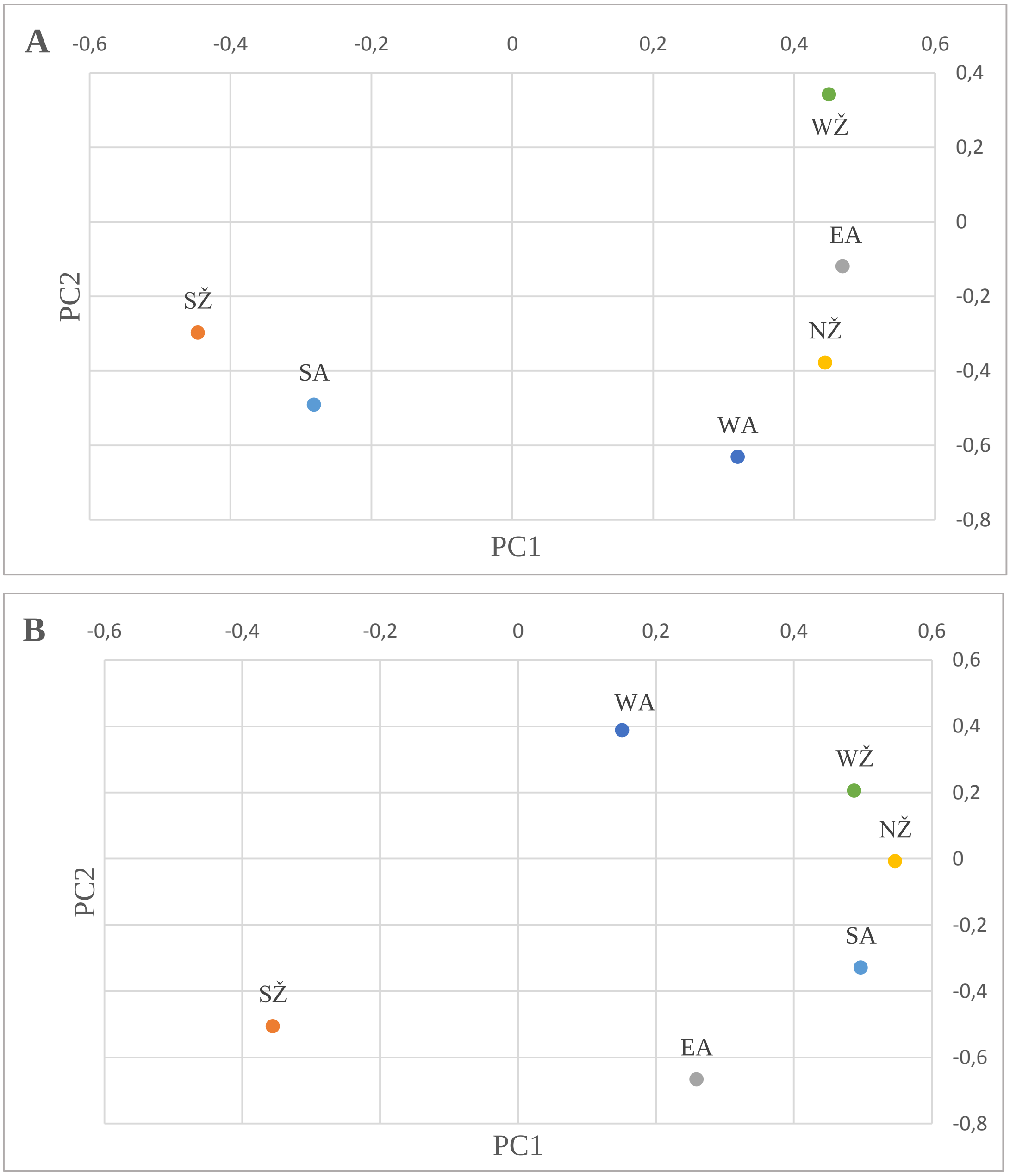


Fig. 3.20. Scatter plot of PCA based on (A) F_{ST} distances (69 haplogroups) and (B) pairwise differences (577−16,023 bp).

Barriers of genetic variation, i.e. mtDNA 69 haplogroups and mtDNA haplotypes (577−16,023 bp), among six ethno-linguistic groups of Lithuania were identified (Fig. 3.21).

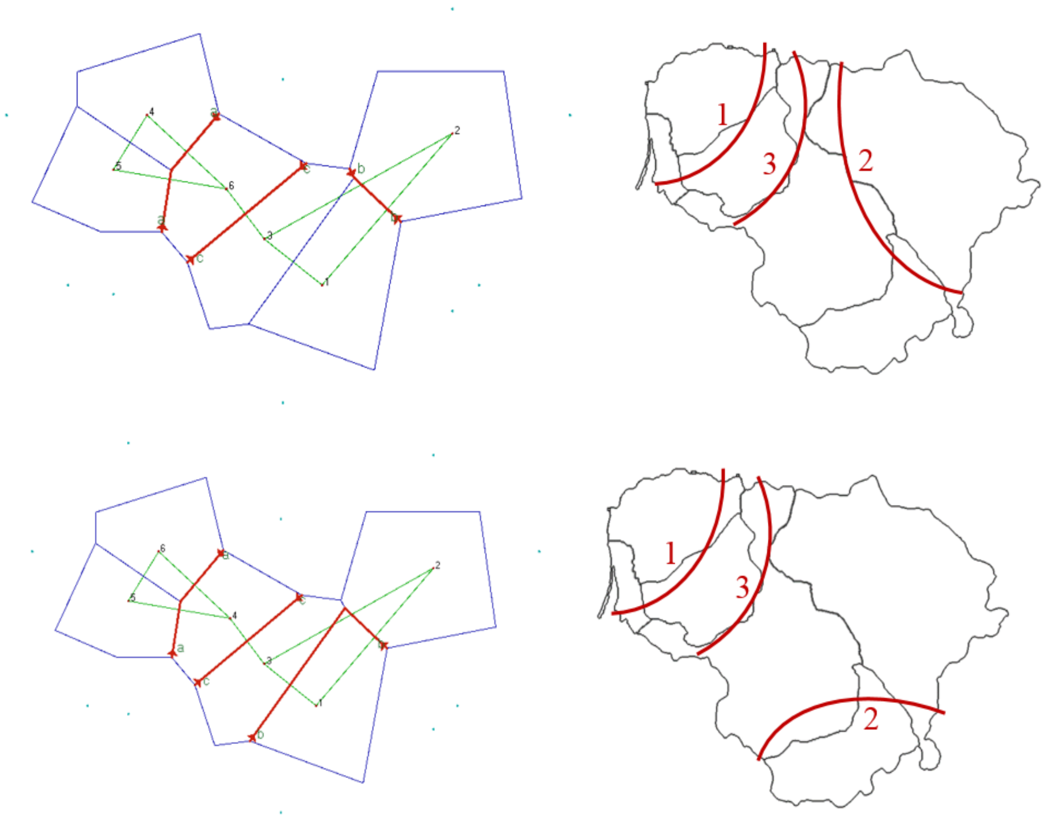


Fig. 3.21. Barriers of genetic variation, i.e. mtDNA haplotypes (577−16,023 bp) (above) and mtDNA 69 haplogroups (below), of the six ethno-linguistic groups of Lithuania. Left: output of the software, i.e. a schematic map as of the ethno-linguistic groups of Lithuania according to submitted coordinates (Annex 1,

Table 1). Right: barriers are numbered in the order of priority. Analysis was based on pairwise differences (top) and F_{ST} (bottom) distances.

Phylogenetic networks based on mtDNA haplotypes (577−16,023 bp) for the most frequent mtDNA phylogenetic lineages (H, U) were constructed. The phylogenetic network of the mtDNA haplogroup H1 had a star-like structure with the most frequent haplotype in the center (4.0%). The cluster of H1 was connected to the rest of the phylogenetic tree through mutation G3010A (Fig. 3.22). In contrast, phylogenetic network of the mtDNA phylogenetic lineage U had no star-like structure (Fig. 3.23).

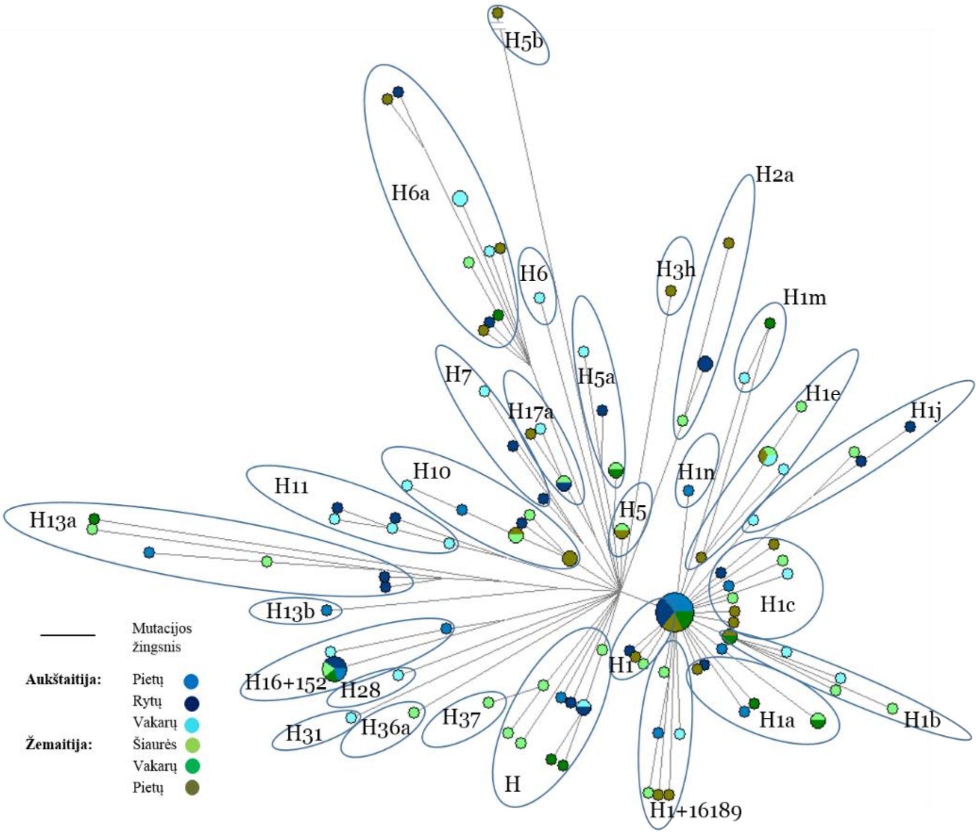


Fig. 3.22. Phylogenetic network of mtDNA phylogenetic lineage H based on the distribution of haplotypes (577−16,023 bp). The Median Joining (MJ) network algorithm (ε = 1) was used to create the network.

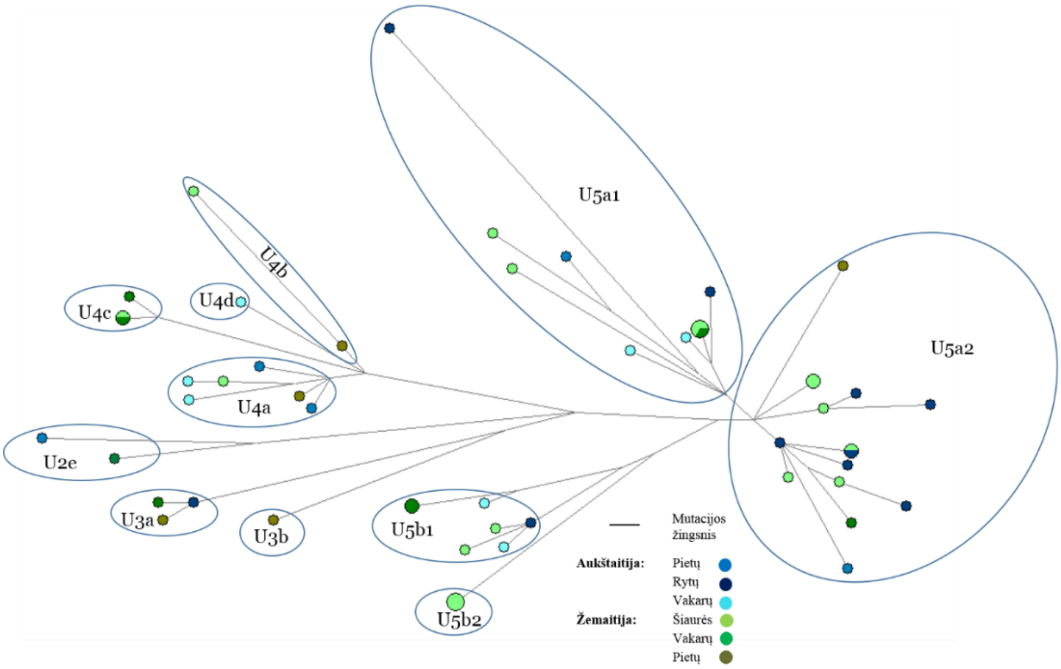


Fig. 3.23. Phylogenetic network of mtDNA phylogenetic lineage U based on the distribution of haplotypes (577−16,023 bp). The Median Joining (MJ) network algorithm (ε = 1) was used to create the network.

The average number of mutations (ρ) per phylogenetic lineage, age and confidence intervals (CI 95%) were calculated based on the distribution of haplotypes (16,568 bp; four positions i.e. 16,182, 16,183, 16,194, 16,519 were removed from the analysis) for mtDNA phylogenetic lineages H and U (Costa *et al.* 2013; Saillard *et al.* 2000; Soares *et al.* 2009) (Table 3.9).

Table 3.9. Estimated age of the mtDNR phylogenetic lineages H and U

Haplogroup	**ρ**[1]	**(±SD)**	**Estimated age**	
			years	**CI 95%**
H	3.80	0.440	1,063	7,723−12,434
H1a	2.14	0.430	5,588	3,363−7,845
H1b	3.66	0.611	9,681	6,448−12,975
H1c	2.08	0.173	5,429	4,530−6,332
H1e	3.66	0.611	9,681	6,448−12,975
H1j	6.00	1.500	16,172	8,054−24,633
H2a	1.75	0.560	4,554	1,682−7,481
H5	6.14	0.878	16,567	11,762−21,489
H6a	3.50	0.450	9,246	6,865−11,660
H7	3.33	1.110	8,784	2,989−14,782
H10	2.38	0.390	6,214	4,187−8,268
H11	5.00	1.000	13,370	800−18,898
H13	8.57	1.220	23,550	16,677−30,633
H16+152	3.43	2.120	955	1,846−20,693
H17a	1.50	0.375	3,895	1,974−5,841
U4	6.66	0.722	18,042	14,053−22,110
U4a	5.83	1.138	15,701	9,524−22,076
U5a1	7.60	0.975	20,909	15,465−26,490
U5a2	5.60	0.453	15,046	12,570−17,554
U5b	9.50	1.330	26,281	18,707−34,097

[1] Average number of mutation within a phylogenetic lineage.

MDS based on pairwise differences (577−16,023 bp) and Φ_{ST} distances (Annex 1 Table 11, Table 12) yielded a genetic map of Europe, Near and Middle East (Annex 1 Table 10) (Fig. 3.24, Fig. 3.25). In the MDS map clusters of European, Near and Middle East populations are seen: (1) Slavs populations (Poland, Ukraine, Russia (Europe), Czech, Slovakia); (2) Southern European populations (Italy, Spain, Sardinia); (3) Ashkenazi; (4) Lithuanian; (5) Saami; (6) Near East, Middle East populations (Armenia, Iran, Azerbaijan, Georgia, Turkey). Positions of the populations in the MDS map were according to the geographic locations except for Belorussia which is placed more far from the other studied Slavs populations in the MDS plot (Fig. 3.24).

The results of the PCA based on pairwise differences (mtDNA haplotype 577−16,023 bp) between Europe, Near and Middle East populations showed that PC1 explained 62.73% and PC2 24.65% of genetic variation. The studied Lithuanian population is close to Slavs populations in the PCA scatter plot (Fig. 3.26). The results of PCA based on Φ_{ST} distances (mtDNA haplotype 577−16,023 bp) between Europe, Near and Middle East populations shows that PC1 explained 88.02 % and PC2 10.53% of genetic variation (Fig. 3.27).

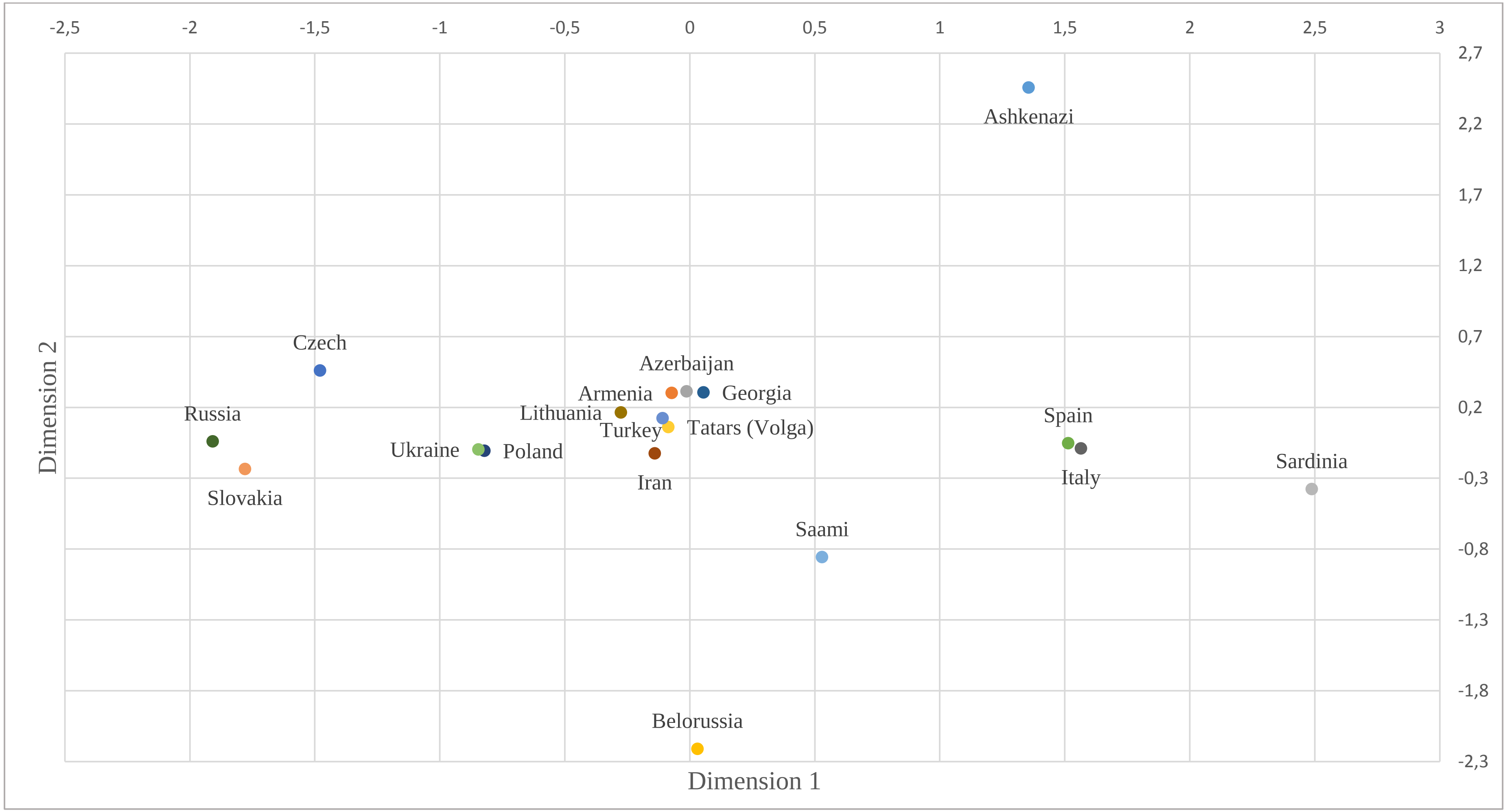


Fig. 3.24. Scatter plot of MDS analysis based on pairwise differences (mtDNA haplotypes 577−16,023 bp) between European populations. MDS stress and correlation coefficient (r^2): stress = 0.12201, r^2 = 0.94647.

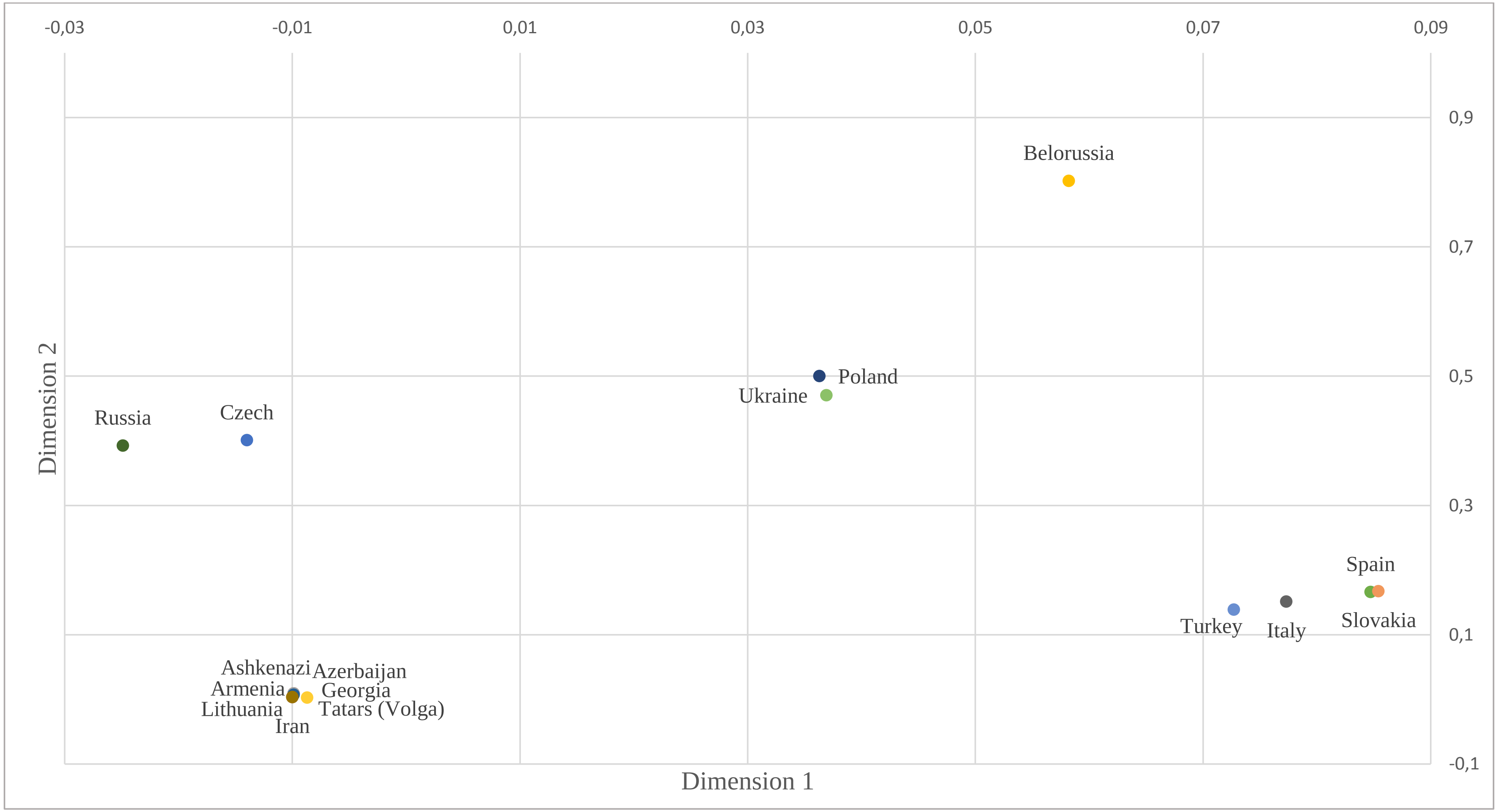


Fig. 3.25. Scatter plot of MDS analysis based on Φ_{ST} distances (mtDNA haplotypes 577−16,023 bp) between European populations (Saami, Sardinia populations removed from analysis). MDS stress and correlation coefficient (r^2): stress = 0.07220, r^2 = 0.99308.

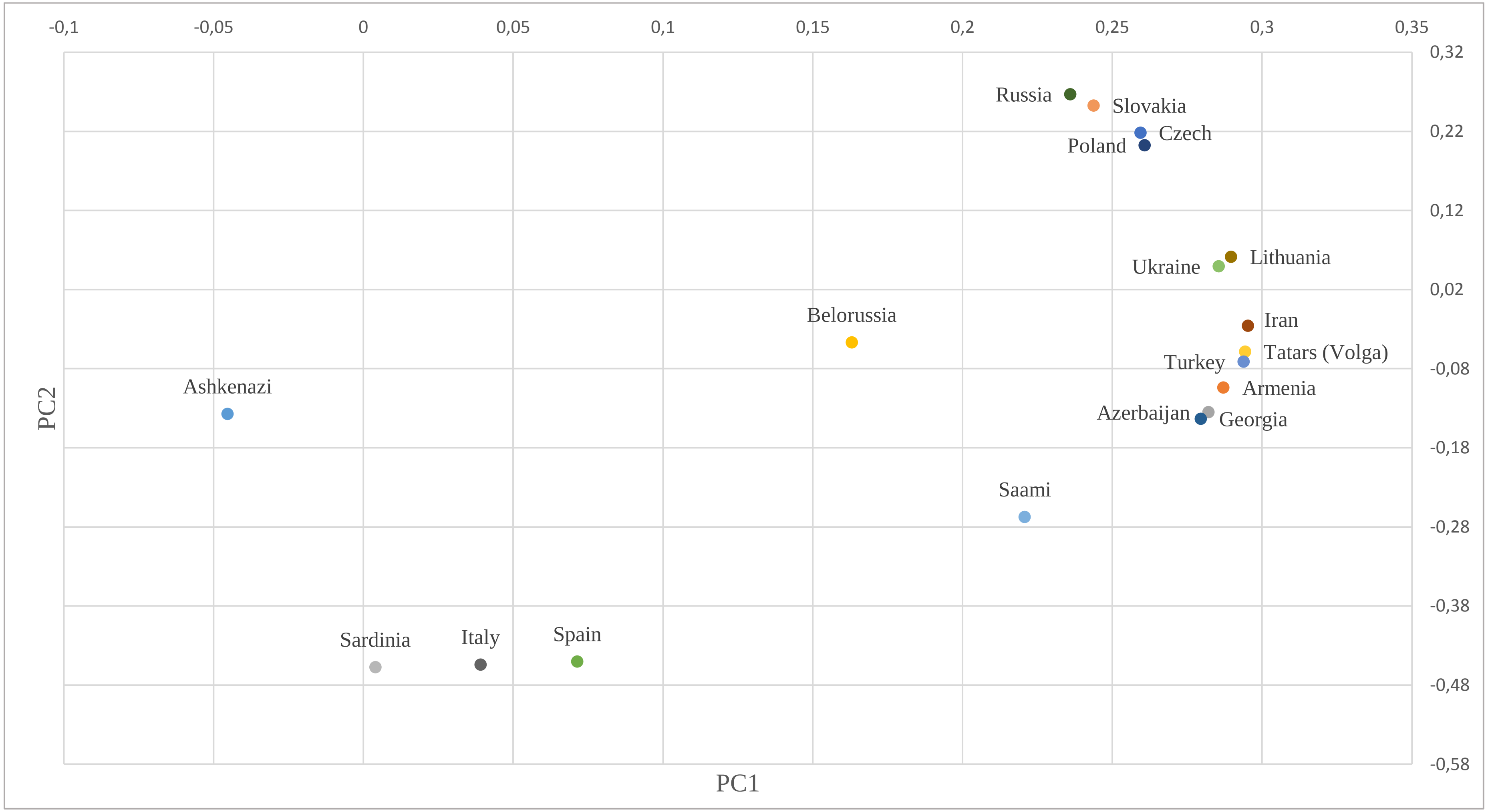


Fig. 3.26. Scatter plot of PCA based on pairwise difference (mtDNA haplotype 577−16,023 bp) between European populations. PC1 explains 62.73%, PC2 24.65% of the genetic variation, i.e. mtDNA haplotype 577−16,023 bp, among the European populations.

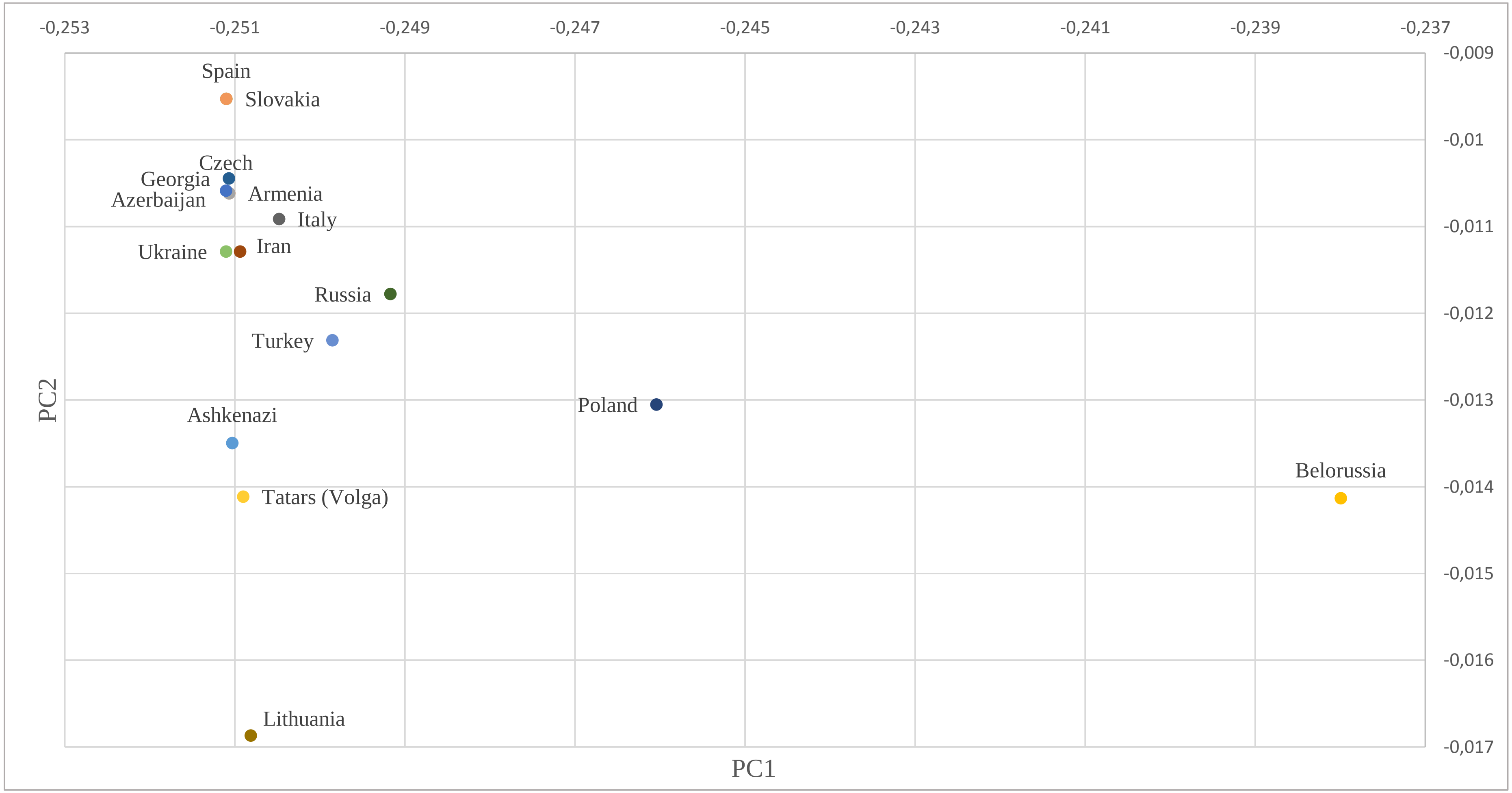


Fig. 3.27. Scatter plot of PCA based on Φ_{ST} distances (mtDNA haplotypes 577−16,023 bp) between European populations (Saami, Sardinia populations removed from analysis). PC1 explains 88.02%, PC2 10.53% of the genetic variation, i.e. mtDNA haplotypes 577−16,023 bp, among the European populations.

SUMMARY

D. Kasperavičiūtė and colleagues studied and described the genetic structure of the Lithuanian population based on the distribution of mtDNA haplogroups and HVRI haplotypes. Also, data on the Lithuanian population were compared with the data on Estonian and Latvian populations (Kasperavičiūtė *et al.* 2004). Other studies in which the population of Lithuania is included analyzed HVRI/ HVRII haplotypes and the main haplogroups of mtDNA (Kushniarevich *et al.* 2013; Lappalainen *et al.* 2008).

The present study for the first time has analyzed complete mtDNA in the largest sample size of six ethno-linguistic groups of Lithuania as compared with the previous studies. The sequencing of complete mtDNA allowed to identify braches of the phylogenetic tree as accurately as possible by the current technologies and analysis methods.

The most frequent phylogenetic lineages identified in this study were H and U (respectively 45.3% and 19.2%). The determined frequencies were close to those from previous studies (respectively 46.1% and 19.5%) (Kasperavičiūtė *et al.* 2004). The phylogenetic lineages T (7.6%), J1 (5.8%) and V (5.8%) together with the most frequent lineages (H and U) comprised 84%, and the rest of mtDNA phylogenetic lineages (A, HV, I, K, M, N, W) comprised 16% of the studied Lithuanian population.

The frequency of the phylogenetic lineage H was close to that of Northern, Central and Eastern Europe (40−50%) (Loogvali *et al.* 2004; Richards *et al.* 2000). The phylogenetic lineage H1 (19.6%) had a similar distribution within Northern, Central and Eastern Europe (10−18%). Furthermore, both phylogenetic lineages (H and H1) are common in Southwest Europe as this part of Europe could be a refugee of phylogenetic lineages H and H1 during LGM. The source region of the phylogenetic lineage H is thought to be Central Europe, and its present distribution in the current populations was reached in the period of the mid Neolite. Indeed, this scenario is still the object to be discussed.

The distribution of the phylogenetic lineage H varied in the range of 18% among six ethno-linguistic groups of Lithuania, the lowest being in West Žemaitija (~35%) and the highest in South Žemaitija and West Aukštaitija (respectively ~53% and ~51%). The previous study showed the lowest frequency in West Žemaitija and South Aukštaitija (each ~40%) and the highest in North and South Žemaitija (each ~53%) (Kasperavičiūtė *et al.* 2004).

The phylogenetic lineage U comprised 19.2% out of which U5 comprises 12.7% of the Lithuanian population studied. The frequency of the phylogenetic lineage U was close to that of Central and Eastern Europe (21%). The phylogenetic lineage U was dominant within hunter-gather populations (~82%) (Bramanti *et al.* 2009), and within the current populations it is detected at a frequency of 12−20%.

The distribution of the phylogenetic lineage U varied also in the range of 18% among six ethno-linguistic groups of Lithuania, the most frequent being in North and West Žemaitija (respectively ~29% and ~23%), and less frequent in South Žemaitija (~11%). The frequency of phylogenetic lineage in the previous study varied within 13.3%, the highest being in South Žemaitija (~27%) and the lowest in North Žemaitija and West Aukštaitija (each ~13%) (Kasperavičiūtė *et al.* 2004). The discrepancy of the frequencies of the phylogenetic lineages H and U in this and in previous studies could be due to a different number of samples, sample sets and the used molecular genetic methods.

Haplotype diversity was even among six ethno-linguistic group of Lithuania. The nucleotide diversity and the mean pairwise haplotype differences based on complete mtDNA were similar except for two ethno-linguistic groups of Lithuania: (1) higher in West Žemaitija; (2) lower in South Žemaitija. Similar results were obtained based on the HVRI region. These results indicated that there were more differences among individuals in West Žemaitija and less in South Žemaitija as compared with the rest ethno-linguistic groups. Statistically significant Tajima`s D value ($P < 0.05$) indicated the possible growth of all six ethno-linguistic groups of Lithuania. A complete mtDNA analysis showed that some regions were outlying for the number of segregating sites, transitions, transversions, synonymous and nonsynonymous sites. The control region and *MT-ATP6* gene had more segregating sites and transitions, whereas *MT-RNR1* (12S), *MT-RNR2* (16S), tRNR, *MT-CO1*, *MT-ND4L* genes had less segregating sites and transitions. Milder differences were seen for transversions, and this could be due to a possible negative effect on the phenotype. Moreover, the ratio of nonsynonymous polymorphisms per nonsynonymous site to synonymous polymorphisms per synonymous site (p_N/p_S) varied across the analyzed regions. More nonsynonymous polymorphisms in *MT-ATP6* gene and less in the *MT-ND6* gene were observed. The higher p_N/p_S ratio in the *MT-ATP6* gene is constant across several populations, whilst the lower p_N/p_S ratio varies in different populations (Gunnarsdottir *et al.* 2011; Schonberg *et al.* 2011). It is more likely that the higher p_N/p_S ratio can be more specific of young phylogenetic lineages (Kivisild *et al.* 2006) than of populations from different climatic zones (Mishmar *et al.* 2003).

Statistically significant F_{ST} distances ($P < 0.05$) based on the distribution of 69 haplogroups were seen between East Aukštaitija and South Žemaitija, North Žemaitija and South Žemaitija. These results could be due to the largest difference of frequencies of the main phylogenetic lineages (H and U) within South Žemaitija.

Results of AMOVA based on the distribution of 23 haplogroups and the mean pairwise differences of haplotypes (16,569 bp) were statistically significant when the ethno-linguistic groups of Lithuania were clustered according to West−East geographic locations in the territory of Lithuania. These results were supported by the analysis of genetic barriers based on the distribution of haplotypes (577−16,023 bp).

The scatter plot of MDS based on the distribution of 23 or 69 haplogroups did not reveal any clusters of ethno-linguistic groups or clusters not supported by results of other analyses. The scatter plot of PCA based on the distribution of haplotypes (577−16,023 bp) showed that the PC1 explained the variation (West−East axis) between South Žemaitija (Middle), East and West Aukštaitija (East), and the rest of ethno-linguistic groups (West−South). The variation explained by PCs among the ethno-linguistic groups accounts for less than 0.5% of all genetic variation.

The phylogenetic tree of the phylogenetic lineage H based on haplotype variation (577−16,023 bp) showed a star-like structure within H1. This could be a sign of the recent population growth in the period of the mid-Neolitic. The source of this phylogenetic lineage may be Central Europe. The phylogenetic tree of the phylogenetic lineage U based on haplotype variation (577−16,023 bp) showed a more scattered structure with longer braches, except for the U5a2. These results have not confirmed the recent growth as it is known that phylogenetic lineages were dominant in the first inhabitants of the territory of Lithuania (Bramanti *et al.* 2009). It is possible that the first settlers were forced to migrate to the West as the frequency of the phylogenetic lineage U remained to be the highest in this part of Lithuania. The reason for a possible migration was the expansion of the farmer culture which could reach the territory of Lithuania from Central and Eastern Europe. This scenario did not explain the complexity of the formation and evolution of maternal lineage in the territory of present-day Lithuania.

The dominance of the phylogenetic lineage U in the first settlers in the territory of the present-day Lithuania was supported by the calculated age of phylogenetic lineages. The age of haplogroups identified within the phylogenetic lineage U varied between 15,000 and 26,000 YBP. The recent growth of the phylogenetic lineage H was supported by the calculated age of haplogroups within this phylogenetic lineage. Most of them (H, H1a, H1b, H1c, H1e, H2a, H6a, H7, H10, H11, H16+152, H17a) fit in the period after LGM, and part of them (H, H1a, H1c, H2a, H16+152, H17a) fit in the period after starting of the Baltic tribe formation, i.e. 5,000 YBP.

The limited number of complete mtDNA sequences from various populations in the databases restricted the analysis of Lithuanian and neighboring populations. The results and conclusions were based on several analyzed populations without a possibility to compare them with all neighboring populations. Moreover, complete mtDNA of specific phylogenetic lineages uploaded in the databases could have an influence on the final results of this study.

The position of the Lithuanian population was between Slavs and Middle East populations in the scatter plot of MDS based on haplotype (577−16,023 bp) variation. All populations are clustered according to their geographic locations, except for Belorussia which was more far away from the rest of Slav populations.

The results of PCA based on haplotype (577−16,023 bp) pairwise differences were similar to MDS; only the position of the Lithuanian population in the scatter plot was more close to Slavs.

A lower differentiation level based on mtDNA haplogroup and haplotype data as compared with Y chromosome data was seen also in previous studies (Nasidze *et al.* 2004). This observation induced the hypothesis of a more frequent maternal lineage migration which leads to smaller differences among populations. This hypothesis was disproved by the study which did not observe lower differences based on mtDNA data as compared with the Y chromosome data (Wilder *et al.* 2004). The most recent study showed that the genetic differences between human populations on the global scale are bigger for the Y chromosome than for mtDNA, although the differences are not as large as previously suggested. Moreover, the same study found substantial regional variation in patterns of mtDNA *versus* Y chromosome variation (Lippold *et al.* 2014). Overall, until now there is no reliable explanations of this phenomenon. Still, genetic differences among the populations are due to the interaction of several evolutionary forces such as mutation, migration, natural selection, and genetic drift.

3.3. Genetic structure and genetic distances of Lithuanian population based on the genome-wide autosomal SNP variation

From the Lithuanian population, 253 samples (Fig. 2.5) were genotyped using the Illumina HumanOmniExpress-12v1.1 arrays (719,666 SNPs). After a systematic primary and secondary quality control (QC) of the generated data, 590,665 autosomal SNPs and 188 samples remained for the subsequent analyses (123 males and 65 females) (Fig. 3.28).

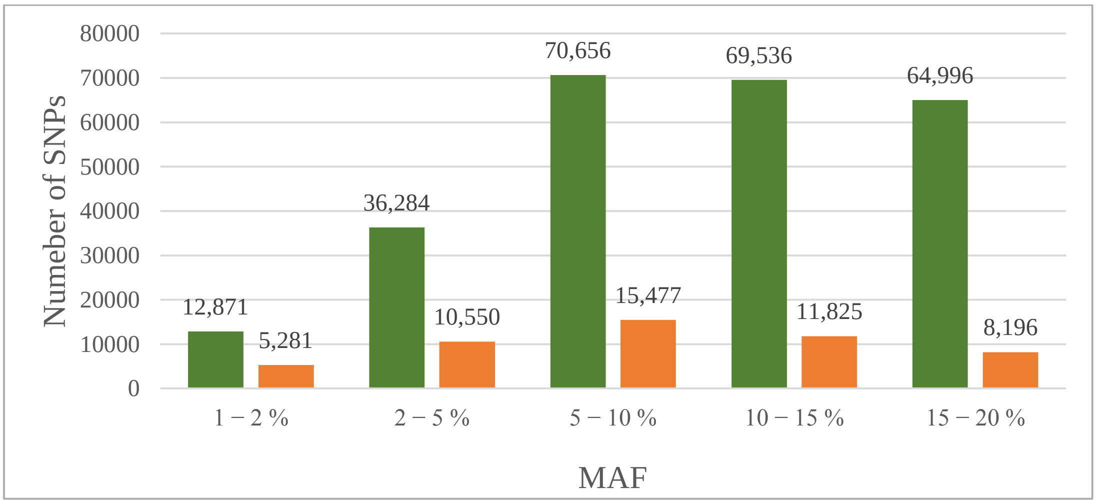


Fig. 3.28. Distribution of SNPs according to MAF before and after QC. Green bars show number of SNPs before removal based on pairwise LD (590,665 SNPs). Orange bars show number of SNPs after removal of SNPs based on pairwise LD (105,387 SNPs). SNPs found to be MAF < 0.01 were removed after QC.

An MDS analysis, based on 105,387 autosomal SNPs after removal of marker set according to pairwise LD ($r^2 < 0.2$) in order to remove correlated SNPs, was performed. No obvious sets of clusters could be seen in the scatter plot of MDS of all six ethno-linguistic groups of Lithuania (Fig. 3.29). Two clearly apparent clusters could be seen in the scatter plot of MDS analysis when only samples from North Žemaitija and South Aukštaitija were included (Fig. 3.30). Other clusters of ethno-linguistic groups that can be distinguished in the scatter plot of MDS analyses are presented in Fig. 3.31, Fig. 3.32, Fig. 3.33, Fig. 3.34. No apparent clusters can be seen in the scatter plot of MDS analysis including samples from Žemaitija only.

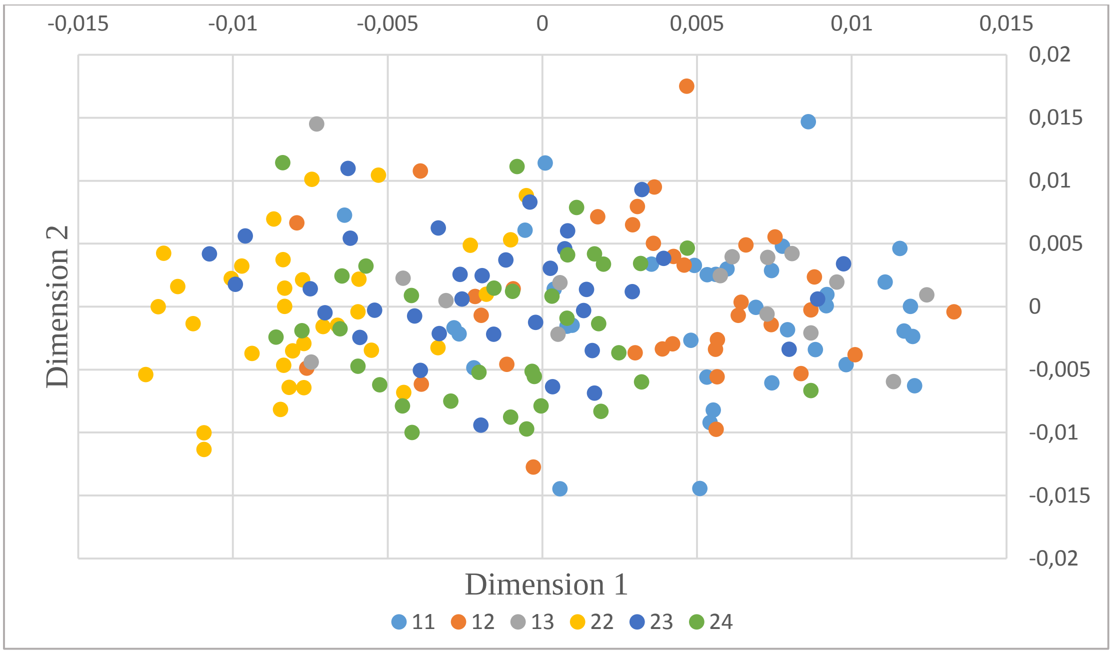


Fig. 3.29. Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 11 – North Žemaitija, 12 – South Žemaitija, 13 – West Žemaitija, 22 – South Aukštaitija, 23 – West Aukštaitija, 24 – East Aukštaitija.

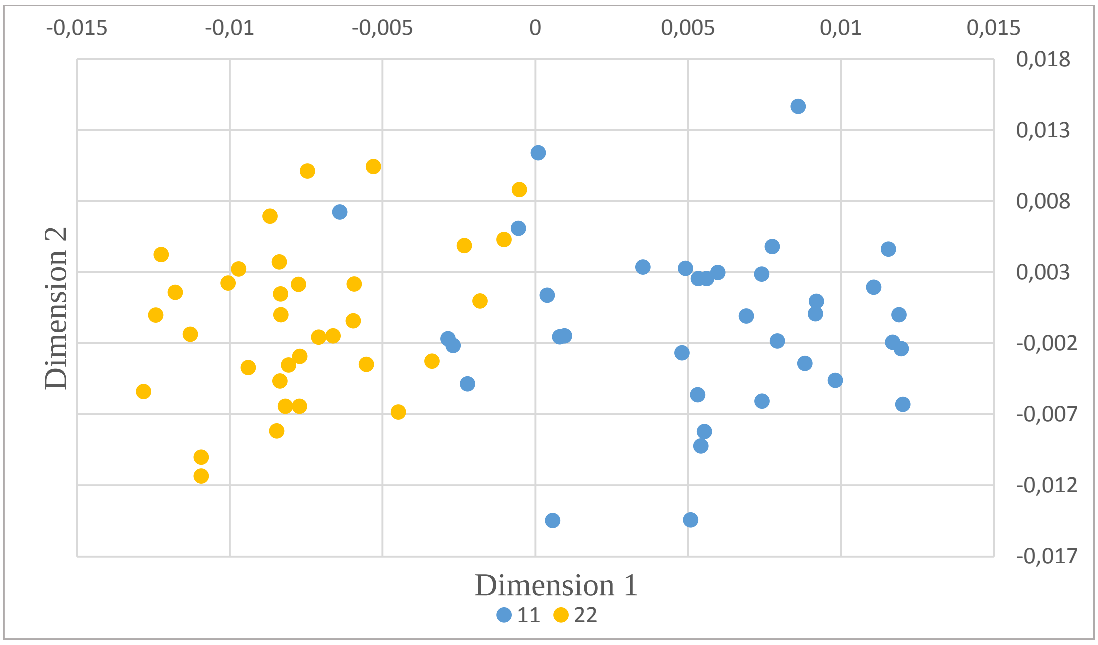


Fig. 3.30. Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 11 – North Žemaitija, 22 – South Aukštaitija.

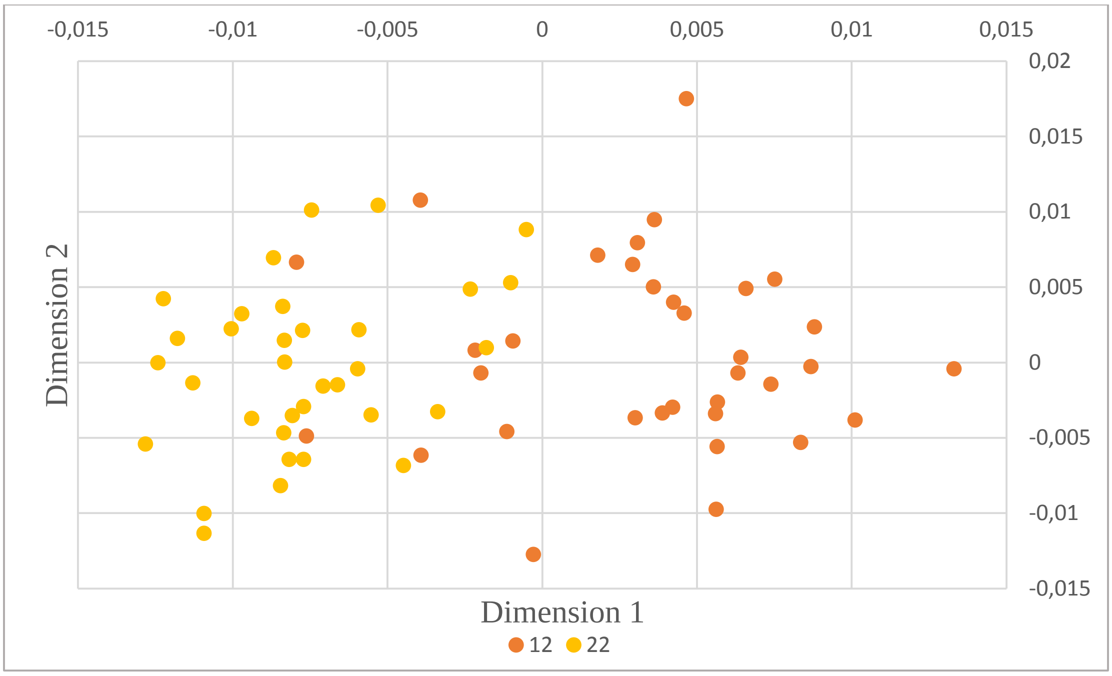


Fig. 3.31. Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 12 – South Žemaitija, 22 – South Aukštaitija.

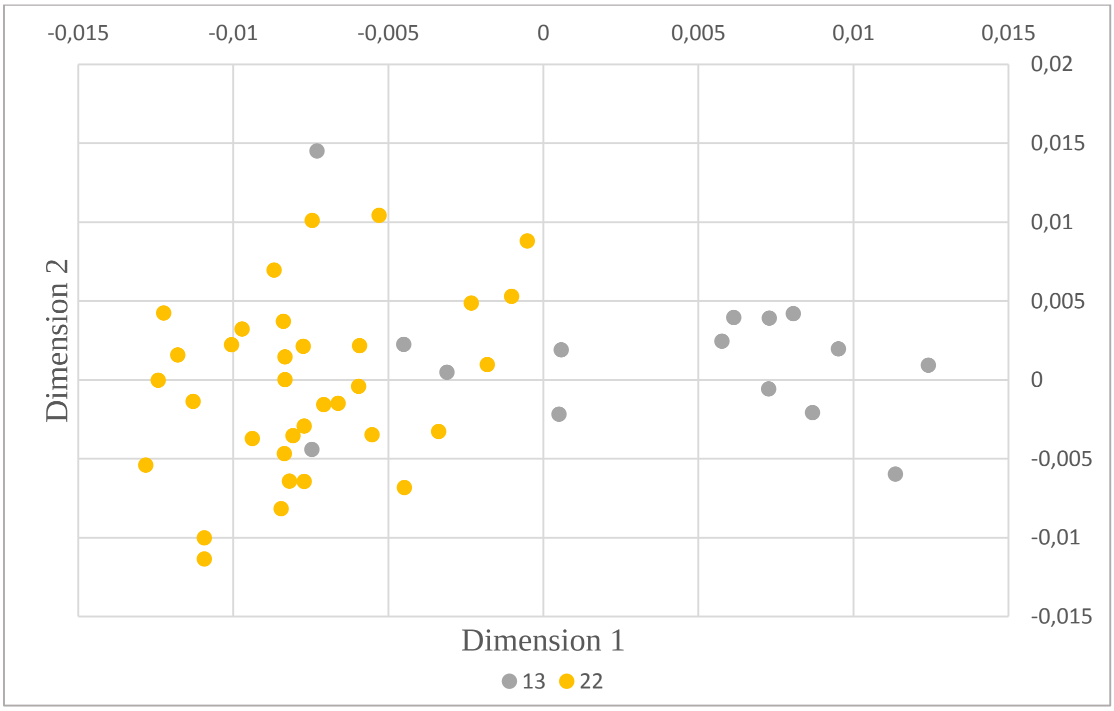


Fig. 3.32 Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 13 – West Žemaitija, 22 – South Aukštaitija.

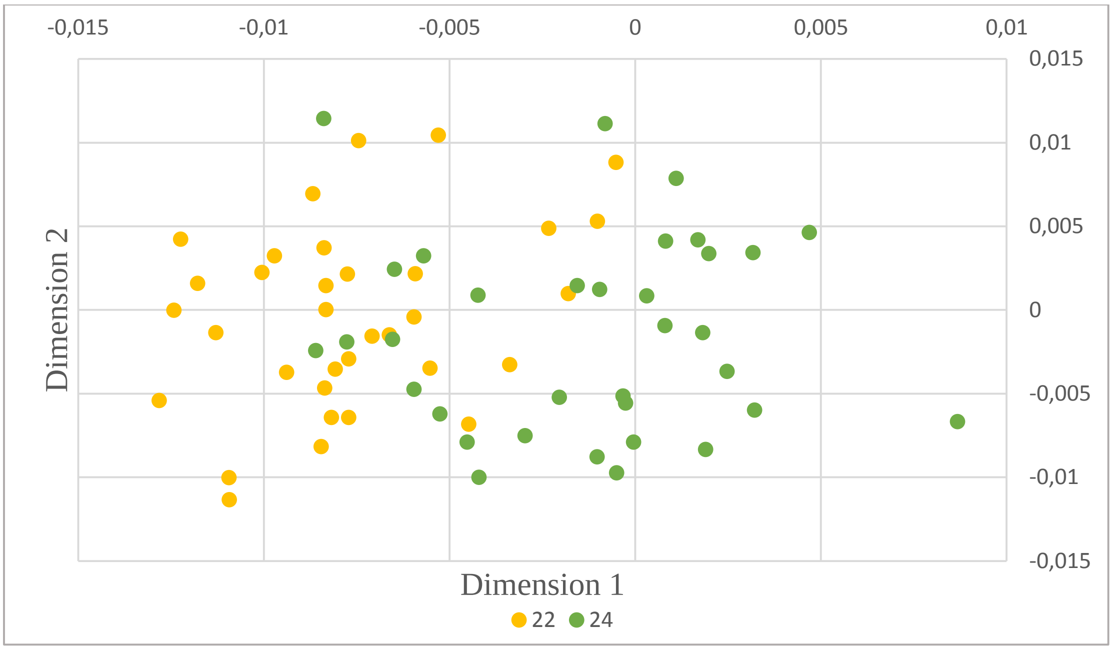


Fig. 3.33 Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 22 – South Aukštaitija, 24 – East Aukštaitija.

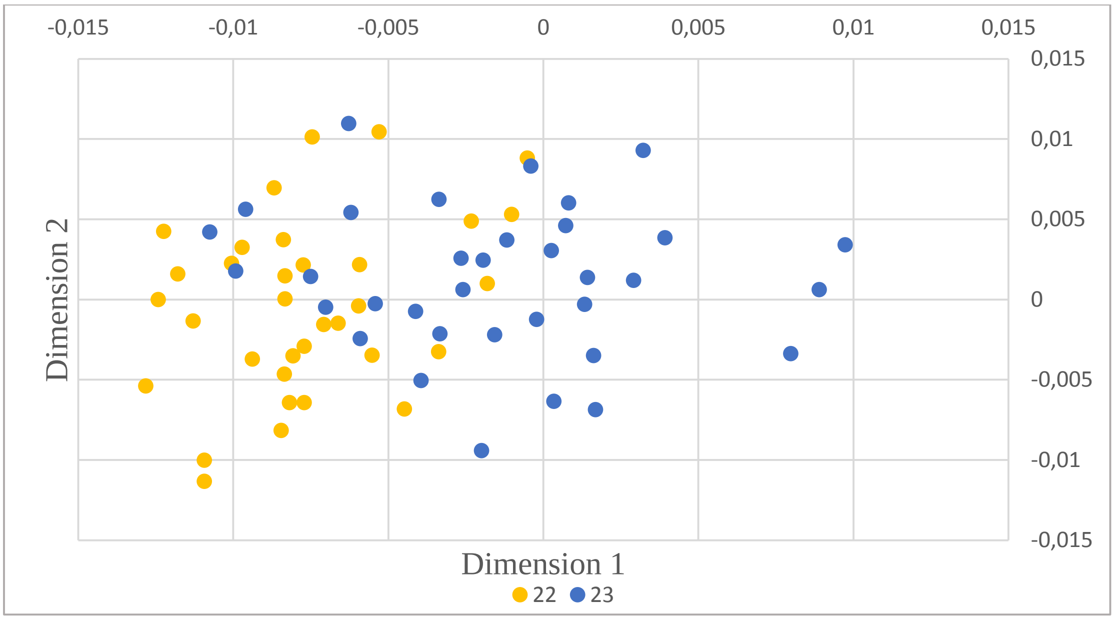


Fig. 3.34 Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 22 – South Aukštaitija, 23 – West Aukštaitija.

Samples from South Aukštaitija, East or West Aukštaitija and North Žemaitija in the scatter plot of MDS analysis were positioned according to their North-South locations on the territory of Lithuania (Fig. 3.35).

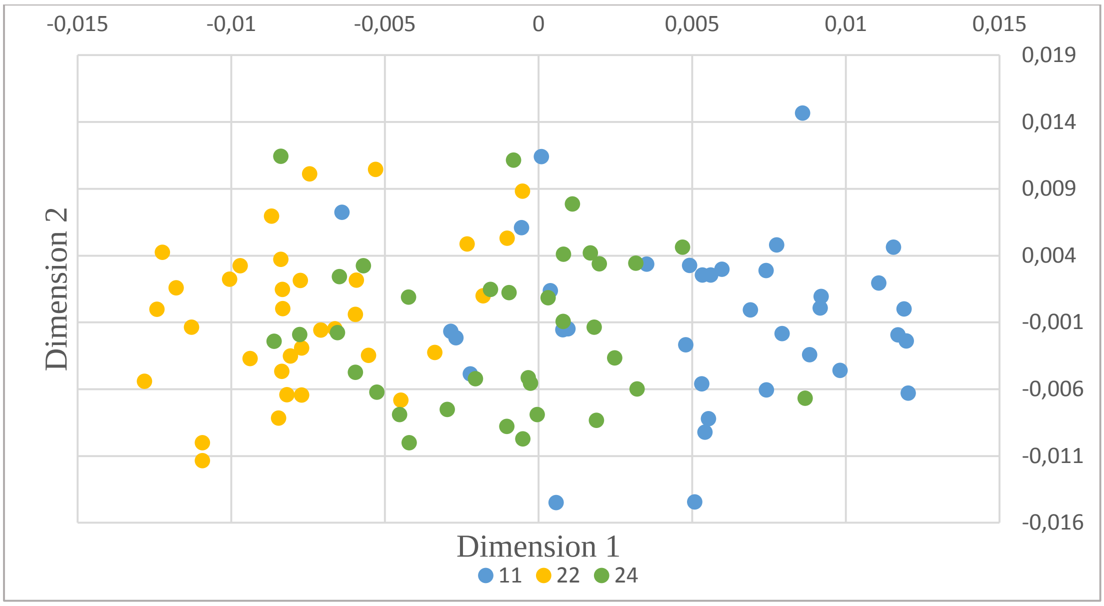


Fig. 3.35. Scatter plot of MDS analysis based on 105,387 autosomal SNPs. Abbreviations: 11 – North Žemaitija, 22 – South Aukštaitija, 24 – East Aukštaitija.

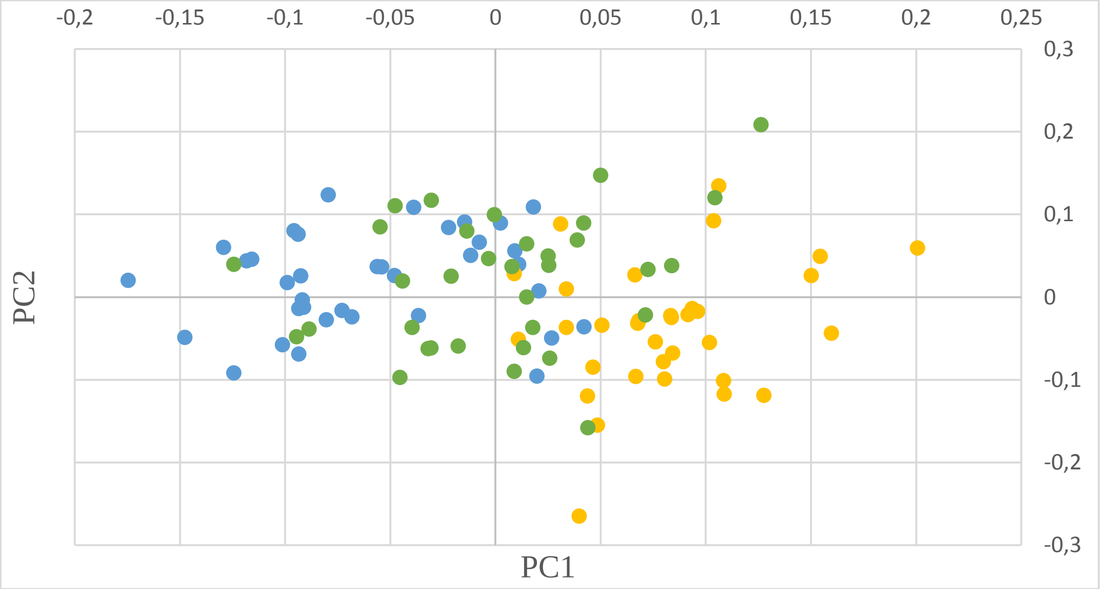


Fig. 3.36. Scatter plot of PCA based on 105,387 autosomal SNPs. PC1 explains 0.61%, PC2 0.59% of genetic variation between 188 samples from six ethno-linguistic groups of Lithuania. Blue label – North Žemaitija, yellow label – South Aukštaitija, green label – East Aukštaitija.

The results of PCA based on 105,387 autosomal SNPs revealed that PC1 explained 0.61% and PC2 – 0.59% of genetic variation among the studied samples (188) of the Lithuanian population. According to the scatter plot of PCA, the greatest differentiation was found among samples from the northwest and southeast of the Lithuanian population (Fig. 3.36). The first PC was statistically significant (Tracy−Widom test, $P < 0.05$).

The most variable SNPs from the statistically significant first PC were identified. Locations of the first 12 most variable SNPs were intragenic. Only six out of the 20 most variable SNPs were in the intronic region of the gene (Table 3.10).

Table 3.10. List of the most variable SNPs from the statistically significant first PC

SNP ID[1]	chr.[3]	Position	SA	WA	EA	SŽ	WŽ	NŽ
rs806269	9		0.3088 (A)[4]	0.1429 (A)	0.07353 (A)	0.1176 (A)	0.06667 (A)	0.02778 (A)
rs2247168	11		0.3824 (A)	0.4143 (G)	0.3971 (G)	0.3235 (G)	0.4 (G)	0.3889 (G)
rs927694	13		0.3676 (A)	0.5 (C)	0.4853 (A)	0.4091 (C)	0.4 (C)	0.4444 (C)
rs2535584	17		0.3971 (A)	0.4714 (G)	0.4265 (A)	0.4848 (A)	0.4333 (A)	0.4167 (G)
rs1162608	9		0.3529 (G)	0.2857 (G)	0.2941 (G)	0.2206 (G)	0.06667 (G)	0.1806 (G)
rs645121	9		0.3529 (A)	0.4714 (A)	0.4265 (C)	0.4559 (C)	0.5 (A)	0.4861 (A)
rs2193769	5		0.3088 (G)	0.2143 (G)	0.2794 (G)	0.1618 (G)	0.1667 (G)	0.06944 (G)
rs17122441	14		0.3088 (A)	0.1286 (A)	0.07576 (A)	0.1029 (A)	0.1333 (A)	0.06944 (A)
rs10742969	11		0.1176 (A)	0.05714 (A)	0.01471 (A)	0 (A)	0 (A)	0.02778 (A)
rs17746916	10		0.2059 (A)	0.2857 (A)	0.2941 (A)	0.3088 (A)	0.5 (A)	0.4444 (A)
rs10769684	11		0.3382 (G)	0.3571 (G)	0.3971 (G)	0.4412 (A)	0.4333 (A)	0.3889 (A)
rs631844	5		0.25 (G)	0.3571 (G)	0.3971 (G)	0.4853 (G)	0.5 (G)	0.4861 (A)
rs221748[2]	6	*PDE10A* (intron)	0.2647 (A)	0.1714 (A)	0.1029 (A)	0.07353 (A)	0.1333 (A)	0.1111 (A)
rs6131629	20	*MACROD2* (intron)	0.1029 (A)	0.2143 (A)	0.1667 (A)	0.2353 (A)	0.3 (A)	0.2639 (A)
rs6997954	8		0.2941 (G)	0.2286 (G)	0.1324 (G)	0.1471 (G)	0.1667 (G)	0.05556 (G)
rs4779631	15	*RYR3* (intron)	0.3382 (A)	0.4714 (A)	0.4706 (G)	0.4412 (G)	0.4333 (G)	0.4444 (G)
rs9609889	22		0.1765 (G)	0.07143 (G)	0.1176 (G)	0.08824 (G)	0 (G)	0.02778 (G)
rs9511987	13	*ATP8A2* (intron)	0.2647 (A)	0.4143 (A)	0.4242 (A)	0.4265 (G)	0.4333 (G)	0.3889 (A)
rs6007216	22	*LINC00229* (intron)	0.3824 (A)	0.2143 (A)	0.1176 (A)	0.2353 (A)	0.2333 (A)	0.1667 (A)
rs6733159	2	*PARD3B* (intron)	0.1324 (C)	0.2429 (C)	0.2059 (C)	0.2794 (C)	0.2333 (C)	0.2361 (C)

[1] Identification number of SNP according to *NCBI dbSNP*.

[2] Hardy–Weinberg (P = 0.03084).

[3] Number of the chromosome.

[4] MAF and allele in the Lithuanian population studied.

The F_{ST} distances based on 105,387 autosomal SNPs between six ethno-linguistic groups of Lithuania were calculated. The significance of ANOVA for six ethno-linguistic groups of Lithuania along PC1 is present in Table 3.11. The geographical and F_{ST} distances were statistically significantly correlated (Mantel test, r = 0.60; *P* = 0.0127; 10,000 permutations).

Table 3.11. F_{ST} distances based on 105,387 autosomal SNPs (below the diagonal). *P*-values of ANOVA statistics for population differences along PC1 (above the diagonal)

	SA	**NŽ**	**WŽ**	**EA**	**WA**	**SŽ**
SA	0	**3.33×10^{-11}**	**1.66×10^{-06}**	**2.84×10^{-07}**	**3.36×10^{-03}**	**1.11×10^{-11}**
NŽ	0.000909	0	0.2082	**4.09×10^{-03}**	**2.06×10^{-02}**	**0.0343**
WŽ	0.000795	0.00024	0	**0.0069**	**0.0092**	0.8915
EA	0.000499	0.000347	0.000488	0	0.5425	**0.0001**
WA	0.00017	0.000336	0.000282	0.000116	0	**0.0002**
SŽ	0.000717	0.000081	0.00025	0.000403	0.000143	0

Barriers of the genome-wide genetic variation, i.e. 105,387 autosomal SNPs, among six ethno-linguistic groups of Lithuania were identified. The first barrier was found between the South Aukštaitija and the rest of Lithuanian population. The second barrier was seen between the West Žemaitija and the rest of Lithuanian population. The third barrier separated the two main groups of Lithuania, i.e. Aukštaitija and Žemaitija (Fig. 3.37).

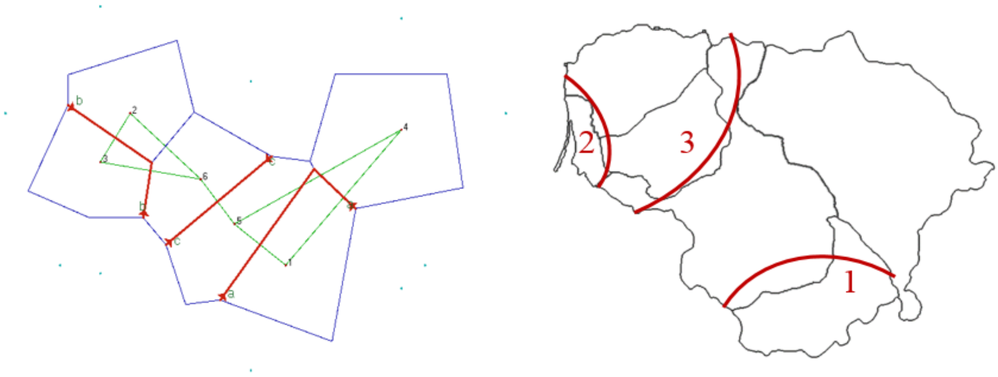


Fig. 3.37. Barriers of genetic variation i.e. 105,387 autosomal SNPs, among six ethno-linguistic groups of Lithuania. Left: output of the software, i.e. a schematic map of the ethno-linguistic groups of Lithuania according to submitted coordinates (Annex 1, Table 1). Right: barriers were numbered in the order of priority. Analysis was based on F_{ST} distances.

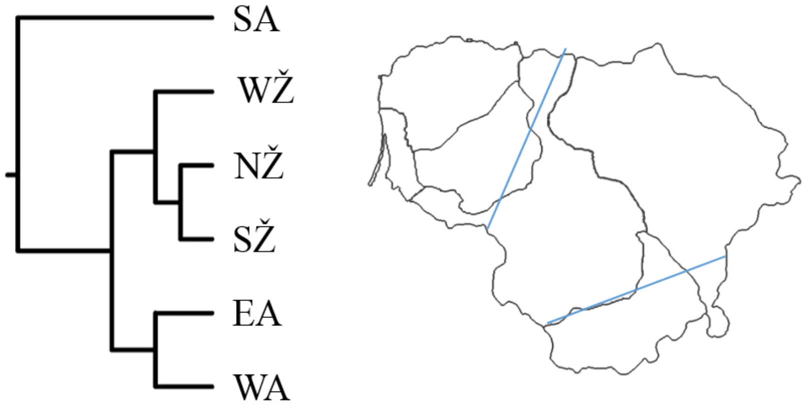


Fig. 3.38. The UPGMA tree based on F_{ST} distances. Right: lines on the map show boundary of the clusters.

The UPGMA clustering based on F_{ST} distances (105,387 autosomal SNPs) was performed. Three clusters of populations could be seen in the constructed UPGMA tree: (1) North, West, South Žemaitija; (2) West, East Aukštaitija; (3) South Aukštaitija. The length of the South Aukštaitija branch reflected the higher genetic differentiation level of this ethno-linguistic group (Fig. 3.38).

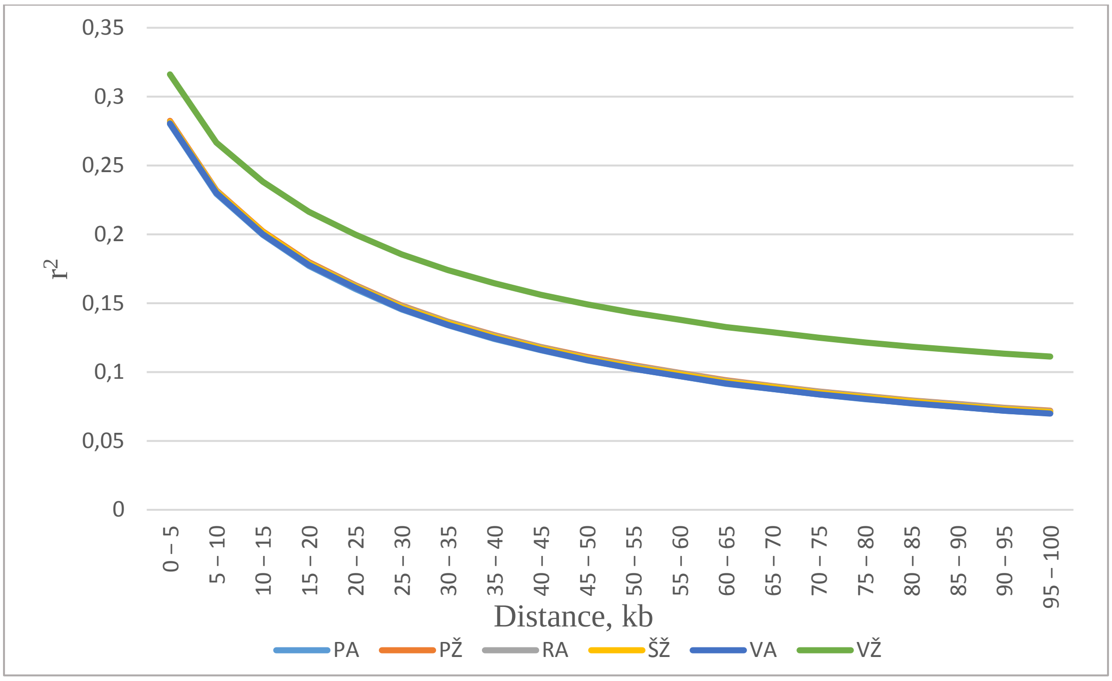


Fig. 3.39. The genome-wide LD pattern (590,665 autosomal SNPs) within Lithuanian population.

The pair-wise LD between SNPs was evaluated as the measurement of the mean of r^2 statistics. The average r^2 statistics were calculated within the inter-marker distances of 5 kb, i.e. 0−5 kb, 5−10 kb, etc. The patterns of the LD structure were quite similar for all ethno-linguistic groups of Lithuania, except for West Žemaitija. The average r^2 values along all distance intervals were higher as compared with the rest of the ethno-linguistic groups (Fig. 3.39). The average r^2 values along all intervals between West Žemaitija and the remaining Lithuanian ethno-linguistic groups differed statistically significantly (t test for independent samples, $P < 0.05$).

An MDS analysis based on 106,545 autosomal SNPs (LD $r^2 < 0.2$) was performed. Population clusters in the MDS scatter plot corresponded to the geographic origin of the populations: (1) East Asia (CHB − Han Chinese in Bejing, China (84 samples); JPT − Japanese in Tokyo, Japan (86 samples); CHD − Chinese in Metropolitan Denver, Colorado (85 samples)); (2) Africa (YRI − Yoruba in Ibadan, Nigeria (78 samples); LWK − Luhya in Webuye, Kenya (88 samples); ASW − African Ancestry in Southwest US (46 samples), MKK − Maasai in Kinyawa, Kenya (118 samples)); (3) Europe (CEU − Utah Residents with Northern and Western European Ancestry (84 samples), TSI − Toscani in Italy (88 samples); LIT – Lithuanian population (188 samples)); (4) North America and South Asia (MXL/MEX − Mexican Ancestry in Los Angeles, California (31 samples), GIH − Gujarati Indians in Houston, Texas (85 samples)). The positions of MXL/MEX and GIH populations in the scatter plot of MDS analysis were in between the groups of Asian and European origin. The Lithuanian population was in the shortest proximity with the other populations of European origin (CEU). The cluster of the African descent populations was the most scattered (Fig. 3.40).

PCA based on genetic variation of 106,545 autosomal SNPs (LD $r^2 < 0.2$) among samples from Lithuanian and HapMap3 populations (1,126 samples) was performed. PC1 explained 5.71% and PC2 3.62% of genetic variation. The scatter plot generated by PCA was very similar to that generated by MDS analysis (Fig. 3.41). The first 38 PCs were statistically significant (Tracy−Widom test, $P < 0.05$).

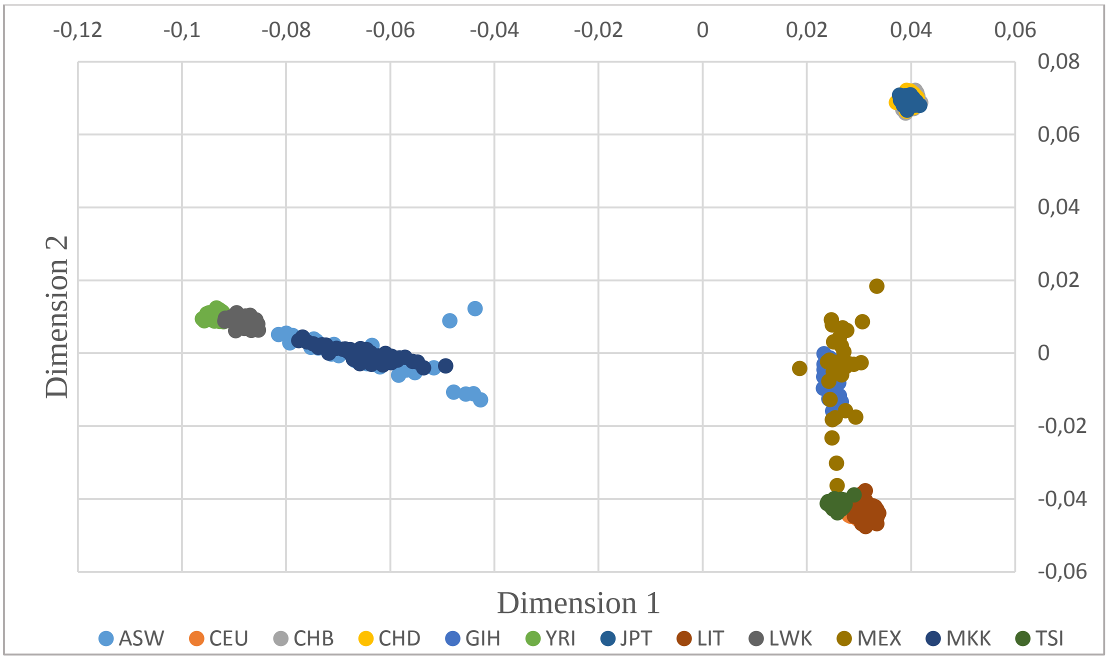


Fig. 3.40. Scatter plot of MDS analysis based on the genetic variation of 106,545 autosomal SNPs between Lithuanian and HapMap3 populations (1,126 samples). Abbreviations: Europe (CEU, TSI, LIT), East Asia (CHB, JPT, CHD), Africa (YRI, LWK, ASW, MKK), North America (MXL/MEX), South Asia (GIH). Detailed descriptions of the samples in the text and abbreviations.

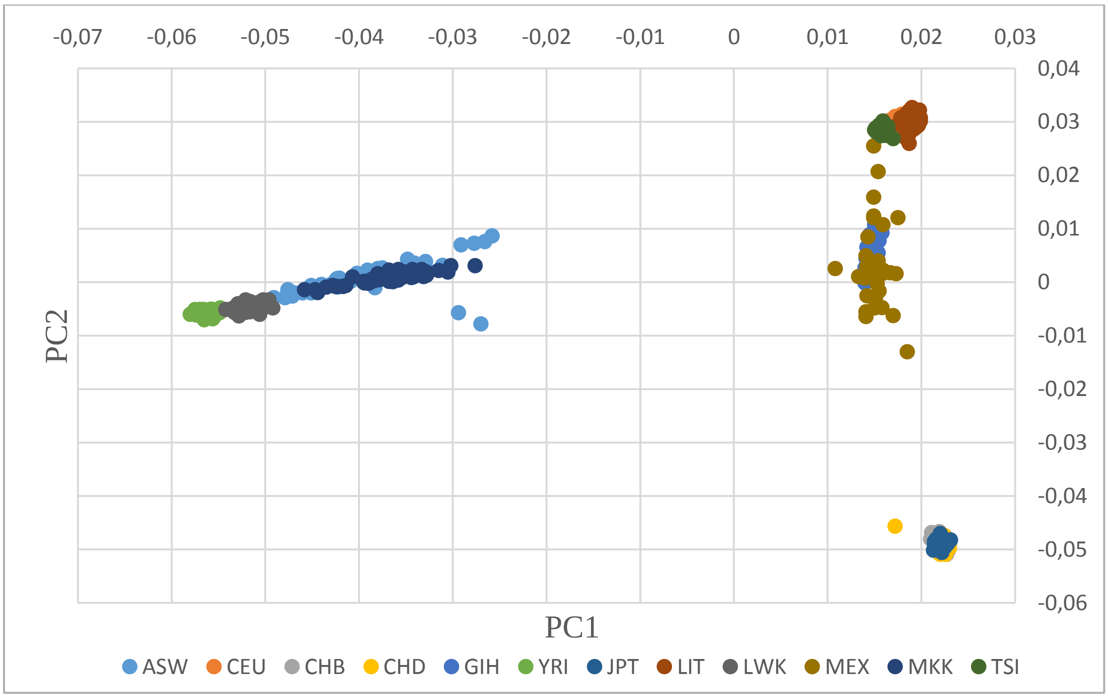


Fig. 3.41. Scatter plot of PCA based on the genetic variation of 106,545 autosomal SNPs between samples from Lithuanian and HapMap3 populations (1,126 samples). PC1 explained 5.71% and PC2 3.62% of genetic variation.

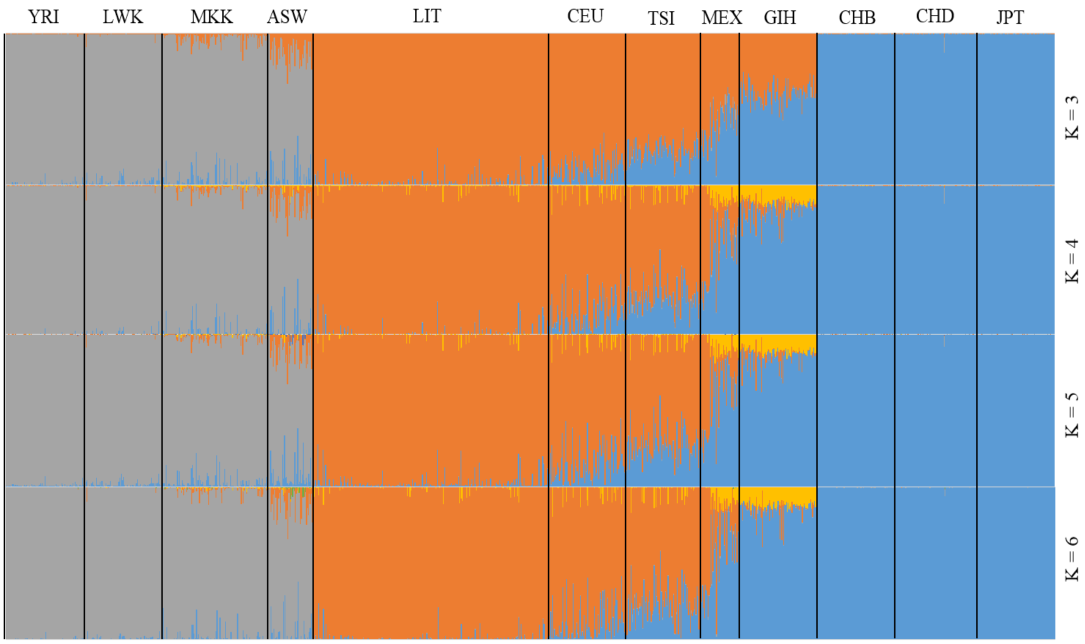


Fig. 3.42. Model-based population structure analysis based on Lithuanian and HapMap3 data (K = 2−6).

The most variable SNPs (1,569 SNPs; LD $r^2 < 0.2$) from the first 100 PCs were selected for the subsequent model-based analysis of Lithuanian and HapMap3 populations' structure (Huckins *et al.* 2014; Tian *et al.* 2008). The number of unknown ancestral populations was determined empirically (K = 2−6) (Annex 1, Fig. 1). The model-based structure analysis (K = 4) showed that even a small number of well selected SNPs enabled to distinguish groups of the same origin (Africa, East Asia, North America, South Asia, Europe) (Fig. 3.42). The model of four unknown ancestral populations (K = 4) was most suitable for group of European populations. Within the group of European descent populations the highest percentage of the genome of East Asian origin was seen in the TSI and the lowest in the Lithuanian population. The model (K = 4) was least suitable for the group of East Asian origin populations as it was impossible to distinguish populations within this group. The results of the model-based (K = 4) population structure analysis were similar to the results of PCA (Fig. 3.41).

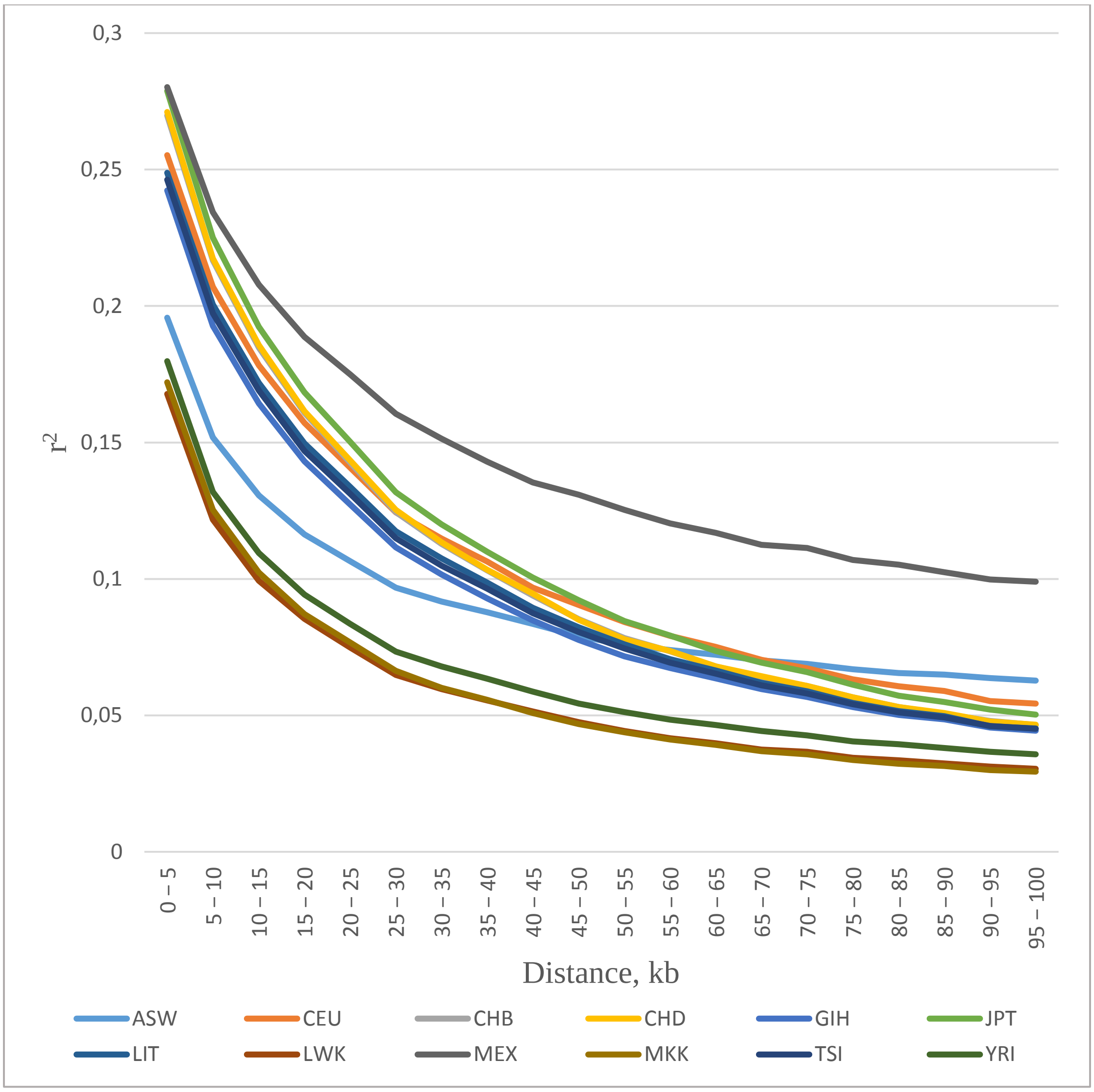


Fig. 3.43. The genome-wide LD pattern (202,997 autosomal SNPs) between Lithuanian and HapMap3 populations (1,126 samples).

Pair-wise LD between SNPs in the Lithuanian and HapMap3 populations was evaluated as a measurement of the mean of r^2 statistics. The average r^2 statistics was calculated within inter-marker distances of the 5 kb, i.e. 0–5 kb, 5–10 kb, etc. The lowest average r^2 values and the shortest LD blocks were seen within the group of African origin populations (YRI, LWK, MKK). The LD structure of the ASW population was somehow different from the other African populations, i.e. the LD blocks were longer. The highest average r^2 values and the longest LD blocks were found within the North American origin population (MXL/MEX). Populations of East Asian (CHB, JPT, CHD), European (CEU, TSI, LIT), and South Asian (GIH) descent showed intermediate average r^2

values. The average r^2 values of these populations were more similar within shorter intervals, and the extent of difference was seen in the longer intervals. The LD pattern of the Lithuanian population was closer to the LD structure of the TSI population than to the other populations of European origin (CEU). The average r^2 values between Lithuanian and North American (MXL/MEX), African (YRI, MKK, LWK) populations along all intervals differed statistically significantly (*t* test for independent samples, $P < 0.05$) (Fig. 3.43).

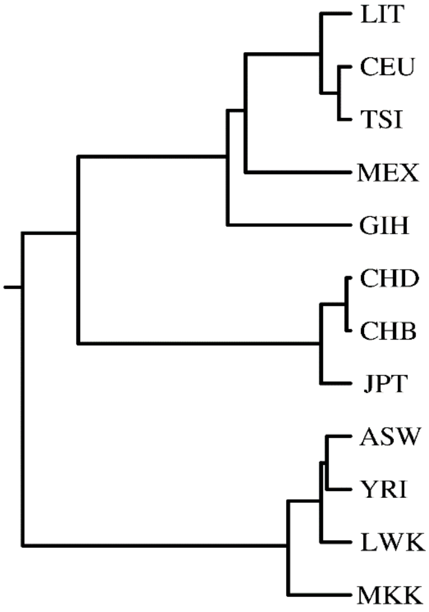


Fig. 3.44. The UPGMA tree based on F_{ST} distances (106,545 autosomal SNPs) between Lithuanian and HapMap3 populations (1,126 samples).

The UPGMA clustering based on F_{ST} distances (Annex 1, Table 15) was performed. Five clusters of populations could be seen in the constructed UPGMA tree: (1) Europe (CEU, TSI, LIT); (2) North America (MXL/MEX); (3) South Asia (GIH); (4) East Asia (CHB, JPT, CHD); (5) Africa (YRI, LWK, ASW, MKK) (Fig. 3.44).

SUMMARY

M. Nelis with colleagues analyzed Northern and Eastern European populations, including Lithuanian population, based on 273,454 autosomal SNPs. The position of Lithuanian population in the scatter plot of PCA corresponded to its geographical location. The Lithuanian population was close to the neighboring populations (Latvia, Estonia) and slightly distanced from Slavs (Poland, Russia) (Nelis *et al.* 2009). Some studies, which also include samples from population of Lithuania, did not disclose in depth the origin of samples or included only samples of certain ethnic origin. That is why the analyzed samples may not represent a certain population at the appropriate level (Behar *et al.* 2010; Mendizabal *et al.* 2012). The present study for the first time analyzes the genetic structure of six ethno-linguistic groups of Lithuania and as compared with HapMap3 populations based on genome-wide autosomal SNPs data.

High throughput technologies made it possible to genotype thousands of SNPs in hundreds of samples. Systematic quality control (QC) must be applied to the data obtained using high throughput technologies. Systematic QC reduces the number of low quality SNPs and samples and thus can improve the final results. The final dataset after primary and secondary QC consisted of 188 samples and 590,655 SNPs or 105,387 SNPs (SNPs removed based on pair-wise LD, $r^2 < 0.2$).

The scatter plot of MDS showed no distinguishable clusters when all six ethno-linguistic groups of Lithuania were positioned on the genetic map. Analysis of certain ethno-linguistic groups revealed clusters in the scatter plot. The highest differences were seen between North Žemaitija and South Aukštaitija. Similar results were obtained when Y chromosome data were analyzed. The other pairs of ethno-linguistic groups that could be distinguished in the scatter plot of MDS were South Aukštaitija and West Žemaitija, South Aukštaitija and South Žemaitija, South and West Aukštaitija, South and East Aukštaitija. No clusters can be seen among the ethno-linguistic groups of Žemaitija. This means that ethno-linguistic groups of Žemaitija are more homogeneous as compared with the ethno-linguistic groups of Aukštaitija. Analyzing more than one pair of ethno-linguistic groups (i.e. North Žemaitija, East Aukštaitija / West Aukštaitija, South Aukštaitija), the Northwest−Southeast gradient in the first dimension could be distinguished.

The PCA based on 105,387 SNPs yielded a scatter plot similar to that of the MDS analysis. A statistically significant PC1 explained 0.61% of all genetic variations, i.e. both among individuals and ethno-linguistic groups. The explained variation by PC1 is due to the Northwest−Southeast genetic variation gradient. Moreover, most variable SNPs from statistically significant PC1 were selected, and only six were within genes, i.e. introns. These results may be due to the used DNA genotyping kit of common SNPs (MAF > 1%). Most variable SNPs, i.e. not within genes, involved into genomic regulation and should not be excluded.

The UPGMA tree based on F_{ST} distances and analyses of genetic barriers supported the results of MDS and PCA. The Northwest−Southeast gradient of genetic variation can be seen in both analyses.

Analysis of genome-wide LD pattern of six ethno-linguistic groups of Lithuania revealed a similar LD structure except for the West Žemaitija ($P < 0.05$). These results may be influenced by unequal sample sizes of West Žemaitija and the rest of ethno-linguistic groups (respectively 15 and ~34).

The genetic structure based on genome-wide autosomal SNPs, which is not specific of maternal or paternal lineages, showed also the Northwest−Southeast gradient of genetic variation. This may be the main cord along which the gene pool of Lithuanian population migrated.

The lack of genome-wide data available in databases restricted the comparison of Lithuanian population with HapMap3 populations.

The scatter plot of MDS showed that the most divergent (scattered) group was that of African origin. Some of African descent populations (MKK and ASW) were closer to the cluster of European origin. Populations of North American (MXL/MEX) and South Asian (GIH) origin were also divergent and were positioned between clusters of the Asian and European origin. The most homogeneous group was of Asian origin with no possibility to distinguish populations within this cluster. The level of divergence within the cluster of European origin was intermediate between these of Asian and African origin. The position of the Lithuanian population within the European origin cluster was closer to that of the CEU population. This could be explained by the results of previous studies which showed the CEU to represent populations of Northern and Western origin (Lao *et al.* 2008; McEvoy *et al.* 2009). Moreover, it is known that the genetic distance correlates with the geographic distances among European populations (Nelis *et al.* 2009). This may explain why the Lithuanian population was closest to the CEU from all the analyzed HapMap3 populations.

As was expected, the scatter plot of PCA (PC1, PC2) was similar to that of MDS analyses (first two dimensions). Most of PC1 explained variation (5.71%) was due to differences of genetic variation between the cluster of African origin and the rest of analyzed populations, whereas, most of PC2 explained variation (3.62%) was due to differences between the cluster of Asian origin and the rest of analyzed populations.

Analysis of the genome-wide LD pattern of Lithuanian and HapMap3 populations revealed the largest LD blocks in the population of North American origin (MXL/MEX) and the smallest in the cluster of African origin. The LD structure of the Lithuanian populations was close to that of the other populations of European (CEU, TSI) and South Asian (GIH) origin. The LD patterns within the Lithuanian population (West Žemaitija was somewhat different from the rest of ethno-linguistic groups) could influence the results of the LD structure analysis of the Lithuanian and HapMap3 populations.

The UPGMA tree based on F_{ST} distances between Lithuanian and HapMap3 populations supported the results of MDS and PC analyses. The populations that originated from the same continent tend to cluster together. Populations of European origin (CEU, TSI) formed a cluster which was joined by Lithuanian population. The next branch consisted of populations of North American and South Asian origin. The furthest branches were of Asian and African origin.

Most variable SNPs from the first 100 PC were selected for the model-based ancestry analysis. At K = 4, clusters of different origin could be distinguished, i.e. those of African, Asian, European, South Asian, North American descent. Moreover, this model was suitable for separating populations of European origin, while it was not as good for populations of African and Asian origin. The smallest Asian ancestry component was seen in the Lithuanian population as compared with the rest of European origin populations (CEU, TSI). Results of this model-based analysis supported the results of MDS and PC analyses.

3.4. Genetic structure of Lithuanian population based on exome variation

The exome was determined and genetic variation identified in 52 samples of the Lithuanian population (Fig. 2.6). Genetic variants after primary, secondary, tertiary QC and with the ≥10-fold coverage depth were used for subsequent analyses (average quality value (QV) of the reference allele 28 (±2.3), average QV of the new allele 28.4 (±1.8)) (Casals *et al.* 2013). The average coverage depth of the identified genetic variants was 40.5 (±49.2), and the average number of variants per sample ~28,890 (±6,528). In total, 163,203 different genetic variants, of them 58,865 (~36%) singletons, were identified in the Lithuanian population. The number of different genetic variants identified in Aukštaitija (26 samples) was 25,860 and in Žemaitija (26 samples) 33,008.

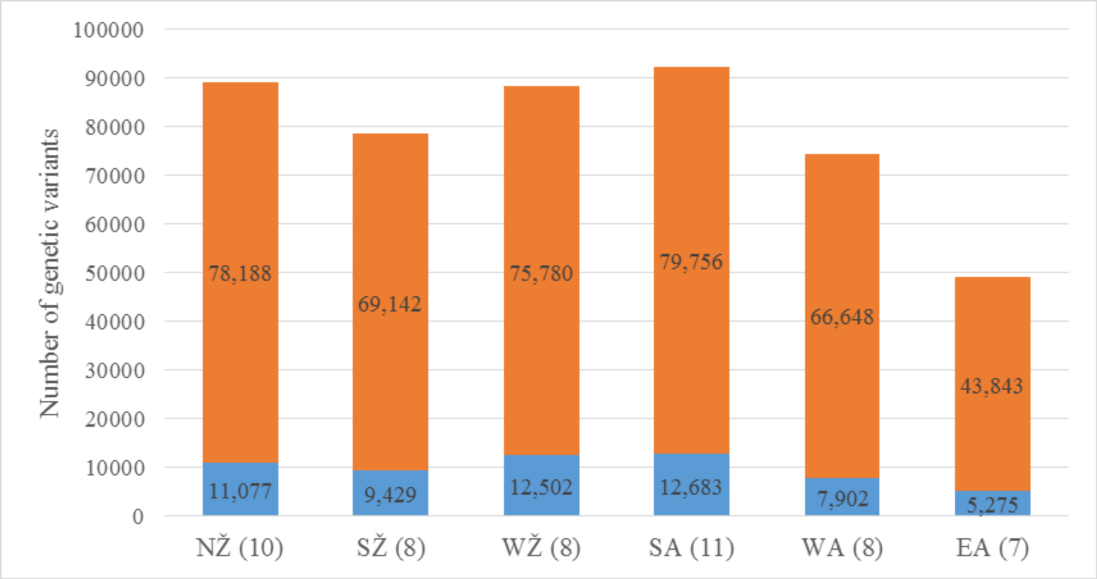


Fig. 3.45. Distribution of different exome variants (orange label) and singletons (blue label) among the ethno-linguistic groups of Lithuania.

The distribution of different exome variants and singletons among the ethno-linguistic groups of Lithuania was estimated. The majority of genetic variants (92,439) were identified in South Aukštaitija (11 samples) and the least (49,118) in East Aukštaitija (7 samples) (Fig. 3.45). The number of singletons per sample was statistically significantly (χ^2 test, $P < 0.05$) greater in West Žemaitija and lower in West and East Aukštaitija than expected (Fig. 3.46).

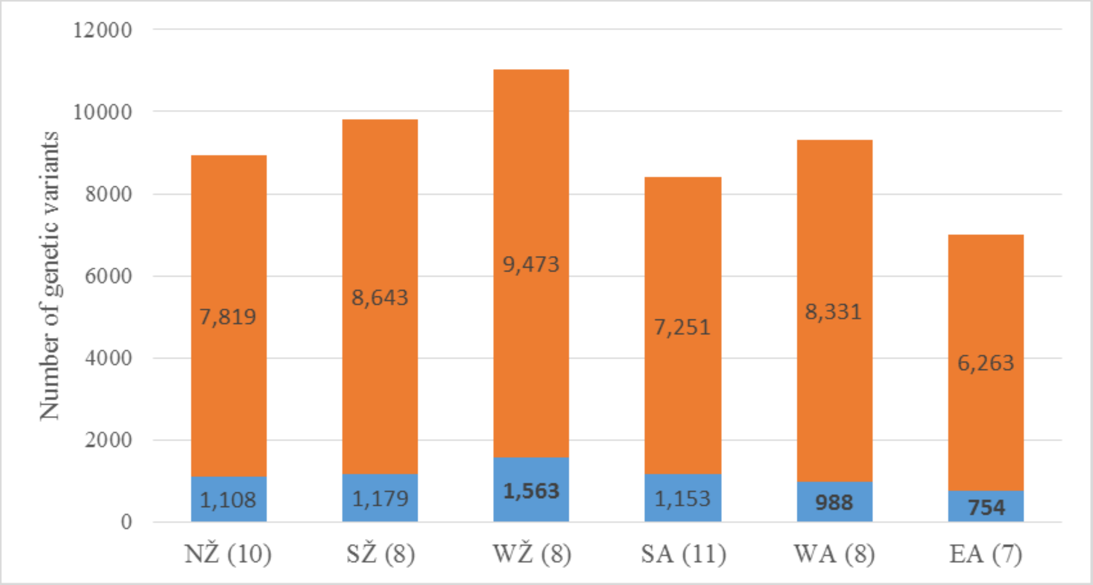


Fig. 3.46 Distribution of different exome variants (orange label) and singletons (blue label) per sample among the ethno-linguistic groups of Lithuania. Significant values are in bold (*P* < 0.05).

The number of exome variants shared between two or more ethno-linguistic groups of Lithuania were calculated. The majority of shared genetic variants were identified in South Aukštaitija (11 samples) and the least number in East Aukštaitija (7 samples). The number of shared genetic variants was higher in Žemaitija compared to Aukštaitija (Fig. 3.47).

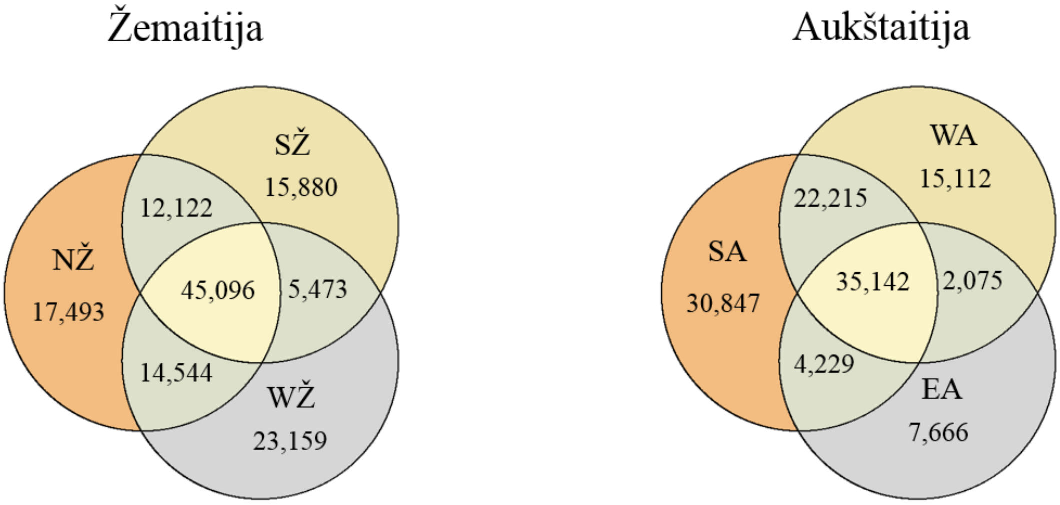


Fig. 3.47. Distribution of shared genetic variants among the ethno-linguistic groups of Lithuania.

The average number of heterozygous exome variants per sample was 18,574 (~64%) and of homozygous exome variants 10,325 (~36%). The average number of transition per sample was 21,039 (the ratio of heterozygous and homozygous genetic variants was 1.78) and transitions per sample 7,851 (the ratio of heterozygous and homozygous genetic variants was 1.84).

The distribution of exome variants according to MAF was evaluated. Exome variants with MAF <1% composed ~35% and exome variants with MAF <5% comprised ~60% of all identified variants (Fig. 3.48).

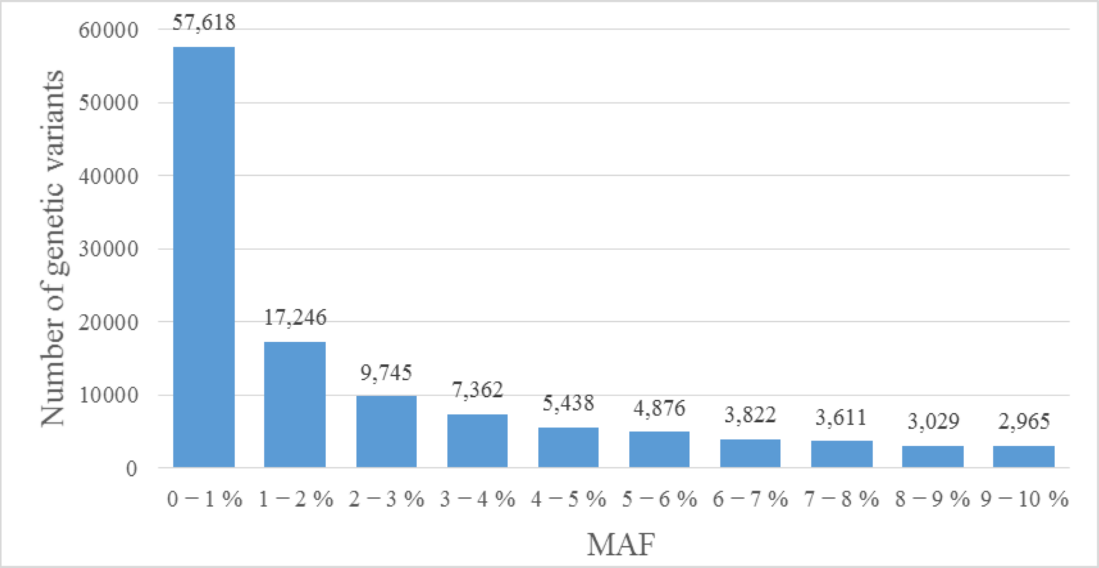


Fig. 3.48. Distribution of exome variants according to MAF.

107,444 exome variants (65.8%) could be assigned to one of groups (Fig. 3.49). The average number of exome variants per sample assigned to one of the group: (1) 3' untranslated region (3' UTR) 1,008 variants; (2) 5' UTR – 337 variants; (3) intron – 10,455 variants; (4) coding region – 6,864 synonymous variants; (5) coding region – 7,057 nonsynonymous variants. About 89% of all exome variants per samples were assigned to one of the five groups.

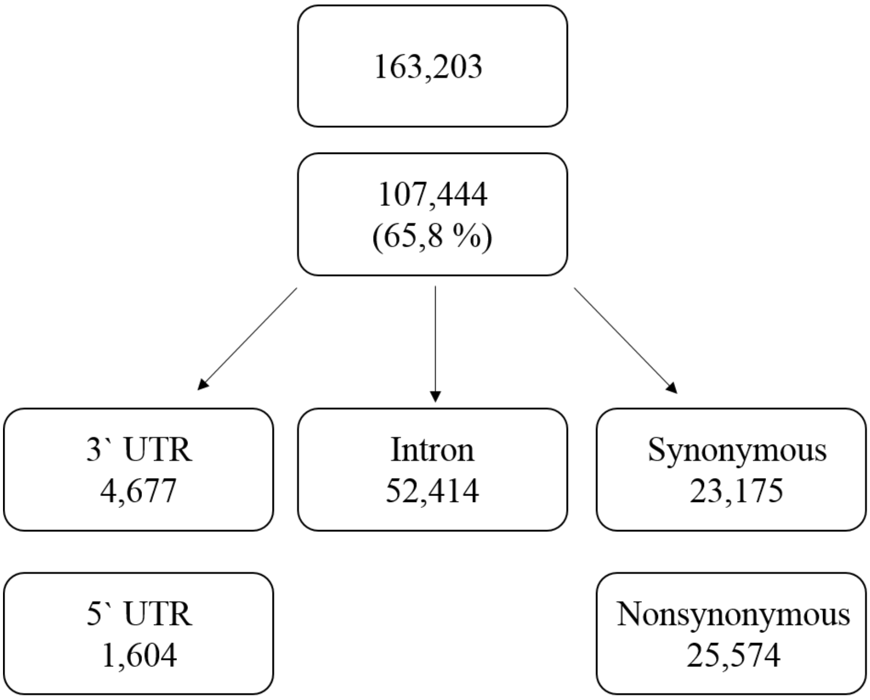


Fig. 3.49. Distribution of exome variants that could be assigned to one of the group.

A linear regression of rare exome variants (MAF < 5%) to total variants per gene in the Lithuanian population was done. Analysis showed that the outlying genes in the Lithuanian population were *MUC4*, *MUC5B*, *TTN*, which had the greatest number of rare variants per total number of variants as compared to the other genes, and *MUC16*, *ZNF717*, *HLA-DRB1*, *PKD1L2*, *MAP2K3*, *DYNC2H1*, which had the least number of rare variants per total number of variants. Moreover, the linear regression of rare exome variants (MAF < 5%) to total variants per gene for each ethno-linguistic group of Lithuania was done. Three ethno-linguistic groups of Lithuania (North, South Žemaitija, South Aukštaitija) had a similar number of rare variants. East and West Aukštaitija had a greater number of rare variants and West Žemaitija had more common variants (Fig. 3.50).

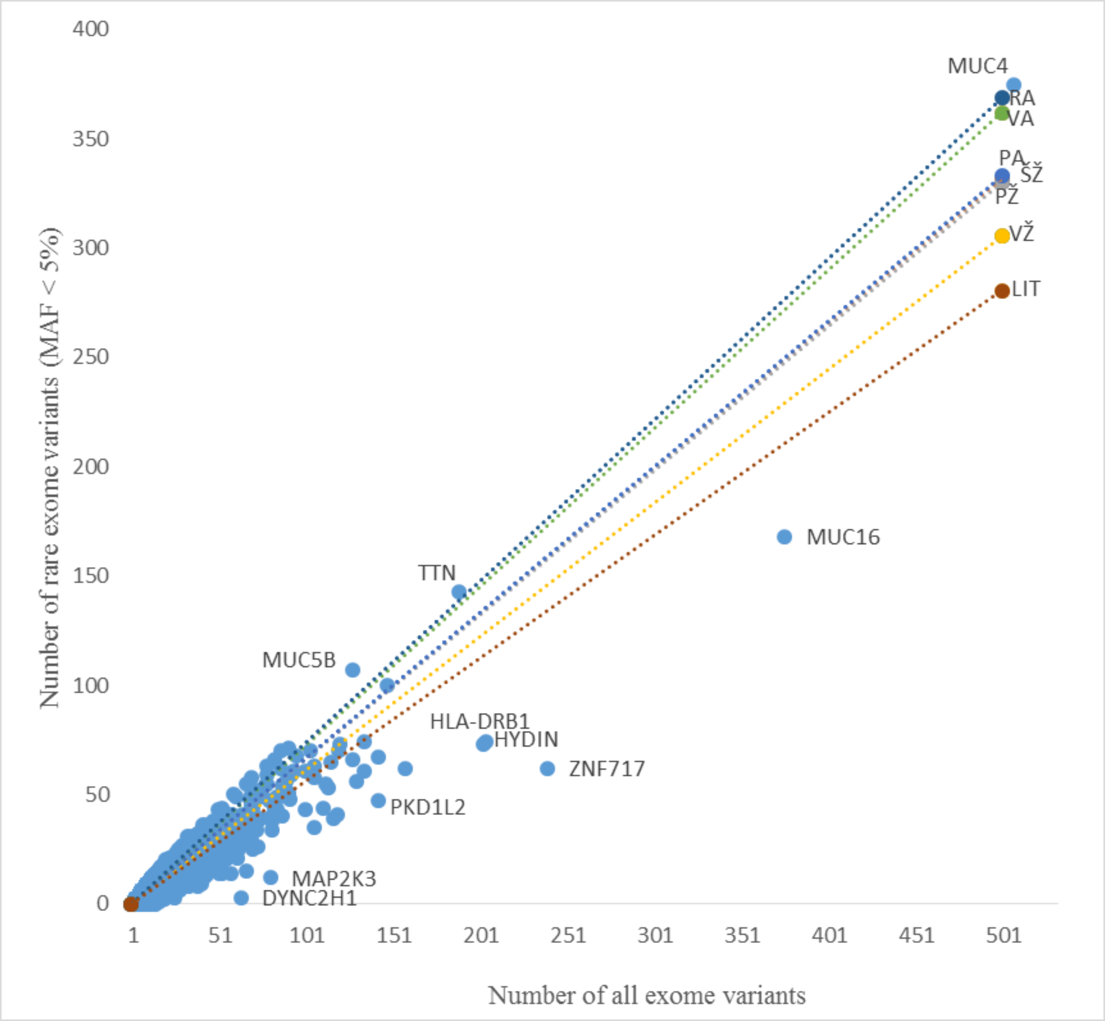


Fig. 3.50. Density of rare variants (MAF < 5%.) in each ethno-linguistic groups and the general Lithuanian population using all available SNPs. Linear regression of all six ethno-linguistic groups and the general Lithuanian population are shown in different colors. Abbreviations: LIT – Lithuanian population.

Exome variants (6,939) that were identified in more than 80% of samples were selected for the further MDS and PCA analyses. Pairwise differences based on selected exome variants were calculated (Table. 3.12).

Table. 3.12. Pairwise differences (below the diagonal) and *P* values (above the diagonal) based on 6,939 exome variants among six ethno-linguistic groups of Lithuania

	SA	**EA**	**WA**	**NŽ**	**WŽ**	**SŽ**
SA	0	**0.0009**	0.7916	0.7415	**0.0065**	0.6282
EA	0.0152	0	**0.0002**	**0.0002**	**0.0002**	**0.0003**
WA	-0.0208	0.0132	0	0.2349	0.0745	0.1782
NŽ	-0.0195	0.0258	-0.0209	0	**0.0289**	0.8319
WŽ	-0.0160	0.0302	-0.0210	-0.0209	0	**0.0005**
SŽ	-0.0225	0.0232	-0.0240	-0.0258	-0.0218	0

The MDS analysis based on pairwise differences of 6,939 exome variants was performed, and the scatter plot of MDS was generated. A tight cluster of three ethno-linguistic groups of Aukštaitija, i.e. South, West, East Aukštaitija, was observed in the MDS scatter plot. Three ethno-linguistic groups of Žemaitija were located farther from the cluster of Aukštaitija (Fig. 3.51).

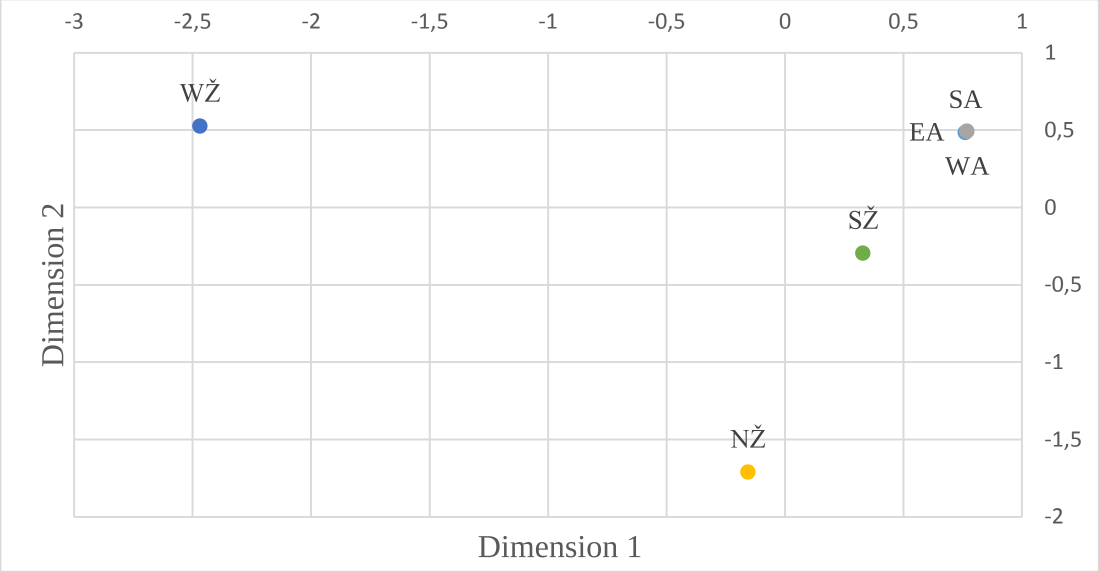


Fig. 3.51. Scatter plot of MDS analysis based on pairwise differences of 6,939 exome variants among six ethno-linguistic groups of Lithuania.

PCA based on pairwise differences of 6,939 exome variants were performed. The first two PCs explained 75.53% and 9.3% of the variation among six ethno-linguistic groups of Lithuania (Fig. 3.52).

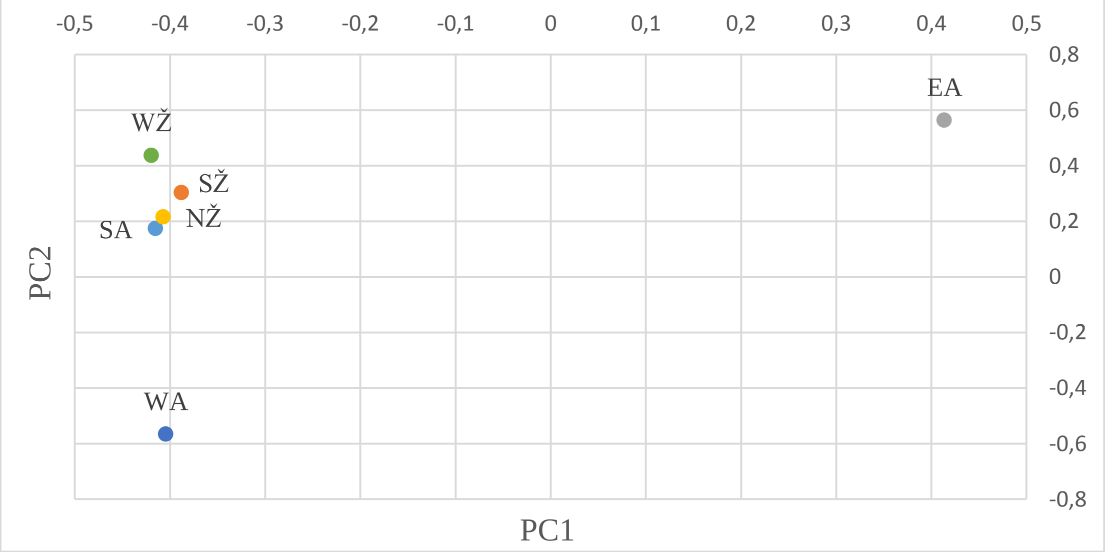


Fig. 3.52. Scatter plot of PCA based on pairwise differences of 6,939 exome variants among six ethno-linguistic groups of Lithuania.

Barriers of genetic variation of 6,939 exome variants among six ethno-linguistic groups of Lithuania were identified. The first barrier separated East Aukštaitija from the rest of Lithuanian population. The second barrier was found between North Žemaitija and the other ethno-linguistic groups of Lithuania. The third barrier was seen between South Aukštaitija and the rest of Lithuanian population (Fig. 3.53).

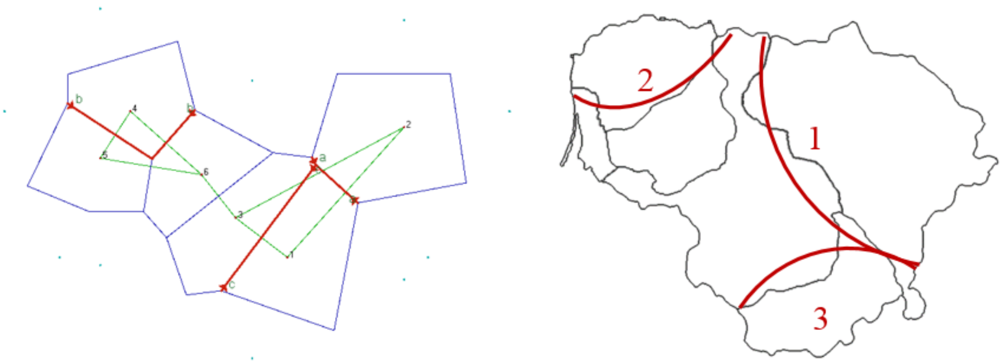


Fig. 3.53. Barriers of genetic variation i.e. 6,939 exome variants, among six ethno-linguistic groups of Lithuania. Left: output of the software, i.e. a schematic map as of the ethno-linguistic groups of Lithuania according to submitted coordinates (Annex 1, Table 1). Right: barriers were numbered in the order of priority. Analysis was based on pairwise differences.

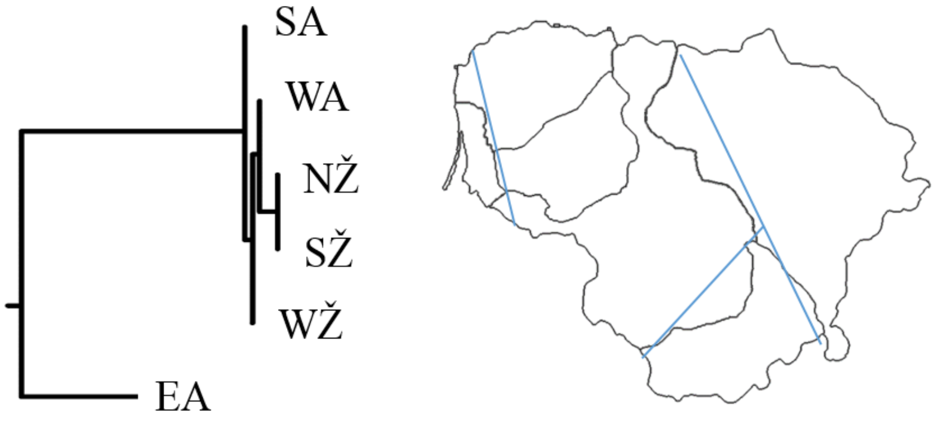


Fig. 3.54. The UPGMA tree based on pairwise differences. Right: lines on the map show boundary of the clusters.

The UPGMA clustering based on pairwise differences of genetic variation of 6,939 exome variants was performed. Two clusters of populations could be seen in the constructed UPGMA tree: (1) South, West Aukštaitija, West, South, North Žemaitija; (2) East Aukštaitija. The length of the South Aukštaitija branch reflected the higher level of genetic differentiation of this ethno-linguistic group (Fig. 3.54).

SUMMARY

The first studies analyzed the coding variants of *CFTR* and *PAH* genes in the group of patients from Lithuanian population (Giannattasio *et al.* 2006), (Tighe *et al.* 2003). The present study for the first time analyzed and described the genetic structure of six ethno-linguistic groups of Lithuania based on a large scale of coding variants, i.e. exome genetic variation.

The average number of variants per sample was ~28,890 (±6,528), and ~64% of which were heterozygous. The ratio of transition and transversions per sample was 2.7. This shows that transversions are less tolerated due to a possible impact on the phenotype. Most of the identified variants were in the coding region, i.e. synonymous (23.8%) and nonsynonymous (24.4%), and less of them were in the noncoding region, i.e. introns (36.2%), 5' UTR (1.2%), 3' UTR (3.5%). The majority of all identified variants per sample (~89%) were assigned to one of the five categories.

Approximately one third (~36%) of all variants (163,203) were singletons. Singletons dominated (~60%) within the group of rare variants (MAF < 5%) which comprised ~60% of all identified variants. The number of singletons was not equally distributed across Aukštaitija and Žemaitija (26 samples each) with the ratio of 5 : 3. Moreover, statistically significantly more singletons were seen in West Žemaitija and less in West and East Aukštaitija. A higher number of shared variants among the ethno-linguistic groups was identified in Žemaitija (33,008) and a lower number in Aukštaitija (25,860). All descriptive statistics could be slightly influenced by different protocols of exome capture, i.e. the usage of different capture kits.

The number of rare variants is not equal in different populations. The possible reason could be natural selection or various demographic events in a certain population. Natural selection affects the functional regions of a genome, while the demographic events affect the genome as a whole. Slope differences of linear regression of rare variants (MAF < 5%) to total variants reflects the demographic events of each population (Raska, Zhu 2011). The results have shown that East and West Aukštaitija have on the average more and West Žemaitija less rare variants as compared with the rest of the ethno-linguistic groups of Lithuania. This fact could be due to different demographic events of these ethno-linguistic groups.

The genes that are most prominent from linear regression could be affected by natural selection. Several genes with more rare variants (*MUC4*, *MUC5B*, *TTN*) were detected. The *MUC* gene family members (*MUC4*, *MUC5B*) encode secreted or cell-associated glycoproteins of epithelial cells. Proteins of these genes play a role in cellular interaction and protecting from extracellular pathogens. Titin (*TTN*) is the gene with the largest cDNA. A large number of rare variants in these genes could be due to the specificity of these genes or to the physical length of a gene.

Also, several genes with more common variants (*MUC16*, *ZNF717*, *HLA-DRB1*, *PKD1L2*, *MAP2K3*, *DYNC2H1*) were detected. *HLA-DRB1* (Major Histocompatibility Complex, Class II, DR Beta 1) gene product is part of the immune system. The *ZNF717* (Zinc Finger Protein 717) gene product is a transcription factor. The *PKD1L2* (Polycystic Kidney Disease 1-Like 2) gene product is a component of cation channel pores. The *MAP2K3* (Mitogen-activated Protein Kinase Kinase 3) gene product participates in the MAP kinase-mediated signaling cascade. The *DYNC2H1* (Dynein, Cytoplasmic 2, Heavy Chain 1) gene product is involved in retrograde transport in the cilium and has a role in intraflagellar transport. A larger number of common variants (MAF > 5%) in these genes may be due to the effect of natural selection on the gene products that interact with a large number of other gene products or participate in the general cell functions.

The statistics using all detected variants could be influenced by slightly different sequencing protocol, i.e. capture kits. This could be avoided by using variants that are detected in more than 80% of samples (variants with MAF < 5%

composed 12.8%). There was 6,939 autosomal variants that were used for subsequent analyses, such as the MDS, PCA, UPGMA trees and barriers of genetic variation.

Ethno-linguistic groups of Aukštaitija formed a tight cluster, and the rest of the ethno-linguistic groups were more scattered in the scatter plot of MDS. These results were different from the other analyses. The scatter plot of PCA showed a higher extent of genetic differences between East Aukštaitija and the rest of the ethno-linguistic groups. These results were supported by detected barriers of genetic variation. The first barrier was seen between East Aukštaitija and the rest of ethno-linguistic groups. Moreover, a higher extent of genetic differentiation of East Aukštaitija was seen as the longest branch in the UPGMA tree based on pairwise differences.

The results of the analyses based on exome autosomal variants were not constant. Most of the analyses showed differences in the genetic variation of East Aukštaitija, whereas differences of variation among the other ethno-linguistic groups could not be identified.

3.5. Discussion

The first differences among the ethno-linguistic groups of Lithuania were detected using phenetical distances (dermatoglyphics) (Klevcova, Kučinskas 1987). Also, results of analyses based on the distribution of blood groups AB0 and Rh(D) showed differences among the ethno-linguistic groups. The results of the mentioned study revealed that the ethno-linguistic groups of Žemaitija were more homogeneous than of Aukštaitija (Kučinskas *et al.* 1994). The reason could be that inhabitants of the territory of Žemaitija were autochthonous as compared with settlers of the territory of Aukštaitija, which could uptake the gene pool of the neighboring populations (Kučinskas 2004). Studies of twelve blood groups (AB0, Rh, MNS, P, LU, KELL, Co, LE, FUT2, FY, JK, LW), which were analyzed using molecular genetic methods, revealed statistically significant differences between South Aukštaitija and the rest of the ethno-linguistic groups of Lithuania according to the distribution of P1 and LW^b alleles. Analyses of serum proteins (TF, GC, PI) showed no statistically significant differences among the ethno-linguistic groups. Statistically significant differences between North Žemaitija and South Aukštaitija were detected when analyzing the distribution of Alu insertions, i.e. *Alu* TPA25 (Kučinskas 2001). The RFLP method was used to identify the genetic markers of mtDNA. Results of this study showed only minimal differences between Aukštaitija and Žemaitija (Kučinskas 1994). The further developments of molecular genetic methods enabled expansion of the genetic markers used to investigate the Lithuanian population (Kasperavičiūtė *et al.* 2004). Previous studies not only analyzed the genetic structure of the Lithuanian population, but also compared it with the other neighboring populations. The distribution of some genetic markers met the frequency gradient within Europe, but there were also exceptions (Kučinskas 2004). Genetic distances between Lithuanian and the neighboring populations, based on mtDNA and Y chromosome genetic markers, corresponded to the geographical location of the analyzed populations (Kasperavičiūtė *et al.* 2004).

A detailed analysis of genetic markers can provide more information about the genetic structure, distribution of genetic markers, haplotypes or haplogroups, the age of a certain phylogenetic lineage or a population (Gunnarsdottir *et al.* 2011; Underhill *et al.* 2010; Underhill *et al.* 2014; Wei *et al.* 2013).

Not only the unequal distribution of Y chromosome or mtDNA but also of autosomal genetic markers was detected (Menozzi *et al.* 1978). Improved or new molecular genetic methods and the development of technologies enabled to genotype thousands of markers in hundreds of samples at a time (IHC 2003; Lander *et al.* 2001). Different genetic markers and analysis methods allowed to study certain aspects of population evolution.

This study analyzed the genetic structure of six ethno-linguistic groups of Lithuania and compared it to that of the other populations based on the distribution of Y chromosome, mtDNA and, for the first time, the genome-wide autosomal and exome genetic markers.

Differences among the ethno-linguistic groups of Lithuania were most evident when the distribution of the genetic markers of Y chromosome were analyzed. The results were constant when analyzing the haplotypes (STRs) in which the genetic variation accumulated in a relatively short period of time or in haplogroups (SNPs) in which genetic variation accumulated in a relatively longer period of time. The frequency gradients of the main haplogroups (N1c1, R1a1a, R1a1a1g) across the North and South of the Lithuanian population were detected. The unequal distribution of these haplogroups across the Lithuanian population led to results that were revealed in this study. Geographic location within the territory of Lithuania but not linguistic or cultural differences could be the main reason for the observed distribution of genetic variation. Genetic variation differences could be due to gene pool migration along the North−South axis within the territory of Lithuania. There was no evidence of different periods of time when the main haplogroups settled on the territory of Lithuania, while different histories of the main haplogroups were supported by some of the observations of this study.

Smaller but constant genetic differences among six ethno-linguistic groups of Lithuania were seen when analyzing genome-wide autosomal genetic markers. Analysis of genetic markers with MAF > 5% can give insights about longer evolutionary processes of a certain population. Systematic QC is necessary to filter out autosomal genetic markers and samples of good quality for subsequent analyses. Some results of different analyses showed the unequal distribution of genetic variation across the Northwest and Southeast of Lithuanian population. The reason for these differences may be geographic location, but not linguistic or cultural differences of six ethno-linguistic groups of Lithuania. Axes of the frequency gradient of Y chromosome and genome-wide genetic markers are similar but not identical. A gradient based on genome-wide autosomal data may show the main axis of gene pool migration in the territory of the present-day Lithuania. Moreover, results of this study have supported the conclusion about the homogeneity of Žemaitija as compared with Aukštaitija of the previous study in which blood groups were analyzed (Kučinskas *et al.* 1994).

Some results of the mtDNA haplogroup and haplotype analysis showed differences among the ethno-linguistic groups of Lithuania. Unfortunately, these results were not constant across different analyses. No evident frequency gradients of the main phylogenetic lineages were seen. The distribution of the phylogenetic lineage U in this study did not correspond with the results of the previous study (Kasperavičiūtė *et al.* 2004). Results of the majority of analyses depended on the distribution of the main and other phylogenetic lineages of mtDNA. The AMOVA results suggested the West−East axis of the distribution of mtDNA haplogroups and haplotypes. The scatter plots of MDS or PCA did not support the AMOVA results, while the barriers of mtDNA genetic variation were similar to those of the AMOVA grouping based on the West−East location of the ethno-linguistic groups. The phylogenetic tree based on the genetic variation of mtDNA haplotypes (577−16,023 bp) with lower mutation rates showed the possible growth of the phylogenetic lineage H1. These observations disappeared when a phylogenetic tree based on the complete mtDNA including HVR with a higher mutation rate was constructed. Complete mtDNA was used to calculate the age of several phylogenetic lineages. The obtained results suggested that the phylogenetic lineage U could be dominant within the first settlers of the territory of present-day Lithuania and later was replaced by the phylogenetic lineage H which could reach the territory of Lithuania from Central or Eastern Europe.

The most contradictory results were obtained when analyzing the exome genetic variation. The descriptive statistics of exome variation was influenced by the use of different methods, i.e. capture kits, when sequencing different sets of samples. This could be improved if more samples using the same method were sequenced. To avoid the bias of different methods, the exome variants which were identified in more than 80% of samples were used for the further analysis. Somehow, East Aukštaitija was different from the rest of the ethno-linguistic groups of Lithuania across several analyses. Differences among the other ethno-linguistic groups could not be reliably determined. Analyzing rare variants (MAF < 5%), the information about the recent events of population evolution can be studied. Comprehensive results of the analyses of rare genetic variants will be obtained in the near future when larger samples set with the same sequencing methods will be proceeded.

4. CONCLUSIONS

1. Analyses of Y chromosome haplogroups and haplotypes revealed:
   1.1. The studied Lithuanian population is homogeneous as differences among its groups according to geographic locations comprised <2% of all genetic variations.
   1.2. The North−South gradients of the most frequent haplogroups (N1c1, R1a1a, R1a1a1g) were the main reason for the observed differences among the ethno-linguistic groups of Lithuania.
   1.3. The evolution of the Y chromosome haplogroups N1c1 and R1a1a could be different in the present-day territory of Lithuania.
   1.4. The Lithuanian population was close to the other two Baltic region populations (Latvia and Estonia), and the genetic distances among the populations depended on geographic distances.
2. Analyses of mtDNA haplogroups and haplotypes (complete mtDNA sequences) showed:
   2.1. The observed genetic differences were mainly (~99%) due to the intrapopulation genetic variation.
   2.2. Genetic differences between West and East of the Lithuanian population were due to the distribution of the two main (H and U) and the other (A, HV, I, J, K, M, N, T, V, W) phylogenetic lineages.
   2.3. The phylogenetic lineage U could be dominant in the first inhabitants of the present-day territory of Lithuania and more recently could be replaced by intruders who carried the phylogenetic lineage H.
   2.4. The studied Lithuanian population was located between Slavs and Middle / Near East populations.
3. Analyses of genome-wide autosomal SNPs demonstrated:
   3.1. The ethno-linguistic groups of Žemaitija were more homogeneous as compared with the ethno-linguistic groups of Aukštaitija.
   3.2. Genetic differences were noted between the Northwest and the Southeast Lithuanian populations.

3.3. The genetic pool of the studied Lithuanian population had a minor impact of Asian origin populations and was closest to the other HapMap3 populations of European origin (CEU and TSI).

4. East Aukštaitija exhibited the highest differentiation from other ethno-linguistic groups of Lithuania according to the results of the analyses of exome genetic variation.

5. This study has shown that the highest genetic differentiation of six ethno-linguistic groups of Lithuania is based on the distribution of Y chromosome haplogroups and haplotypes with a slightly lower differentiation based on the distribution of genome-wide autosomal SNPs and mtDNA haplogroups and haplotypes.

ACKNOWLEDGMENT

Special thanks to all Lithuanians who donated blood samples to be studied in the LITGEN project (VP1-3.1-ŠMM-07-K-01-013) which is funded by the European Social Fund under the Global Grant measure.

5. REFERENCES

1. Athey T. W. (2005). Haplogroup Prediction from Y-STR Values Using an Allele-Frequency Approach. *Journal of Genetic Genealogy*(1): 1-7.
2. Ballantyne K. N., Goedbloed M., Fang R., Schaap O., Lao O., Wollstein A., Choi Y., van Duijn K., Vermeulen M., Brauer S., Decorte R., Poetsch M., von Wurmb-Schwark N., de Knijff P., Labuda D., Vezina H., Knoblauch H., Lessig R., Roewer L., Ploski R., Dobosz T., Henke L., Henke J., Furtado M. R. and Kayser M. (2010). Mutability of Y-chromosomal Microsatellites: Rates, Characteristics, Molecular Bases, and Forensic Implications. *Am J Hum Genet* 87(3): 341-53.
3. Barbieri C., Whitten M., Beyer K., Schreiber H., Li M. and Pakendorf B. (2012). Contrasting Maternal and Paternal Histories in the Linguistic Context of Burkina Faso. *Mol Biol Evol* 29(4): 1213-23.
4. Behar D. M., Yunusbayev B., Metspalu M., Metspalu E., Rosset S., Parik J., Rootsi S., Chaubey G., Kutuev I., Yudkovsky G., Khusnutdinova E. K., Balanovsky O., Semino O., Pereira L., Comas D., Gurwitz D., Bonne-Tamir B., Parfitt T., Hammer M. F., Skorecki K. and Villems R. (2010). The Genome-Wide Structure of the Jewish People. *Nature* 466(7303): 238-42.
5. Belle E. M. S., Shah S., Parfitt T. and Thomas M. G. (2010). Y chromosomes of Self-Identified Syeds from the Indian Subcontinent Show Evidence of Elevated Arab Ancestry but Not of a Recent Common Patrilineal Origin. *Archaeol Anthropol Sci* (2): 217–24.
6. Bramanti B., Thomas M. G., Haak W., Unterlaender M., Jores P., Tambets K., Antanaitis-Jacobs I., Haidle M. N., Jankauskas R., Kind C. J., Lueth F., Terberger T., Hiller J., Matsumura S., Forster P. and Burger J. (2009). Genetic Discontinuity between Local Hunter-Gatherers and Central Europe's First Farmers. *Science* 326(5949): 137-40.
7. Casals F., Hodgkinson A., Hussin J., Idaghdour Y., Bruat V., de Maillard T., Grenier J. C., Gbeha E., Hamdan F. F., Girard S., Spinella J. F., Lariviere M., Saillour V., Healy J., Fernandez I., Sinnett D., Michaud J. L., Rouleau G. A., Haddad E., Le Deist F. and Awadalla P. (2013). Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. *PLoS Genet* 9(9): e1003815.
8. Costa M. D., Pereira J. B., Pala M., Fernandes V., Olivieri A., Achilli A., Perego U. A., Rychkov S., Naumova O., Hatina J., Woodward S. R., Eng K. K., Macaulay V., Carr M., Soares P., Pereira L. and Richards M. B. (2013). A Substantial Prehistoric European Ancestry Amongst Ashkenazi Maternal Lineages. *Nat Commun* 4: 2543.
9. Edgar R. C. (2004). Muscle: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Res* 32(5): 1792-7.
10. Excoffier L. and Lischer H. E. (2010). Arlequin Suite Ver 3.5: A New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol Ecol Resour* 10(3): 564-7.

11. Felsenstein J. (2004). Phylip (Phylogeny Inference Package) Version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.
12. Giannattasio S., Bobba A., Jurgelevičius V., Vacca R. A., Lattanzio P., Merafina R. S., Utkus A., Kučinskas V. and Marra E. (2006). Molecular Basis of Cystic Fibrosis in Lithuania: Incomplete CFTR Mutation Detection by PCR-Based Screening Protocols. *Genetic Testing* 10(3): 169-73.
13. Girdenis A. and Zinkevičius Z. (1966). Dėl Lietuvių Tarmių Klasifikacijos. *Kalbotyra*(14): 139-47.
14. Gunnarsdottir E. D., Li M., Bauchet M., Finstermeier K. and Stoneking M. (2011). High-Throughput Sequencing of Complete Human mtDNA Genomes from the Philippines. *Genome Res* 21(1): 1-11.
15. Hall T. A. (1999). Bioedit: A User-Friendly Biological Sequence Alignment Editor and Analysis Program for Windows 95/98/Nt. *Nucl. Acids. Symp. Ser*(41): 95-98.
16. Henn B. M., Gravel S., Moreno-Estrada A., Acevedo-Acevedo S. and Bustamante C. D. (2010). Fine-Scale Population Structure and the Era of Next-Generation Sequencing. *Hum Mol Genet* 19(R2): R221-6.
17. Huckins L. M., Boraska V., Franklin C. S., Floyd J. A. B., Southam L., GCAN, WTCCC, Sullivan P. F., Bulik C. M., Collier D. A., Tyler-Smith C., Zeggini E. and Tachmazidou I. (2014). Using Ancestry-Informative Markers to Identify Fine Structure across 15 Populations of European Origin. *Eur J Hum Genet*.
18. IHC (2003). The International Hapmap Project. *Nature* 426(6968): 789-96.
19. Jakkula E., Rehnstrom K., Varilo T., Pietilainen O. P., Paunio T., Pedersen N. L., deFaire U., Jarvelin M. R., Saharinen J., Freimer N., Ripatti S., Purcell S., Collins A., Daly M. J., Palotie A. and Peltonen L. (2008). The Genome-Wide Patterns of Variation Expose Significant Substructure in a Founder Population. *Am J Hum Genet* 83(6): 787-94.
20. Kasperavičiūtė D., Kučinskas V. and Stoneking M. (2004). Y chromosome and Mitochondrial DNA Variation in Lithuanians. *Ann Hum Genet* 68(Pt 5): 438-52.
21. Kivisild T., Shen P., Wall D. P., Do B., Sung R., Davis K., Passarino G., Underhill P. A., Scharfe C., Torroni A., Scozzari R., Modiano D., Coppa A., de Knijff P., Feldman M., Cavalli-Sforza L. L. and Oefner P. J. (2006). The Role of Selection in the Evolution of Human Mitochondrial Genomes. *Genetics* 172(1): 373-87.
22. Klevcova N. I. and Kučinskas V. (1987). Dermatoglifičeskije Ocobenosti Litovcev. *Voprosy Antropologii* (81): 74-88.
23. Kloss-Brandstatter A., Pacher D., Schonherr S., Weissensteiner H., Binna R., Specht G. and Kronenberg F. (2011). Haplogrep: A Fast and Reliable Algorithm for Automatic Classification of Mitochondrial DNA Haplogroups. *Hum Mutat* 32(1): 25-32.

24. Kučinskas V. (1994). Human Mitochondrial DNA Variation in Lithuania. *Anthropol Anz* 52(4): 289-95.
25. Kučinskas V. (2001). Population Genetics of Lithuanians. *Ann Hum Biol* 28(1): 1-14.
26. Kučinskas V. (2004). Genomo Įvairovė: Lietuviai Europoje. Vilnius, Splavų šalis.
27. Kučinskas V., Radikas J. and Rasmuson M. (1994). Genetic Diversity in the Lithuanian Rural Population as Illustrated by Variation in the AB0 and Rh(D) Blood Groups. *Hum Hered* 44(6): 344-9.
28. Kushniarevich A., Sivitskaya L., Danilenko N., Novogrodskii T., Tsybovsky I., Kiseleva A., Kotova S., Chaubey G., Metspalu E., Sahakyan H., Bahmanimehr A., Reidla M., Rootsi S., Parik J., Reisberg T., Achilli A., Hooshiar Kashani B., Gandini F., Olivieri A., Behar D. M., Torroni A., Davydenko O. and Villems R. (2013). Uniparental Genetic Heritage of Belarusians: Encounter of Rare Middle Eastern Matrilineages with a Central European Mitochondrial DNA Pool. *PLoS One* 8(6): e66499.
29. Laitinen V., Lahermo P., Sistonen P. and Savontaus M. L. (2002). Y-chromosomal Diversity Suggests That Baltic Males Share Common Finno-Ugric-Speaking Forefathers. *Hum Hered* 53(2): 68-78.
30. Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J. P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann N., Stojanovic N., Subramanian A., Wyman D., Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Coulson A., Deadman R., Deloukas P., Dunham A., Dunham I., Durbin R., French L., Grafham D., Gregory S., Hubbard T., Humphray S., Hunt A., Jones M., Lloyd C., McMurray A., Matthews L., Mercer S., Milne S., Mullikin J. C., Mungall A., Plumb R., Ross M., Shownkeen R., Sims S., Waterston R. H., Wilson R. K., Hillier L. W., McPherson J. D., Marra M. A., Mardis E. R., Fulton L. A., Chinwalla A. T., Pepin K. H., Gish W. R., Chissoe S. L., Wendl M. C., Delehaunty K. D., Miner T. L., Delehaunty A., Kramer J. B., Cook L. L., Fulton R. S., Johnson D. L., Minx P. J., Clifton S. W., Hawkins T., Branscomb E., Predki P., Richardson P., Wenning S., Slezak T., Doggett N., Cheng J. F., Olsen A., Lucas S., Elkin C., Uberbacher E., Frazier M., Gibbs R. A., Muzny D. M., Scherer S. E., Bouck J. B., Sodergren E. J., Worley K. C., Rives C. M., Gorrell J. H., Metzker M. L., Naylor S. L., Kucherlapati R. S., Nelson D. L., Weinstock G. M., Sakaki Y., Fujiyama A., Hattori M., Yada T., Toyoda A., Itoh T., Kawagoe C., Watanabe H., Totoki Y., Taylor T., Weissenbach J., Heilig R., Saurin W., Artiguenave F., Brottier P., Bruls T., Pelletier E., Robert C., Wincker P., Smith D. R., Doucette-Stamm L., Rubenfield M., Weinstock K., Lee H. M., Dubois J., Rosenthal A., Platzer M., Nyakatura G., Taudien S., Rump A., Yang H., Yu J., Wang J., Huang G., Gu J., Hood L., Rowen L., Madan A., Qin S., Davis R. W., Federspiel N. A., Abola A. P., Proctor

M. J., Myers R. M., Schmutz J., Dickson M., Grimwood J., Cox D. R., Olson M. V., Kaul R., Shimizu N., Kawasaki K., Minoshima S., Evans G. A., Athanasiou M., Schultz R., Roe B. A., Chen F., Pan H., Ramser J., Lehrach H., Reinhardt R., McCombie W. R., de la Bastide M., Dedhia N., Blocker H., Hornischer K., Nordsiek G., Agarwala R., Aravind L., Bailey J. A., Bateman A., Batzoglou S., Birney E., Bork P., Brown D. G., Burge C. B., Cerutti L., Chen H. C., Church D., Clamp M., Copley R. R., Doerks T., Eddy S. R., Eichler E. E., Furey T. S., Galagan J., Gilbert J. G., Harmon C., Hayashizaki Y., Haussler D., Hermjakob H., Hokamp K., Jang W., Johnson L. S., Jones T. A., Kasif S., Kaspryzk A., Kennedy S., Kent W. J., Kitts P., Koonin E. V., Korf I., Kulp D., Lancet D., Lowe T. M., McLysaght A., Mikkelsen T., Moran J. V., Mulder N., Pollara V. J., Ponting C. P., Schuler G., Schultz J., Slater G., Smit A. F., Stupka E., Szustakowski J., Thierry-Mieg D., Thierry-Mieg J., Wagner L., Wallis J., Wheeler R., Williams A., Wolf Y. I., Wolfe K. H., Yang S. P., Yeh R. F., Collins F., Guyer M. S., Peterson J., Felsenfeld A., Wetterstrand K. A., Patrinos A., Morgan M. J., de Jong P., Catanese J. J., Osoegawa K., Shizuya H., Choi S. and Chen Y. J. (2001). Initial Sequencing and Analysis of the Human Genome. *Nature* 409(6822): 860-921.

31. Lao O., Lu T. T., Nothnagel M., Junge O., Freitag-Wolf S., Caliebe A., Balascakova M., Bertranpetit J., Bindoff L. A., Comas D., Holmlund G., Kouvatsi A., Macek M., Mollet I., Parson W., Palo J., Ploski R., Sajantila A., Tagliabracci A., Gether U., Werge T., Rivadeneira F., Hofman A., Uitterlinden A. G., Gieger C., Wichmann H. E., Ruther A., Schreiber S., Becker C., Nurnberg P., Nelson M. R., Krawczak M. and Kayser M. (2008). Correlation between Genetic and Geographic Structure in Europe. *Curr Biol* 18(16): 1241-8.
32. Lappalainen T., Laitinen V., Salmela E., Andersen P., Huoponen K., Savontaus M. L. and Lahermo P. (2008). Migration Waves to the Baltic Sea Region. *Ann Hum Genet* 72(Pt 3): 337-48.
33. Librado P. and Rozas J. (2009). DnaSP V5: A Software for Comprehensive Analysis of DNA Polymorphism Data. *Bioinformatics* 25(11): 1451-2.
34. Lippold S., Xu H., Ko A., Li M., Renaud G., Butthof A., Schroder R. and Stoneking M. (2014). Human Paternal and Maternal Demographic Histories: Insights from High-Resolution Y chromosome and mtDNA Sequences. *Investig Genet* 5: 13.
35. Loogvali E. L., Roostalu U., Malyarchuk B. A., Derenko M. V., Kivisild T., Metspalu E., Tambets K., Reidla M., Tolk H. V., Parik J., Pennarun E., Laos S., Lunkina A., Golubenko M., Barac L., Pericic M., Balanovsky O. P., Gusar V., Khusnutdinova E. K., Stepanov V., Puzyrev V., Rudan P., Balanovska E. V., Grechanina E., Richard C., Moisan J. P., Chaventre A., Anagnou N. P., Pappa K. I., Michalodimitrakis E. N., Claustres M., Golge M., Mikerezi I., Usanga E. and Villems R. (2004). Disuniting Uniformity: A Pied Cladistic Canvas of mtDNA Haplogroup H in Eurasia. *Mol Biol Evol* 21(11): 2012-21.

36. Luca F., Di Giacomo F., Benincasa T., Popa L. O., Banyko J., Kracmarova A., Malaspina P., Novelletto A. and Brdicka R. (2007). Y-chromosomal Variation in the Czech Republic. *Am J Phys Anthropol* 132(1): 132-9.
37. Manni F., Guerard E. and Heyer E. (2004). Geographic Patterns of (Genetic, Morphologic, Linguistic) Variation: How Barriers Can Be Detected by Using Monmonier's Algorithm. *Hum Biol* 76(2): 173-90.
38. McEvoy B. P., Montgomery G. W., McRae A. F., Ripatti S., Perola M., Spector T. D., Cherkas L., Ahmadi K. R., Boomsma D., Willemsen G., Hottenga J. J., Pedersen N. L., Magnusson P. K., Kyvik K. O., Christensen K., Kaprio J., Heikkila K., Palotie A., Widen E., Muilu J., Syvanen A. C., Liljedahl U., Hardiman O., Cronin S., Peltonen L., Martin N. G. and Visscher P. M. (2009). Geographical Structure and Differential Natural Selection among North European Populations. *Genome Res* 19(5): 804-14.
39. Mendizabal I., Lao O., Marigorta U. M., Wollstein A., Gusmao L., Ferak V., Ioana M., Jordanova A., Kaneva R., Kouvatsi A., Kucinskas V., Makukh H., Metspalu A., Netea M. G., de Pablo R., Pamjav H., Radojkovic D., Rolleston S. J., Sertic J., Macek M., Jr., Comas D. and Kayser M. (2012). Reconstructing the Population History of European Romani from Genome-Wide Data. *Curr Biol* 22(24): 2342-9.
40. Menozzi P., Piazza A. and Cavalli-Sforza L. (1978). Synthetic Maps of Human Gene Frequencies in Europeans. *Science* 201(4358): 786-92.
41. Mishmar D., Ruiz-Pesini E., Golik P., Macaulay V., Clark A. G., Hosseini S., Brandon M., Easley K., Chen E., Brown M. D., Sukernik R. I., Olckers A. and Wallace D. C. (2003). Natural Selection Shaped Regional mtDNA Variation in Humans. *Proc Natl Acad Sci U S A* 100(1): 171-6.
42. Nasidze I., Ling E. Y., Quinque D., Dupanloup I., Cordaux R., Rychkov S., Naumova O., Zhukova O., Sarraf-Zadegan N., Naderi G. A., Asgary S., Sardas S., Farhud D. D., Sarkisian T., Asadov C., Kerimov A. and Stoneking M. (2004). Mitochondrial DNA and Y-chromosome Variation in the Caucasus. *Ann Hum Genet* 68(Pt 3): 205-21.
43. Nei M. (1972). Genetic Distance between Populations. *American Naturalist*(106): 283-92.
44. Nelis M., Esko T., Magi R., Zimprich F., Zimprich A., Toncheva D., Karachanak S., Piskackova T., Balascak I., Peltonen L., Jakkula E., Rehnstrom K., Lathrop M., Heath S., Galan P., Schreiber S., Meitinger T., Pfeufer A., Wichmann H. E., Melegh B., Polgar N., Toniolo D., Gasparini P., D'Adamo P., Klovins J., Nikitina-Zake L., Kucinskas V., Kasnauskiene J., Lubinski J., Debniak T., Limborska S., Khrunin A., Estivill X., Rabionet R., Marsal S., Julia A., Antonarakis S. E., Deutsch S., Borel C., Attar H., Gagnebin M., Macek M., Krawczak M., Remm M. and Metspalu A. (2009). Genetic Structure of Europeans: A View from the North-East. *PLoS One* 4(5): e5472.
45. Patterson N., Price A. L. and Reich D. (2006). Population Structure and Eigenanalysis. *PLoS Genet* 2(12): e190.

46. Pritchard J. K., Stephens M. and Donnelly P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155(2): 945-59.

47. Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A., Bender D., Maller J., Sklar P., de Bakker P. I., Daly M. J. and Sham P. C. (2007). Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81(3): 559-75.

48. Raska P. and Zhu X. (2011). Rare Variant Density across the Genome and across Populations. *BMC Proc* 5 Suppl 9: S39.

49. Richards M., Macaulay V., Hickey E., Vega E., Sykes B., Guida V., Rengo C., Sellitto D., Cruciani F., Kivisild T., Villems R., Thomas M., Rychkov S., Rychkov O., Rychkov Y., Golge M., Dimitrov D., Hill E., Bradley D., Romano V., Cali F., Vona G., Demaine A., Papiha S., Triantaphyllidis C., Stefanescu G., Hatina J., Belledi M., Di Rienzo A., Novelletto A., Oppenheim A., Norby S., Al-Zaheri N., Santachiara-Benerecetti S., Scozari R., Torroni A. and Bandelt H. J. (2000). Tracing European Founder Lineages in the near Eastern mtDNA Pool. *Am J Hum Genet* 67(5): 1251-76.

50. Saillard J., Forster P., Lynnerup N., Bandelt H. J. and Norby S. (2000). mtDNA Variation among Greenland Eskimos: The Edge of the Beringian Expansion. *Am J Hum Genet* 67(3): 718-26.

51. Schonberg A., Theunert C., Li M., Stoneking M. and Nasidze I. (2011). High-Throughput Sequencing of Complete Human mtDNA Genomes from the Caucasus and West Asia: High Diversity and Demographic Inferences. *Eur J Hum Genet* 19(9): 988-94.

52. Soares P., Ermini L., Thomson N., Mormina M., Rito T., Rohl A., Salas A., Oppenheimer S., Macaulay V. and Richards M. B. (2009). Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. *Am J Hum Genet* 84(6): 740-59.

53. Tian C., Plenge R. M., Ransom M., Lee A., Villoslada P., Selmi C., Klareskog L., Pulver A. E., Qi L., Gregersen P. K. and Seldin M. F. (2008). Analysis and Application of European Genetic Substructure Using 300 K SNP Information. *PLoS Genet* 4(1): e4.

54. Tighe O., Dunican D., O'Neill C., Bertorelle G., Beattie D., Graham C., Zschocke J., Cali F., Romano V., Hrabincova E., Kozak L., Nechyporenko M., Livshits L., Guldberg P., Jurkowska M., Zekanowski C., Perez B., Desviat L. R., Ugarte M., Kucinskas V., Knappskog P., Treacy E., Naughten E., Tyfield L., Byck S., Scriver C. R., Mayne P. D. and Croke D. T. (2003). Genetic Diversity within the R408W Phenylketonuria Mutation Lineages in Europe. *Hum Mutat* 21(4): 387-93.

55. Underhill P. A., Myres N. M., Rootsi S., Metspalu M., Zhivotovsky L. A., King R. J., Lin A. A., Chow C. E., Semino O., Battaglia V., Kutuev I., Jarve M., Chaubey G., Ayub Q., Mohyuddin A., Mehdi S. Q., Sengupta S., Rogaev E. I., Khusnutdinova E. K., Pshenichnov A., Balanovsky O., Balanovska E., Jeran N., Augustin D. H., Baldovic M., Herrera R. J., Thangaraj K., Singh V., Singh L., Majumder P., Rudan P., Primorac D.,

Villems R. and Kivisild T. (2010). Separating the Post-Glacial Coancestry of European and Asian Y chromosomes within Haplogroup R1a. *Eur J Hum Genet* 18(4): 479-84.

56. Underhill P. A., Poznik G. D., Rootsi S., Jarve M., Lin A. A., Wang J., Passarelli B., Kanbar J., Myres N. M., King R. J., Di Cristofaro J., Sahakyan H., Behar D. M., Kushniarevich A., Sarac J., Saric T., Rudan P., Pathak A. K., Chaubey G., Grugni V., Semino O., Yepiskoposyan L., Bahmanimehr A., Farjadian S., Balanovsky O., Khusnutdinova E. K., Herrera R. J., Chiaroni J., Bustamante C. D., Quake S. R., Kivisild T. and Villems R. (2014). The Phylogenetic and Geographic Structure of Y-chromosome Haplogroup R1a. *Eur J Hum Genet*.
57. Wei W., Ayub Q., Chen Y., McCarthy S., Hou Y., Carbone I., Xue Y. and Tyler-Smith C. (2013). A Calibrated Human Y-chromosomal Phylogeny Based on Resequencing. *Genome Res* 23(2): 388-95.
58. Wiik K. (2008). Where Did European Men Come From? *Journal of Genetic Genealogy*(4): 35-85.
59. Wilder J. A., Kingan S. B., Mobasher Z., Pilkington M. M. and Hammer M. F. (2004). Global Patterns of Human Mitochondrial DNA and Y-chromosome Structure Are Not Influenced by Higher Migration Rates of Females Versus Males. *Nat Genet* 36(10): 1122-5.
60. Wilson I., Weale M. and Balding D. (2003). Inferences from DNA Data: Population Histories, Evolutionary Processes and Forensic Match Probabilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*(166): 155-88.
61. Zhivotovsky L. A., Underhill P. A., Cinnioglu C., Kayser M., Morar B., Kivisild T., Scozzari R., Cruciani F., Destro-Bisol G., Spedini G., Chambers G. K., Herrera R. J., Yong K. K., Gresham D., Tournev I., Feldman M. W. and Kalaydjieva L. (2004). The Effective Mutation Rate at Y chromosome Short Tandem Repeats, with Application to Human Population-Divergence Time. *Am J Hum Genet* 74(1): 50-61.
62. Zoossmann-Diskin A. (2010). The Origin of Eastern European Jews Revealed by Autosomal, Sex Chromosomal and mtDNA Polymorphisms. *Biol Direct* 5: 57.

6. SANTRAUKA

Nevienodą genetinę įvairovę tarp populiacijų ir jų viduje lemia kalbiniai, kultūriniai ar geografiniai skirtumai. Pastarasis, manoma, gali būti viena iš pagrindinių esamo skirtingo populiacijų genetinio fondo priežasčių. Tiriant daugiau genetinių žymenų ir pasirinkus tinkamus duomenų kokybės parametrus galima nustatyti tarp artimų populiacijų ar populiacijos viduje esamus mažus genetinės įvairovės skirtumus.

Šis darbas, remiantis detaliau nustatyta mtDNR, Y chromosomos haplogrupių ir haplotipų įvairove ir pirmą kartą tolygiai po genomą pasiskirsčiusių ir egzomo DNR sekos variantų pasiskirstymu, įvertina šešių Lietuvos etnolingvistinių grupių genetinę struktūrą, taip pat lietuvių populiacijos padėtį kitų analizuotų populiacijų atžvilgiu. Pirmą kartą pateiktas darbo metu naudotų skirtingų genomo žymenų, pritaikytų populiacijos struktūrai nustatyti, palyginimas.

Tiriamųjų grupes sudaro atsitiktinai atrinkti negiminingi asmenys, kurie nurodė bent tris lietuvių tautybės kartas iš bendros lietuvių populiacijos. Genetiniams ir genominiams tyrimams vykdyti buvo gauti Vilniaus regioninio biomedicininių tyrimų etikos komiteto leidimai.

Darbo metu genominė DNR išskirta iš veninio kraujo leukocitų. Įvertinus išskirtos genominės DNR kiekybinius ir kokybinius rodiklius atlikti genominių žymenų (Y chromosomos, mtDNA, tolygiai po genomą pasiskirsčiusių ir egzomo variantų) nustatymas naudojant šiuolaikinius molekulinius genetinius tyrimo metodus ir naujausią laboratorinę įrangą Vilniaus universiteto Medicinos fakulteto Žmogaus ir medicininės genetikos katedroje ar užsienio įstaigose. Duomenų analizei panaudoti plačiai pasaulyje taikomi biostatistiniai metodai.

Analizuojant Y chromosomos haplogrupių ir haplotipų pasiskirstymą, galima manyti, kad genetinės įvairovės skirtumus tarp tirtų šešių etnolingvistinių grupių lėmė ne kalbinės ar kultūrinės priežastys, bet geografinė padėtis Lietuvos teritorijoje. Gauti rezultatai rodo, kad Lietuvos etnolingvistinių grupių Y chromosomos genetinės įvairovės pasiskirstymas pagal geografinę padėtį

galėjo susidaryti dėl genetinio fondo judėjimo šiaurės–pietų ašimi Lietuvos teritorijoje. Gauti tyrimo rezultatai neleidžia daryti išvadų apie skirtingą dažniausių haplogrupių (N1c1, R1a1a) formavimosi pradžią, bet leidžia manyti, kad formavimosi istorijos buvo skirtingos.

Dažniausių mtDNR filogenetinių linijų H (45,3 %) ir U (19,2 %) pasiskirstymas neturi aiškios gradiento ašies, ir skirtingų analizės metodų rezultatai priklauso nuo daugelio haplogrupių pasiskirstymo tarp tirtų šešių etnolingvistinių grupių. Dalies analizių rezultatai rodo mtDNR haplogrupių ir haplotipų pasiskirstymo vakarų–rytų ašį. Gauti mtDNR filogenetinių linijų analizės rezultatai neprieštarauja galimam filogenetinės linijos U dominavimui tarp pirmųjų gyventojų dabartinėje Lietuvos teritorijoje ir vėlesniam filogenetinės linijos H populiacijos atsikraustymui.

Gauti tolygiai po genomą pasiskirsčiusių genetinių žymenų analizės rezultatai, naudojant skirtingus metodus, rodo genetinės įvairovės skirtumus šiaurės vakarų–pietryčių ašimi. Galima skirtumų priežastis yra ne kalbinė ar kultūrinė, bet geografinė padėtis. Nustatyta etnolingvistinių grupių išsidėstymo ašis artima, bet netapati etnolingvistinių grupių padėčiai pagal Y chromosomos įvairovės skirtumus. Nustatyta ašis rodo galimą pirmųjų gyventojų genetinio fondo pagrindinę judėjimo dabartinėje Lietuvos teritorijoje ašį. Taip pat matyti, kad Žemaitija yra homogeniška, o asmenis iš Aukštaitijos, nors ir esant didelei paklaidos tikimybei, galima priskirti etnolingvistinei grupei.

Egzomo variantų analizės rezultatai, naudojant skirtingus metodus, nėra tolygūs ir matomi tik Rytų Aukštaitijos genetinės įvairovės skirtumai nuo kitų etnolingvistinių grupių, kurie pasikartoja atlikus analizę keliais metodais. Nustatytų skirtumų patvirtinimui reikalinga išplėsta tiriamųjų imtis ir tolygus duomenų pasiskirstymas.

Darbo metu gauti rezultatai papildo ankstesnių tyrimų rezultatus ir taps pagrindu tolesniems lietuvių populiacijos genetinės struktūros tyrimams. Lietuvių populiacijos genetinės struktūros nustatymas, analizuojant daugelio genomo žymenų alelių pasiskirstymą, bus aktualus kitiems planuojamiems ar vykdomiems lietuvių populiacijos tyrimams, kurių strategijos sudarymui ir rezultatams svarbus esamas netolygus genetinės įvairovės pasiskirstymas.

ANNEX 1. SUPPLEMENTARY MATERIAL

Supplementary material based on Y chromosome haplogroup and haplotype data

Table 1. Geographic coordinates of six ethno-linguistic groups of Lithuania (decimal degree format) used for analyses of genetic barriers with the *Barrier v2.2* software (Manni *et al.* 2004)

Ethno-linguistic group	Latitude	Longitude	Nearby City
E. Aukštaitija	55.95	25.46	Rokiškis
W. Aukštaitija	54.73	23.18	Kazlų Rūda
S. Aukštaitija	54.2	23.88	Seirijai
N. Žemaitija	56.16	21.76	Mosėdis
S. Žemaitija	55.31	22.72	Eržvilkas
W. Žemaitija	55.53	21.35	Priekulė

Table 2. Geographic distance (km) matrix of six ethno-linguistic groups of Lithuania, used for evaluating the correlation of genetic and geographical distances (Mantel test with *R v3.0*.3 (*Ade4*)

	RA	ŠŽ	PŽ	VŽ	VA	PA
RA	0	230.9	74.58	112.3	182.5	256.1
ŠŽ	230.9	0	261.4	186.1	197.9	184.6
PŽ	74.58	261.4	0	89.85	146.4	200.7
VŽ	112.3	186.1	89.85	0	70.85	143.8
VA	182.5	197.9	146.4	70.85	0	74.29
PA	256.1	184.6	200.7	143.8	74.29	0

Table 3. Mutation rates, number of investigated meiosis and weighting factor of Y chromosome 15 STRs for constructing the phylogenetic tree of three main haplogroups N1c1, R1a1a, R1a1a1g using *Network v4.6.1.2* software (www.fluxus-engineering.com) (Ballantyne *et al.* 2010)

Genetic marker (DYS)	393	390	19	391	439	389I	392	389II
Length of the motif (bp)	4	4	4	4	4	4	3	4
Number of mutation	3	2	7	5	6	9	1	6
Number of meiosis	1750	1758	1756	1759	1736	1751	1728	1743
Weighting factor	7	8	4	6	5	2	9	5
Genetic marker (DYS)	**458**	**437**	**448**	**H4**	**456**	**438**	**635**	**Avg.**
Length of the motif (bp)	4	4	6	4	4	5	4	
Number of mutation	14	2	0	5	8	1	6	5
Number of meiosis	1756	1760	1747	1755	1757	1751	1732	1749
Weighting factor	1	8	10	6	3	9	5	

Table 4. Prior parameters for TMRCA calculations of three main Y chromosome haplogroups using *BATWING* software (Wilson *et al.* 2003)

Mutation rate – 8.5×10^{-3} mutation/marker/generation (Ballantyne et al. 2010)					
Haplogroup	**N**	**N prior[1]**	**mu prior[2]**	**alfa prior[3]**	**beta prior[4]**
N1c1	119	gamma(1;0.001)	gamma(5;1,749)	uniform(0;0.04)	uniform(0;1)
R1a1a	97	gamma(1;0.001)	gamma(5;1,749)	uniform(0;0.04)	uniform(0;1)
R1a1a1g	25	gamma(1;0.0025)	gamma(5;1,749)	uniform(0;0.04)	uniform(0;1)
Mutation rate – 6.9×10^{-4} mutation/marker/generation (Zhivotovsky et al. 2004)					
Haplogroup	**N**	**N prior**	**mu prior**	**alfa prior**	**beta prior**
N1c1	119	gamma(1;0.001)	gamma(1.47;2,130)	uniform(0;0.04)	uniform(0;1)
R1a1a	97	gamma(1;0.001)	gamma(1.47;2,130)	uniform(0;0.04)	uniform(0;1)
R1a1a1g	25	gamma(1;0.0025)	gamma(1.47;2,130)	uniform(0;0.04)	uniform(0;1)
Mutation rate – 2×10^{-3} mutation/marker/generation (Luca et al. 2007)					
Haplogroup	**N**	**N prior**	**mu prior**	**alfa prior**	**beta prior**
N1c1	119	gamma(1;0.001)	gamma(2;1,000)	uniform(0;0.04)	uniform(0;1)
R1a1a	97	gamma(1;0.001)	gamma(2;1,000)	uniform(0;0.04)	uniform(0;1)
R1a1a1g	25	gamma(1;0.0025)	gamma(2;1,000)	uniform(0;0.04)	uniform(0;1)

[1] Ancestral population size (*N prior*).

[2] Mutation rate, mutation/marker/generation (*mu prior*).

[3] Population growth rate (*alfa prior*).

[4] The time at which population growth starts (*beta prior*) .

Table 5. Distribution of Y chromosome haplogroups among the European populations

	R1a	**R1b**	**N3**	**I1**	**J2**	**E1**	**G**	**Reference**
Latvia	0.389	0.097	0.416	0.071	0	0.009	0	(Lappalainen et al. 2008)
Estonia	0.373	0.042	0.339	0.169	0.017	0.025	0	(Lappalainen et al. 2008)
Poland	0.57	0.116	0.037	0.173	0.025	0.005	0	(Kayser et al. 2005)
Belorussia	0.456	0.044	0.088	0.176	0.044	0.044	0.015	(Kharkov et al. 2005)
E. Finland	0.059	0.026	0.709	0.197	0	0	0	(Lappalainen et al. 2008)
W. Finland	0.087	0.052	0.413	0.413	0	0.009	0	(Lappalainen et al. 2008)
Russia	0.483	0.068	0.14	0.159	0.014	0.048	0.012	(Malyarchuk. Derenko 2008)
Lithuania	0.4164	0.044	0.406	0.0478	0.01365	0.0273	0.0102	This study
Ukraine	0.397	0.208	0.057	0.038	0.076	0	0	(Varzari et al. 2013)
Sweden	0.244	0.131	0.144	0.375	0	0.013	0	(Lappalainen et al. 2008)
Germany	0.179	0.389	0.016	0.236	0.04	0.062	0	(Kayser et al. 2005)
Moldavia	0.304	0.16	0.016	0.048	0.04	0.128	0.008	(Varzari et al. 2013)
Romania	0.2037	0.1296	0	0.037	0.055	0.074	0.055	(Varzari et al. 2013)
Norway	0.263	0.05	0.038	0.325	0.011	0.011	0	(Dupuy et al. 2006)
Greece	0.113	0.133	0.0066	0.026	0.153	0.32	0.053	(King et al. 2011)
Turkey	0.055	0.1775	0.0475	0.025	0.2475	0.1075	0.115	(King et al. 2011)
Italy	0.12	0.233	0	0.0067	0.193	0	0.09	(Brisighelli et al. 2012)
Denmark	0.165	0.361	0.005	0.387	0.026	0	0	(Sanchez et al. 2004)
Greenland	0.088	0.199	0	0.167	0.019	0	0	(Sanchez et al. 2004)
Czech	0.342	0.28	0.016	0.051	0.035	0.008	0.051	(Luca et al. 2007)
Britain	0.054	0.59	0.009	0.172	0.009	0.009	0.036	(King. Jobling 2009)
Croatia	0.339	0.156	0	0.376	0.018	0.055	0.009	(Barac et al. 2003)

Table 6. The Nei (D) distance matrix based on the distribution of Y chromosome haplogroups (E1, G, I1, J2, N3, R1a, R1b) among the European populations

	Latvia	**Estonia**	**Poland**	**Belorussia**	**E. Finland**	**W. Finland**	**Russia**	**Lithuania**	**Ukraine**	**Sweden**	**Germany**
Latvia	0	0.003395	0.032796	0.022759	0.036619	0.037339	0.016834	0.000835	0.025829	0.033708	0.058159
Estonia	0.003395	0	0.023638	0.012501	0.040461	0.026169	0.009374	0.003673	0.022454	0.018834	0.048798
Poland	0.032796	0.023638	0	0.003885	0.128073	0.077964	0.003832	0.031576	0.010195	0.027803	0.041433
Belorussia	0.022759	0.012501	0.003885	0	0.096403	0.053715	0.000900	0.021101	0.009209	0.017312	0.036530
E. Finland	0.036619	0.040461	0.128073	0.096403	0	0.022714	0.089503	0.041407	0.106003	0.068989	0.113248
W. Finland	0.037339	0.026169	0.077964	0.053715	0.022714	0	0.053315	0.043067	0.070074	0.018480	0.056871
Russia	0.016834	0.009374	0.003832	0.000900	0.089503	0.053315	0	0.015547	0.009420	0.019378	0.039043
Lithuania	0.000835	0.003673	0.031576	0.021101	0.041407	0.043067	0.015547	0	0.026782	0.038076	0.066102
Ukraine	0.025829	0.022454	0.010195	0.009209	0.106003	0.070074	0.009420	0.026782	0	0.027337	0.021989
Sweden	0.033708	0.018834	0.027803	0.017312	0.068989	0.018480	0.019378	0.038076	0.027337	0	0.019735
Germany	0.058159	0.048798	0.041433	0.036530	0.113248	0.056871	0.039043	0.066102	0.021989	0.019735	0
Moldavia	0.032829	0.026009	0.017878	0.011239	0.105077	0.064886	0.012996	0.033013	0.005206	0.024890	0.018718
Romania	0.037783	0.029969	0.027635	0.016955	0.098834	0.059468	0.020361	0.038359	0.009309	0.024844	0.018664
Norway	0.039186	0.021810	0.020776	0.010904	0.088038	0.030958	0.015003	0.040996	0.022188	0.003250	0.022956
Greece	0.066976	0.057111	0.062146	0.043454	0.118917	0.080816	0.048675	0.066124	0.035072	0.050173	0.035154
Turkey	0.061465	0.054055	0.064809	0.046109	0.102093	0.069609	0.052188	0.062974	0.030138	0.045583	0.029329
Italy	0.055093	0.049476	0.048755	0.037253	0.110480	0.073455	0.042236	0.057829	0.017637	0.039736	0.019895
Denmark	0.070494	0.055085	0.047666	0.042461	0.116927	0.048067	0.046433	0.079822	0.035777	0.014032	0.004899
Greenland	0.047928	0.036965	0.040273	0.028153	0.090817	0.042543	0.032480	0.052836	0.019471	0.014834	0.007985
Czech	0.035168	0.031549	0.016658	0.015990	0.114951	0.073236	0.016596	0.037916	0.002460	0.027568	0.013575
Britain	0.094955	0.091646	0.086534	0.082770	0.143541	0.091088	0.084353	0.108044	0.049948	0.053730	0.010781
Croatia	0.049468	0.031071	0.017304	0.013127	0.113147	0.044525	0.016859	0.052443	0.022644	0.005766	0.017909

Table 7. The Nei (D) distance matrix based on the distribution of Y chromosome haplogroups (E1, G, I1, J2, N3, R1a, R1b) among the European populations

	Moldavia	**Romania**	**Norway**	**Greece**	**Turkey**	**Italy**	**Denmark**	**Greenland**	**Czech**	**Britain**	**Croatia**
Latvia	0.032829	0.037783	0.039186	0.066976	0.061465	0.055093	0.070494	0.047928	0.035168	0.094955	0.049468
Estonia	0.026009	0.029969	0.021810	0.057111	0.054055	0.049476	0.055085	0.036965	0.031549	0.091646	0.031071
Poland	0.017878	0.027635	0.020776	0.062146	0.064809	0.048755	0.047666	0.040273	0.016658	0.086534	0.017304
Belorussia	0.011239	0.016955	0.010904	0.043454	0.046109	0.037253	0.042461	0.028153	0.015990	0.082770	0.013127
E. Finland	0.105077	0.098834	0.088038	0.118917	0.102093	0.110480	0.116927	0.090817	0.114951	0.143541	0.113147
W. Finland	0.064886	0.059468	0.030958	0.080816	0.069609	0.073455	0.048067	0.042543	0.073236	0.091088	0.044525
Russia	0.012996	0.020361	0.015003	0.048675	0.052188	0.042236	0.046433	0.032480	0.016596	0.084353	0.016859
Lithuania	0.033013	0.038359	0.040996	0.066124	0.062974	0.057829	0.079822	0.052836	0.037916	0.108044	0.052443
Ukraine	0.005206	0.009309	0.022188	0.035072	0.030138	0.017637	0.035777	0.019471	0.002460	0.049948	0.022644
Sweden	0.024890	0.024844	0.003250	0.050173	0.045583	0.039736	0.014032	0.014834	0.027568	0.053730	0.005766
Germany	0.018718	0.018664	0.022956	0.035154	0.029329	0.019895	0.004899	0.007985	0.013575	0.010781	0.017909
Moldavia	0	0.002716	0.017980	0.015474	0.020637	0.014935	0.033492	0.012866	0.005500	0.048013	0.019967
Romania	0.002716	0	0.017333	0.013084	0.011350	0.007407	0.031961	0.007266	0.007809	0.044040	0.023336
Norway	0.017980	0.017333	0	0.041444	0.039653	0.033836	0.019078	0.012941	0.023593	0.062186	0.003654
Greece	0.015474	0.013084	0.041444	0	0.011452	0.019882	0.055052	0.024304	0.032687	0.061538	0.047630
Turkey	0.020637	0.011350	0.039653	0.011452	0	0.004090	0.045198	0.016199	0.026935	0.046187	0.049300
Italy	0.014935	0.007407	0.033836	0.019882	0.004090	0	0.034537	0.010734	0.013680	0.033353	0.040240
Denmark	0.033492	0.031961	0.019078	0.055052	0.045198	0.034537	0	0.012808	0.026883	0.019109	0.013350
Greenland	0.012866	0.007266	0.012941	0.024304	0.016199	0.010734	0.012808	0	0.014208	0.025685	0.017945
Czech	0.005500	0.007809	0.023593	0.032687	0.026935	0.013680	0.026883	0.014208	0	0.033317	0.021955
Britain	0.048013	0.044040	0.062186	0.061538	0.046187	0.033353	0.019109	0.025685	0.033317	0	0.054925
Croatia	0.019967	0.023336	0.003654	0.04763	0.0493	0.04024	0.01335	0.017945	0.021955	0.054925	0

Supplementary material based on mtDNA haplogroup and haplotype data

Table 8. The Nei (D) distance matrix based on the distribution of 23 mtDNA haplogroups (above the diagonal) and 69 mtDNA haplogroups (below the diagonal) among six ethno-linguistic groups of Lithuania

	EA	**SA**	**WA**	**NŽ**	**SŽ**	**WŽ**
EA	0	0.002739	0.001627	0.001231	0.002881	0.000936
SA	0.000784	0	0.001554	0.0027	0.00207	0.001566
WA	0.000577	0.000789	0	0.002281	0.001921	0.001386
NŽ	0.000353	0.000731	0.000635	0	0.003716	0.001123
SŽ	0.000885	0.000823	0.000779	0.000846	0	0.002387
WŽ	0.000592	0.000895	0.000695	0.000464	0.000881	0

Table 9. Pairwise distance matrix on the distribution of mtDNA haplotypes 577−16,568 bp (above the diagonal) and mtDNA haplotypes 16,568 bp (below the diagonal) among six ethno-linguistic groups of Lithuania. Values in italics *P* = 0.07, values in bold and italics *P* = 0.06

	EA	**SA**	**WA**	**NŽ**	**SŽ**	**WŽ**
EA	0	0.00242	-0.00546	0.00101	0.0035	-0.00012
SA	0.00094	0	-0.00661	0.00139	0.01087	-0.00849
WA	-0.00438	-0.00419	0	0.00198	-0.00017	-0.00237
NŽ	0.00128	0.00285	0.00439	0	***0.01244***	-0.00433
SŽ	0.00236	0.00942	-0.00057	*0.00972*	0	0.009
WŽ	-0.00147	-0.00732	-0.00001	-0.00325	0.00788	0

Table 10. mtDNA haplotypes (16,569 bp) from European population used in this study

Population	N	Number of db *GenBank*[1]	Reference
Saami	17	AY882379, DQ902708	(Achilli et al. 2005; Ingman, Gyllensten 2007)
Belorussia	33	KC867103	(Kushniarevich et al. 2013)
Kazan Tatars	73	GU122975	(Malyarchuk et al. 2010)
Poland	26	JX307099, JX128041, JX266260	(Mielnik-Sikorska et al. 2013)
Ukraine	7	JX307099, JX128041, JX266260	(Mielnik-Sikorska et al. 2013)
Russia	21	JX307099, JX128041, JX266260	(Mielnik-Sikorska et al. 2013)
Slovakia	11	JX307099, JX128041, JX266260	(Mielnik-Sikorska et al. 2013)
Czech	8	JX307099, JX128041, JX266260	(Mielnik-Sikorska et al. 2013)
Georgia	28	HM852756	(Schonberg et al. 2011)
Turkey	29	HM852756	(Schonberg et al. 2011)
Armenia	30	HM852756	(Schonberg et al. 2011)
Azerbaijan	30	HM852756	(Schonberg et al. 2011)
Iran	30	HM852756	(Schonberg et al. 2011)
Ashkenazi[2]	72	KC878709, JX273243	(Schonberg et al. 2011)
Spain	11	AY882379	(Costa et al. 2013)
Italy	20	AY882379	(Achilli et al. 2005)
Sardinia	9	GQ129143	(Achilli et al. 2005)

[1] Datasets available at *NCBI DNA&RNA PopSet*.

[2] Near East.

Table 11. Pairwise distance (below the diagonal) and Φ_{ST} (above the diagonal) matrix based on the distribution of mtDNA haplotype (577−16,023 bp) among European populations. Abbreviations: AKZ − Ashkenazi, ARM − Armenia, AZR − Azerbaijan, BUY − Belorussia, CZH − Czech, ESP − Spain, GEO − Georgia, IRN − Iran, ITA − Italy, LIT − Lithuania, PLN − Poland, RUS − Russia, SAA − Saami, SLO − Slovakia, SRD − Sardinia, TAT − Tatars, TUR − Turkey, UKR − Ukraine. Statistically significant values in bold (*P* < 0.05; 10,100 permutations)

	AKZ	**ARM**	**AZR**	**BUY**	**CZH**	**ESP**	**GEO**	**IRN**	**ITA**
AKZ	0	0.0002	0.0002	**0.01225**	0.00022	0.00021	0.0002	0.00133	0.00273
ARM	**0.38361**	0	0	**0.01236**	0	0	0	0.00115	0.00259
AZR	**0.36423**	0.00312	0	**0.01236**	0	0	0	0.00115	0.00259
BUY	**0.59565**	**0.26931**	**0.24261**	0	0.0136	0.01315	**0.01239**	**0.0135**	**0.01517**
CZH	**0.59314**	**0.13284**	**0.1354**	**0.44405**	0	0	0	0.00126	0.00284
ESP	**0.46992**	**0.15571**	**0.13799**	**0.4405**	**0.43708**	0	0	0.00122	0.00274
GEO	**0.36082**	-0.0069	0.00328	**0.24868**	**0.13948**	**0.13045**	0	0.00115	0.00259
IRN	**0.44062**	0.01242	0.0119	**0.23806**	**0.10874**	**0.19527**	**0.02597**	0	0.00375
ITA	**0.44296**	**0.18812**	**0.1741**	**0.45519**	**0.44039**	-0.02004	**0.16682**	**0.23161**	0
LIT	**0.39475**	**0.02431**	**0.06029**	**0.27173**	**0.05121**	**0.19551**	**0.04336**	**0.02932**	**0.22035**
PLN	**0.52225**	**0.1154**	**0.11778**	**0.33908**	0.03756	**0.30631**	**0.1276**	**0.07772**	**0.33917**
RUS	**0.61405**	**0.19384**	**0.20112**	**0.48309**	**0.0651**	**0.50884**	**0.20583**	**0.16924**	**0.49841**
SAA	**0.50439**	**0.13872**	**0.11078**	**0.37377**	**0.30169**	**0.16666**	**0.13456**	**0.11288**	**0.18456**
SLO	**0.60467**	**0.16367**	**0.16593**	**0.46706**	0.03741	**0.47725**	**0.17151**	**0.13913**	**0.47009**
SRD	**0.59411**	**0.3133**	**0.26542**	**0.57491**	**0.75291**	**0.13441**	**0.28425**	**0.33206**	**0.11124**
TAT	**0.3518**	0.00223	0.01102	**0.21615**	**0.092**	**0.13743**	0.00734	0.00244	**0.17063**
TUR	**0.40001**	0.00317	0.00986	**0.25891**	**0.10572**	**0.1494**	0.01179	0.00572	**0.18474**
UKR	**0.54835**	**0.10711**	**0.06681**	**0.34821**	**0.14995**	**0.31754**	**0.10803**	0.037	**0.35776**

Table 12. Pairwise distance (below the diagonal) and Φ_{ST} (above the diagonal) matrix based on the distribution of mtDNA haplotype (577−16,023 bp) among European populations. Abbreviations: AKZ − Ashkenazi, ARM − Armenia, AZR − Azerbaijan, BUY − Belorussia, CZH − Czech, ESP − Spain, GEO − Georgia, IRN − Iran, ITA − Italy, LIT − Lithuania, PLN − Poland, RUS − Russia, SAA − Saami, SLO − Slovakia, SRD − Sardinia, TAT − Tatars, TUR − Turkey, UKR − Ukraine. Statistically significant values in bold (*P* < 0.05; 10,100 permutations)

	LIT	**PLN**	**RUS**	**SAA**	**SLO**	**SRD**	**TAT**	**TUR**	**UKR**
AKZ	**0.00157**	**0.00768**	**0.00479**	**0.05448**	0.00021	**0.04843**	**0.00134**	**0.00381**	0.00022
ARM	0.0014	**0.00765**	**0.00469**	**0.05642**	0	**0.04977**	0.00116	0.00369	0
AZR	0.0014	**0.00765**	**0.00469**	**0.05642**	0	**0.04977**	0.00116	0.00369	0
BUY	**0.01326**	**0.02005**	**0.01723**	**0.06881**	0.01315	**0.06288**	**0.01321**	**0.01605**	0.01383
CZH	0.00152	0.00841	0.00515	**0.0641**	0	0.05654	0.00127	0.00405	0
ESP	0.00148	0.00813	0.00498	**0.0615**	0	**0.0541**	0.00123	0.00391	0
GEO	0.0014	**0.00767**	**0.0047**	**0.05666**	0	**0.04996**	0.00117	0.0037	0
IRN	**0.00251**	**0.0088**	**0.00586**	**0.0576**	0.00122	**0.05101**	**0.00229**	0.00484	0.00128
ITA	**0.00393**	**0.01039**	0.0074	**0.06068**	0.00274	**0.05385**	**0.00372**	0.00634	0.00289
LIT	0	**0.00878**	**0.00595**	**0.05439**	0.00148	**0.04885**	**0.0025**	**0.00497**	0.00155
PLN	**0.07788**	0	**0.01249**	**0.06479**	0.00813	**0.05843**	**0.00864**	**0.01137**	0.00855
RUS	**0.09034**	**0.04201**	0	**0.06263**	0.00498	**0.05595**	**0.00577**	**0.00844**	0.00524
SAA	**0.15383**	**0.20189**	**0.37653**	0	**0.0615**	**0.11466**	**0.05544**	**0.06032**	**0.0654**
SLO	**0.07535**	0.01402	-0.01279	**0.3382**	0	**0.0541**	0.00123	0.00391	0
SRD	**0.3129**	**0.44151**	**0.72082**	**0.33237**	**0.76362**	0	**0.04944**	**0.05383**	0.05779
TAT	**0.02239**	**0.0812**	**0.13773**	**0.09933**	**0.11586**	**0.25855**	0	**0.00478**	0.00129
TUR	**0.0183**	**0.08655**	**0.16777**	**0.11016**	**0.13736**	**0.30432**	0.00284	0	0.00411
UKR	**0.09891**	0.0336	**0.24015**	**0.14466**	**0.20313**	**0.54435**	**0.04779**	**0.07329**	0

Supplementary material based on genome-wide SNPs

Table 13. QC parameters (*GenomeStudio v2011.1*)

QC values	
SAMPLES	
Parameter	**Notes**
Call rate	Range [0.97−100] (all samples); 0.97 accepted if all other QC values are of good quality
p95Grn. p95Red	Red [3,874−9,931]; Green [3,002−7,751]
p10GC	>0.7 (all samples); =0.4 (LTG-1075) eliminated
P-P-C Error Rate	Range [0−0.07]
Sample graph	All samples except for the LTG-1075
GENOTYPES	
Call frequency. Call Freq	Range [0.13−1]; Based on 1% rule. All that are <0.9845 eliminated. Number of eliminated SNPs - 10,262

Heritability error	Rep = 0 (accepted);
	P-C kl.=0 (accepted);
	P-P-C cl. assess

GenTrain score	Range [0.35−0.98] (all samples);

Het excess	<(-0.3) − 31 SNPs (accepted);
	>0.2 − 25 SNPs (accepted);
	(X. XY. Y. 0 chromosomes assessed separately)

ClusterSep	<0.27 eliminate or asses and leave as accepted; 0.27<x<0.4; 2,128 cases were assessed and eliminated or accepted

AB T Mean	<0.3 − 2 209 SNPs (accepted);
	<0.2 – 30 SNPs (accepted);
	>0.7 – 6,722 SNPs (accepted);
	>0.8 none

AB R Mean	Range [0.16−4.85]; <0.145 eliminated
Minor Freq	MAF<0.1 autosome – 140,839 SNPs; MAF<0.001 none
	Y chr. – 1,372 SNPs; all SNPs were assessed and corrected if necessary; X chr. – 17,814 SNPs out of which 14,036 SNPs with AB ≠ 0; SNPs with heterozygous values should be eliminated; XY (PAR) – 463 SNPs (accepted).

Table 14. QC parameters (*GenomeStudio v2011.1*)

PROCESS OF GENOTYPING	
Parameter	**Notes**
Staining Control	Intensity:
	>3 000 (all samples); LTG-1075 eliminated
	>2 000 (all samples); LTG-1075 eliminated
Extension Control	Intensity:
	>3 000 (all samples); LTG-1075 eliminated
	>2 000 (all samples); LTG-1075 eliminated
Target Removal Control	Intensity:
	Only *green* system was assessed
	<200 (all samples); LTG-1075 eliminated
Hybridization Control	Intensity:
	Only *green* system was assessed

	J ~2 000;
	M ~3−7 000;
	Ž ~ 5−12 000 (all samples);
	LTG-1075 eliminated

Restoration Control	Intensity:
	Only *green* system was assessed
	<500 (all samples); LTG-1075 eliminated
Stringency Control	Intensity:

	PM > 6 000;
	MM < 1 000 (all samples);
	LTG-1075 eliminated

	Only *red* system was assessed
Non-specific Binding Control	Intensity:
	<500 (all samples); LTG-1075 eliminated
	<400 (all samples); LTG-1075 eliminated
Non-polymorphic Control	Intensity:
	>3 000 (all samples); LTG-1388?; LTG-1075 eliminated
	>3 000 (all samples); LTG-1075 eliminated

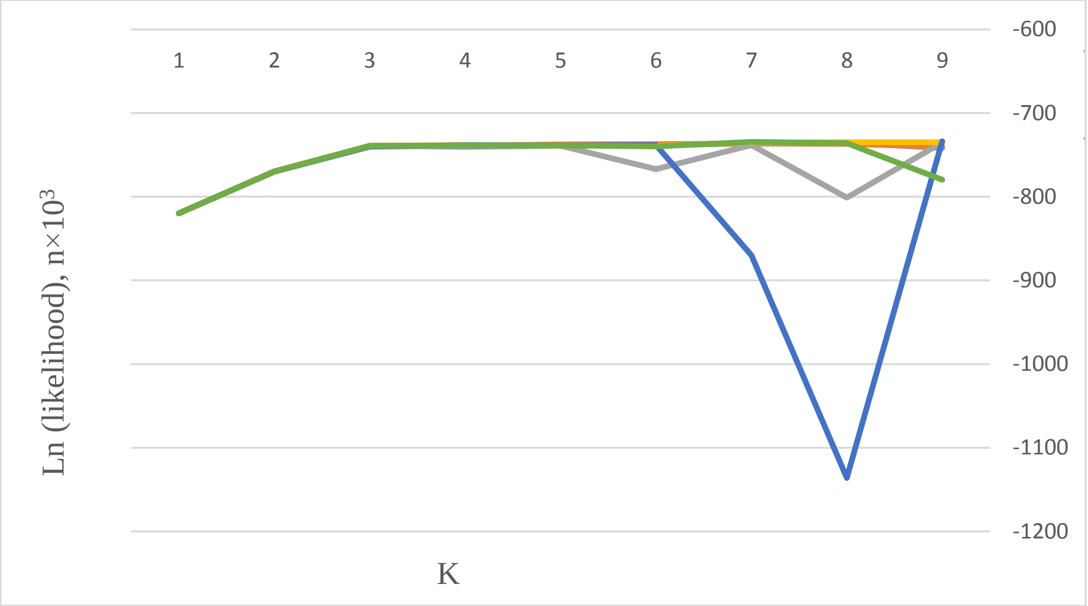


Fig. 1. Distribution of likelihood of model-based structure analyses (K = 1−9) of Lithuanian and HapMap3 populations.

Table 15. F_{ST} distance matrix based on the variation of 106,545 autosomal SNPs (below the diagonal) and *P*-values of ANOVA analysis of PC1 (PC2) (above the diagonal). Statistically significant values in bold

	LIT	**CEU**	**ASW**	**MKK**	**MEX**	**CHD**	**CHB**	**JPT**	**LWK**	**TSI**	**GIH**	**YRI**
LIT	0	**0**	**0**	**0**	**0**	**0**	**1.11×10^{-16}**	**2.22×10^{-16}**	**0**	**1.11×10^{-16}**	**2.22×10^{-16}**	**0**
CEU	0.005	0	**0**	**0**	**0**	**0**	**0**	**1.11×10^{-16}**	**0**	**0**	**0**	**0**
ASW	0.061	0.058	0	**0.0002** (PK2)	**0** (PK2)	**0**	**0**	**0**	**0**	**1.11×10^{-16}**	**0**	**0**
MKK	0.066	0.062	0.013	0	**0**	**0**	**0**	**1.11×10^{-16}**	**2.22×10^{-16}**	**0**	**0**	**0**
MEX	0.029	0.025	0.058	0.063	0	**0**	**0**	**0**	**0**	**9.61×10^{-5}**	**0.0107** (PK2)	**0**
CHD	0.08	0.079	0.088	0.092	0.054	0	0.155089	0.340182	**0**	**0**	**0**	**0**
CHB	0.079	0.078	0.088	0.091	0.053	0.001	0	**0.00071** (PK2)	**1.11×10^{-16}**	**0**	**0**	**0**
JPT	0.081	0.08	0.089	0.093	0.054	0.008	0.007	0	**0**	**0**	**0**	**0**
LWK	0.089	0.086	0.008	0.013	0.082	0.108	0.107	0.109	0	**0**	**1.11×10^{-16}**	**0**
TSI	0.01	0.003	0.058	0.059	0.027	0.079	0.078	0.08	0.084	0	**0**	**0**
GIH	0.033	0.03	0.06	0.062	0.033	0.058	0.058	0.059	0.082	0.03	0	**0**
YRI	0.096	0.093	0.006	0.022	0.089	0.114	0.114	0.115	0.007	0.092	0.089	0

Supplementary material based on exome variation

Table 16. Parameters of primary, secondary, tertiary analysis of exome sequencing data using *LifeScope™ Genomic Analysis Software v2.5*

PARAMETER	VALUE	DESCRIPTION
Analysis Assembly Name	hg 19	(GRCh37–2009-02)
Annotation dbSNP File	dbSNP_b132_00-All.vcf	Build132
Analysis Regions	TargetSeq_exome_named_targets_hg19.bed/ SureSelect_regions2.bed	
Analysis Space	Auto	Base / Color
Analysis Platform	SOLiD	
SAET		
PARAMETER	**VALUE**	
Update Quality Values	True	
Trusted Quality Values	25	
Support Votes	3	
Maximum Correction per Read	0	
K-mer size	0	
Genome Length	30,000,000	
Position of Error Inflation Point	0	
Disable Random Sampling for Large Data	False	
Trusted Frequency	0	
On Target Ratio	0.5	
Number of Recursive Runs	1	
Fragment Mapping		
PARAMETER	**VALUE**	
Mapping QV Threshold	0	
Create Unmapped BAM Files	False	
Reference Weight	8	
Second Map Gapped Algorithm	GLOBAL	
Base Quality Filter Threshold	10	
Map in Base Space	False	
Add Color Sequence	True	
BAM soft clip	False	
BAM Stats		
PARAMETER	**VALUE**	
Whether to combine Data from Both the Strands for Coverage in WIG format	1	
Bin Size for Coverage in WIG file format	100	
Maximum insert size	100,000	
Insert size bin	100	
Primary Alignments only for Coverage in WIG file format	1	
Maximum Coverage	500 000	
Small Indel		
PARAMETER	**VALUE**	
Detail Level	3	
Zygosity Profile Name	Max-mapping	
Genomic Region		
Display Base Qvs	False	
Number Alignment per Pileup	1,000	

Random Seed	94,404
Min Num Evid	2
Max Num Evid	-1
ConsGroup	1
Max Reported Alignments	-1
Min Mapping Quality (MAPQ)	8
Min Best Mapping Quality	10
Min Anchor Mapping Quality	-1
Ungapped BAM Flag Filter	ProperPair, Primary
Gapped BAM Flag Filter	Primary
Edge Length Deletions	0
Edge Length Insertion	0
Perform Filtering	True
Indel Size Distribution Allowed	Can-cluster
Remove Singletons	True
Alignment Compatibility Filter	1
Max Coverage Ratio	12
Max Nonreds 4Fit	2
Min From End Position	9.1
Min Insertion Size	0
Min Deletion Size	0
Max Insertion Size	1,000,000,000
Max Deletion Size	1,000,000,000
Small Indel Annotation	
PARAMETER	**VALUE**
Show Only Variants not in dbSNP	False
Show only coding variants	False
dbSNP concordance for dbSNP SNPs	False
dbSNP Indel border slack	5
Show Only Variants in dbSNP	False
dbSNP concordance for dbSNP Indels	True
Show only variant in genes	False
Enrichment	
PARAMETER	**VALUE**
Extend Bases	0
Min Mapping Score	8
Minimum Target Overlap	0.0001
Minimum Target Overlap Reverse	0.0001
Summary Report	True
Target Coverage Stats	True
Coverage Frequency	False
Coverage Bedgraph	False
Genome Coverage Frequency	False
SNP Finding Annotation	
PARAMETER	**VALUE**
Show Only Variants not in dbSNP	False
Show Only Coding Variants	False
dbSNP Concordance for dbSNP SNPs	True
dbSNP Indel Border Slack	5
Show Only Variants in dbSNP	False
dbSNP Concordance for dbSNP Indels	False

Show Only Variant in Genes	False
SNP Finding	
PARAMETER	**VALUE**
Call Stringency	Medium
Skip High Coverage Positions (Het)	False
Minimum Mapping QV	8
Detect Adjacent SNPs	False
Polymorphism rate	0.001
Include Reads with unmapped Mate	False
Exclude Reads with Indels	True
Require Only Uniquely Mapped Reads	False
Ignore Reads with a Higher Mismatch Count to Alignment Length Ratio	1.0
Ignore Reads with a Lower Alignment Length to Read Length Ratio	0.0
Minimum Ratio of the Filtered Reads and Raw reads	0
Require alleles to be present in both strands	False
Minimum Base QV a Read for a Position	28
Minimum Color QV of a Read for a Position	7
Min Base QV of the non-reference allele of the position	28
Minimum Unique Start Positions of Less Common Allele	0
The Less Common Allele on Both Strands	False
Maximum Difference of Color QVs of the Most Common and Less Common Alleles	99
Minimum Average Color QV of Less Common Allele	0
Minimum Allele Ratio (Het)	0.15
Minimum Coverage (Het)	2
Minimum Unique Start Positions (Het)	2
Minimum non-Reference Color QV (Het)	7
Minimum non-Reference Base QV (Het)	28
Minimum Ratio of Valid Reads (Het)	0.65
Minimum Valid Tricolor Count (Het)	2
Minimum Coverage (Hom)	1
Minimum Count of the Non-Reference Allele (Hom)	2
Minimum Average non-Reference Base QV (Hom)	28
Minimum Average non-Reference Color QV (Hom)	7
Minimum Unique Start Position of the non-Reference allele (Hom)	2
Compress the Consensus File	False
Output Consensus File	True
Output FASTA File	True

LIST OF PUBLICATIONS

Published articles:

(1) I. Lazaridis, N. Patterson, A. Mittnik, <...>, V. Kučinskas, <...>, **I. Uktverytė**, <...>. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature, 2014; 513: 409−13.

(2) V. Kučinskas, **I. Uktverytė**, A. Molytė. Šiuolaikinių molekulinių genetinių ir biostatistinių metodų taikymas dabartinėse ir istorinėse populiacijose. Metodai Lietuvos archeologijoje. Mokslas ir technologijos praeičiai pažinti, Vilniaus universiteto leidykla, 2013, p. 22−64 (ISBN 978-609-459-278-2).

(3) **I. Uktverytė**, A. Molytė, V. Kučinskas. Lietuvos populiacijos etnolingvistinių grupių genetinių ir geografinių atstumų analizė pagal Y chromosomos trumpas tandemines kartotines sekas. Laboratorinė medicina, 2013, t. 15, nr. 1(57), p. 3–8.

(4) V. Kučinskas, **I. Uktverytė**. Genetic variation and genomic origin of Lithuanians. IFEH 12th World Congress on Environmental Health Vilnius, Lithuania 22−27 May 2012. Medimond Publisher proceedings, 2012, p. 7−11, (ISBN-978-88-7587-664-7).

(5) **I. Uktverytė**, O. Balanovsky, E. Balanovska, V. Kučinskas. Genetinė įvairovė tarp Lietuvos etnolingvistinių grupių remiantis Y chromosomos DNR sekų tyrimais. Laboratorinė medicina, 2011, t. 13, nr. 2(50), p. 75–9.

Poster presentations:

(1) **I. Uktverytė**, R. Meškienė, L. Ambrozaitytė, I. Domarkienė, A. Pranculis, N. Burokienė, A. Coj, A. Mažeikienė, V. Kasiulevičius, Z. A. Kučinskienė, V. Kučinskas. LITGEN − revealing genetic structure of the population of Lithuania. Europos žmogaus genetikos draugijos konferencija, Paryžius, Prancūzija, 2013. Vol. 20, Suppl. 2, p. 394 (P16.067).

(2) **I. Uktverytė**, M. Li, M. Stoneking, V. Kučinskas. mtDNA haplogroups in the population of Lithuania. Europos žmogaus genetikos draugijos

konferencija, Niurnbergas, Vokietija, 2012. Vol. 20, Suppl. 1, p. 257 (P10.41).

(3) R. Meškienė, **I. Uktverytė**, J. Arasimavičius, L. Ambrozaitytė, L. Viniarskaitė, A. Irnius, V. Kučinskas. Research of the Nutritional Genomics Markers Specific in the Population of Lithuania. Konferencija „11th Baltic Congress of Laboratory Medicine", Vilnius, Lietuva, 2012. Vol. 14, spec. suppl., p. 42 (ISSN 1392-6470).

(4) **I. Uktverytė**, O. Balanovsky, S. Frolova, M. Kuznetsova, E. Balanovska, V. Kučinskas. The place of the population of Lithuania between Northern and Eastern Europe: Y chromosome analysis. Europos žmogaus genetikos draugijos konferencija, Amsterdamas, Olandija, 2011. Vol. 19, Suppl. 2, p. 345 (P10.79).

Oral presentations:

(1) A. Molytė, **I. Uktverytė**, V. Kučinskas. Kritinė genetinių ir geografinių atstumų analizė pagal Y chromosomos trumpus tandeminius pasikartojimus žmonių populiacijose. Žodinis pranešimas jaunųjų mokslininkų konferencijoje „Bioateitis: gyvybės ir geomokslų perspektyvos", Vilnius, Lietuva, 2012.

(2) **I. Uktverytė**, O. Balanovsky, S. Frolova, M. Kuznetsova, E. Balanovska, V. Kučinskas. Application of Evolving Y chromosome Genetic Markers for Analysis of Lithuanian Population. Tarptautinė konferencija „Evoliucinė medicina: nauji senųjų problemų sprendimai", Vilnius, Lietuva, 2012.

(3) **I. Uktverytė**, O. Balanovsky, S. Frolova, M. Kuznetsova, E. Balanovska, V. Kučinskas. Analysis of the Population of Lithuania using Y Chromosome Genetic Markers. Konferencija „11th Baltic Congress of Laboratory Medicine", Vilnius, Lietuva, 2012.

(4) I. Pepalytė, **I. Uktverytė**, V. Dirsė, V. Kučinskas. Characteristics of the Genomic Structural Variation in the Lithuanian Population. Tarptautinė

konferencija „Evoliucinė medicina: nauji senųjų problemų sprendimai“, Vilnius, Lietuva, 2012.

APIE AUTORĘ

Vardas:	Ingrida
Pavardė:	Uktverytė
Gimimo data:	1985 08 29
Darbovietės adresas:	Santariškių g. 2, LT-08661, Vilnius
Telefono nr.:	+37060064870
El. paštas:	ingrida.uktveryte@mf.vu.lt

Išsilavinimas:

2008−2010 m. Vilniaus universitetas Gamtos mokslų fakultetas Genetikos studijų programa, biologijos magistro kvalifikacinis laipsnis.

2004−2008 m. Vilniaus universitetas Gamtos mokslų fakultetas Molekulinės biologijos studijų programa, biologijos bakalauro kvalifikacinis laipsnis.

2000−2004 m. Klaipėdos „Ąžuolyno“ gimnazija.

1992−2000 m. Klaipėdos „Saulėtekio“ pagrindinė mokykla.

Papildomi mokymai:

2014 m. rugsėjo 8−11 d. kursai „Next-generation sequencing in a diagnostic setting“, Atėnai, Graikija.

2013 m. birželio 8–11 d. Europos žmogaus genetikos draugijos konferencija, Paryžius, Prancūzija.

2012 m. gruodžio 3–4 d. konferencija „Genomics in health and disease - Towards personal genomics“, Naijmegenas, Olandija.

2012 m. rugsėjo 18 d. 4-asis Žmogaus identifikavimo molekulinių produktų vartotojų susitikimas, Ryga, Latvija.

2012 m birželio 12–15 d. tarptautinė mokslinė konferencija „Evoliucinė medicina: nauji senųjų problemų sprendimai“, Vilnius, Lietuva.

2012 m. birželio 8–9 d. konferencija „Gene Forum 2012“, Tartu, Estija.

2012 m. gegužės 10–12 d. konferencija „11th Baltic Congress of Laboratory Medicine“, Vilnius, Lietuva.

2012 m vasario 29 d. – kovo 1 d. kursai „5500 LifeScope Bioinformatics training“, Darmštadas, Vokietija.

2011 m. rugsėjo 5–31 d. mokslinė išvyka į Makso Planko evoliucinės antropologijos institutą, Leipcigas, Vokietija.

2011 m. birželio 13–17 d. mokslinė išvyka į Tartu universiteto Molekulinės ir ląstelės biologijos institutą, Tartu, Estija.

2011 m. gegužės 28–31 d. Europos žmogaus genetikos draugijos konferencija, Amsterdamas, Olandija.

2010 m. spalio 17 d. – lapkričio 20 d. mokslinė išvyka Nacionalinės geografijos projekte „Genografija“ Rusijos Medicinos mokslų akademijos Medicininės genetikos tyrimų centrą (RAMN), Maskva, Rusija.

Narystė profesinėse draugijose:

Lietuvos žmogaus genetikos draugija,

Europos žmogaus genetikos draugija.

ABOUT THE AUTHOR

Name:	Ingrida
Surname:	Uktverytė
Date of Birth:	29 08 1985
Workplace Address:	Santariskiu Str. 2, LT-08661, Vilnius
Mobile.:	+37060064870
E-mail:	ingrida.uktveryte@mf.vu.lt

Academic background:

2008 − 2010 Faculty of Natural Sciences, Vilnius University, Genetics study program. Master of Biology.

2004 − 2008 Faculty of Natural Sciences, Vilnius University, Molecular Biology study program. Bachelor of Biology.

2000 − 2004 Klaipėda "Ąžuolynas" gymnasium.

1992 − 2000 Klaipėda "Saulėtekis" primary school.

Trainings, conferences, scientific visits:

8−11 September 2014, NGS Course "Next-generation sequencing in a diagnostic setting", Athens, Greece.

8–11 June 2013, European Human Genetic Conference, Paris, France.

3–4 December 2012, Conference "Genomics in health and disease - Towards personal genomics", Nijmegen, The Netherlands.

18 September 2012, 4th Baltic HID User Meeting, Riga, Latvia.

12–15 June 2012, International Scientific Conference "Evolutionary Medicine: New Solution for the Old Problems", Vilnius, Lithuania.

8–9 June 2012, Conference "Gene Forum 2012", Tartu, Estonia.

10–12 May 2012, Conference "11th Baltic Congress of Laboratory Medicine", Vilnius, Lithuania.

29 February – 1 March 2012, Course "5500 LifeScope Bioinformatics training", Darmstadt, Germany.

5–31 September 2011, Scientific visit to Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany.

13–17 June 2011, Scientific visit to Tartu University, Institute for Molecular and Cell Biology, Tartu, Estonia.

28–31 May 2011, European Human Genetic Conference, Amsterdam, The Netherlands.

17 October – 20 November 2010, Scientific visit to the Research Centre for Medical Genetics of Russian Academy of Medical Sciences, Moscow, Russian Federation.

Memberships in Professional Societies:

Lithuanian Society of Human Genetics,

European Society of Human Genetics.