VILNIAUS UNIVERSITETAS

Alina
URNIKYTĖ

LIETUVOS POPULIACIJOS GENETINĖS STRUKTŪROS IR EVOLIUCINIŲ VEIKSNIŲ ANALIZĖ, REMIANTIS PLATAUS MASTO GENOTIPAVIMO DUOMENIMIS: PRAEITIS IR DABARTIS

DAKTARO DISERTACIJOS SANTRAUKA

Biomedicinos mokslai,
medicina (06 B)

VILNIUS 2018

Disertacija rengta 2014–2018 metais Vilniaus universiteto Medicinos fakulteto Biomedicinos mokslų instituto Žmogaus ir medicininės genetikos katedroje. Mokslinius tyrimus rėmė Lietuvos mokslo taryba, doktorantūra buvo finansuojama ES struktūrinių fondų lėšomis, 2018 m. buvo gauta stipendija už akademinius pasiekimus.

Mokslinis vadovas
akad. prof. habil. dr. Vaidutis Kučinskas (Vilniaus universitetas, biomedicinos mokslai, biologija – 01 B, medicina – 06 B)

Gynimo taryba
pirmininkas – **prof. dr. Algirdas Utkus** (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B)
Nariai
prof. dr. Loreta Cimbalistienė (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B)
doc. dr. Elena Bosch Fuste (Pompeu Fabra universitetas, biomedicinos mokslai, biologija – 01 B)
akad. prof. dr. Rimantas Jankauskas (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B)
akad. prof. habil. dr. Limas Kupčinskas (Lietuvos sveikatos mokslų universitetas, biomedicinos mokslai, medicina – 06 B)

Disertacija ginama viešame Gynimo tarybos posėdyje 2018 m. gruodžio mėn. 6 d. 10 val. Vilniaus universiteto ligoninės Santaros klinikų Raudonojoje auditorijoje (E122). Adresas: Santariškių g. 2, LT-08661 Vilnius, Lietuva, tel. (+370 5) 2501788; el. paštas genetika(at)mf.vu.lt.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekose ir VU interneto svetainėje adresu: *https://www.vu.lt/naujienos/ivykiu-kalendorius*

VILNIUS UNIVERSITY

Alina
URNIKYTĖ

AN EVALUATION OF THE GENETIC STRUCTURE AND EVOLUTIONARY FORCES OF THE LITHUANIAN POPULATION ACCORDING TO HIGH-DENSITY GENOTYPING DATA: PAST AND FUTURE

SUMMARY OF DOCTORAL DISSERTATION

Biomedical Sciences,
Medicine (06 B)

VILNIUS 2018

This dissertation was written between 2014 and 2018 at the Department of Human and Medical Genetics of the Biomedical Sciences Institute, Faculty of Medicine, Vilnius University. The research was supported by the Research Council of Lithuania; the doctoral studies were financed from the EU structural funds, and a scholarship was granted for academic accomplishments in 2018.

Academic supervisor
Acad. Prof. Habil. Dr. Vaidutis Kučinskas (Vilnius University, Biomedical Sciences, Biology 01 B, Medicine 06 B)

This doctoral dissertation will be defended during a public meeting of the Dissertation Defence Panel:

Chairman – Prof. Dr. Algirdas Utkus (Vilnius University, Biomedical Sciences, Medicine 06 B)
Members
Prof. Dr. Loreta Cimbalistienė (Vilnius University, Biomedical Sciences, Medicine 06 B)
Assoc. Prof. Dr. Elena Bosch Fuste (Pompeu Fabra University, Biomedical Sciences, Biology – 01 B)
Acad. Prof. Dr. Rimantas Jankauskas (Vilnius University, Biomedical Sciences, Medicine 06 B)
Acad. Prof. Habil. Dr. Limas Kupčinskas (Lithuanian University of Health Sciences, Biomedical Sciences, Medicine 06 B)

The dissertation will be defended at a public meeting of the Dissertation Defence Panel at 10 p.m. on December 6, 2018 in the Red Auditorium (E122) at the Vilnius University Hospital Santaros Klinikos.
Address: 2 Santariškių Str., LT-08661 Vilnius, Lithuania.
Tel. (+370 5) 2501788; email: genetika(at)mf.vu.lt.
The text of this dissertation can be accessed through the Library of Vilnius University as well as on the website of Vilnius University at www.vu.lt/lt/naujienos/ivykiu-kalendorius.

CONTENT

1. INTRODUCTION

The contemporary Lithuanian population is composed of a complex mixture of the former Baltic tribes. Thus, the roots of the present Lithuanian population are deep, and it is highly probable that the inhabitants of present-day Lithuania have preserved this ancient genetic composition. The availability of high-throughput genotyping platforms and next-generation sequencing techniques and the development of new statistical and computational methods in the field of evolutionary genomics allow us to infer evolutionary forces and more finely scale the genetic structure of a population. The analysis of geographically specific regions and the characterization of fine-scale patterns of genetic diversity may facilitate a much better understanding of the microevolutionary processes affecting local human populations.

Aim of the Study

An evaluation of the the local patterns of population structure, the signatures of adaptive positive selection and evolutionary demographic parameters from high-density SNP genotyping data generated in the Lithuanian population.

Tasks of the Research

1. To infer and evaluate the genetic structure of the Lithuanian population using high-density SNP genotyping data.
2. To investigate the signatures of positive selection in the Lithuanian population using high-density SNP genotyping data.
3. To infer the demographic parameters of the Lithuanian population and the changes in long-term effective population size including the date of the Lithuanian split in comparison with other populations.

4. To reconstruct past events between the two main ethnolinguistic groups (the Aukštaičiai and Žemaičiai) of Lithuania using effective population size and estimated divergence time.
5. To infer the recent effective population size using inferred long segments of identity by descent and to estimate the effective/census size ratio in the Lithuanian population.

Statements to be Defended

1. The Lithuanian population is homogeneous and genetically differentiated from its neighboring populations but only within the expected general European context.
2. Specific signals of positive selection do exist in the Lithuanian population.
3. The long-term effective population size is small compared to other European populations.
4. A statistically significant difference in effective population size may suggest a potential genetic difference between the Aukstaičiai and Žemaičiai groups.
5. The population of Lithuania is small and has historically suffered the effects of population bottlenecks and expansions, which might produce very small *Ne*/*N* values.

This research contributes to the progress of scientific knowledge and is of special importance in clarifying the relationship between natural selection and disease as well as improving our understanding of the evolutionary mechanisms observed at the individual and population levels.

2. MATERIALS AND METHODS

2.1. Samples

The data set consisted of 425 samples from unrelated Lithuanian individuals who indicated at least three generations of Lithuanian nationality. The samples were collected randomly from six regions of Lithuania: three groups from Aukštaitija (the Western (n = 79), Southern (n = 67) and Eastern (n = 79) regions) and three groups from Žemaitija (the Northern (n = 79), Western (n = 43) and Southern (n = 78) regions) (Fig. 2.1). In accordance with the Declaration of Helsinki, forms of written informed consent were received from all of the study participants.

Genomic DNA was extracted from whole venous blood using either the phenol-chloroform extraction method or the automated DNA extraction platform TECAN Freedom EVO (TECAN Group Ltd., Männedorf, Switzerland) based on the paramagnetic particle method. DNA concentration and quality were measured with a NanoDropR ND-1000 spectrophotometer (NanoDrop Technologies Inc., US).

This study is part of the LITGEN project, which was approved by the Vilnius Regional Research Ethics Committee 235 No. 158200-05-329-79 on May 3, 2011.

Figure 2.1. Map of Lithuanian ethnolinguistic groups.

2.2. Genotyping

Genotyping was performed at the Department of Human and Medical Genetics of the Biomedical Science Institute, Faculty of Medicine, Vilnius University, Lithuania using the Illumina HumanOmniExpress-12v1.1 (296 samples) and the Infinium OmniExpress-24 (129 samples) arrays (Illumina, San Diego, CA, US), which include an overlap of 707 138 SNPs that were distributed genome-wide. Genotyping data quality control was performed according to the standard recommendations of the manufacturer. Individuals and SNPs with > 10% missing data and SNPs with minor allele frequency (MAF) < 0.01 were excluded from the analysis. SNPs with deviations from the Hardy-Weinberg equilibrium ($P < 10^{-4}$) were eliminated from the study. After quality control, 1 individual was excluded with more than 10% missing genotypes (MIND > 0.1), and 532 836 autosomal SNPs remained out of 589 752.

2.3. Data Analysis

2.3.1. Population Genetic Structure Analysis

To characterize the Lithuanian population in a broader genetic context, we merged these SNP genotyping data to those downloaded from the 1 000 Genomes Project Phase3 dataset [1], generating a pooled dataset of 264 950 genome-wide distributed autosomal SNPs in a total of 2 928 individuals from 20 populations and 4 main geographical regions: African populations, including the Yoruba in Ibadan, Nigeria (YRI), the Luhya in Webuye, Kenya (LWK), the Gambian in the Western Divisions in the Gambia (GWD), the Mende in Sierra Leone (MSL) and the Esan in Nigeria (ESN); European populations, including Utah residents with ancestries from Northern and Western Europe (CEU), the Toscani in Italy (TSI), the Finnish in

Finland (FIN), the British in England and Scotland (GBR) and the Lithuanians (LT); East Asian populations, including the Han Chinese in Bejing, China (CHB), the Japanese in Tokyo, Japan (JPT), the Southern Han Chinese, China (CHS), the Chinese Dai in Xishuangbanna, China (CDX) and the Kinh in Ho Chi Minh City, Vietnam (KHV); South Asian populations, including the Gujarati Indians in Houston, Texas (GIH), the Punjabi from Lahore, Pakistan (PJL), the Bengali from Bangladesh (BEB), the Sri Lankan Tamil from the UK (STU) and the Indian Telugu from the UK (ITU).

A principal component analysis (PCA) was carried out with independently pruned SNPs using SmartPCA from EIGENSOFT 7.2.1 [2]. SNPs in linkage disequilibrium were removed with the indep-pairwise option of PLINK v1.07 using a window size of 50 SNPs, a step size of 5 and an r^2 threshold of 0.5 [3]. The PCA was performed for the six ethnolinguistic groups of the Lithuanian sample set alone as well as on the merged Lithuanian 1 000 Genomes Project Phase3 dataset.

Ancestry analysis was performed with ADMIXTURE v.1.3.0 varying the number of ancestral populations K between 2 and 9 [4]. The best K was identified using the cross-error estimation implemented in ADMIXTURE.

2.3.2. The Detection of Signals of Positive Selection

Previous to the analysis of selection, genetic relationships and consanguinity were inferred through the kinship and the inbreeding coefficients, which were estimated with KING v.2.1 [5] and PLINK v1.07 [3], respectively. Negative F values were converted to zero, as they probably represent sampling errors [6]. Outlier samples on the PCA plots and individuals with inbreeding coefficients higher than those expected for offspring from second cousin mating (F values ≥ 0.0156) were removed from the further study.

Genotyping data were phased with SHAPEIT2 [7]. Any signatures of recent or ongoing positive selection were investigated using the locus fixation index (F_{ST}) [8] and the XP-EHH statistic [9], which were computed between the following pairs of populations: LT-CEU, LT-FIN, LT-YRI, CEU-FIN, CEU-YRI and FIN-YRI. The F_{ST} values were calculated using VCFtools v.0.1.13 [10]. XP-EHH was run using Selscan v1.2.0a [11]. For each comparison, a score of XP-EHH per SNP was obtained, and any XP- EHH scores > 2 were considered as indicative of positive selection. We selected as candidates for positive selection any genomic region with two or more SNPs located at the 0.1% top extreme of the XP-EHH genome-wide empirical distribution and with at least one SNP presenting an F_{ST} p-value < 0.01.

Older signals of selection were investigated through the Tajima's D neutrality statistic, which was calculated with the PopGenome package implemented in R [12] considering 100 kb sliding windows across all autosomal regions with a step size of 10 kb. Windows containing missing variants were ignored. As for F_{ST}, the extreme negatives of Tajima's D values were identified considering the rank of the score in the genomic distribution. In particular, the windows were sorted in an ascending order based on Tajima's D values; we considered for further analysis those with empirical p-values less than 0.01.

Variant annotation in the candidate regions for selection was performed with ANNOVAR [13] using GRCh37 (hg19), RefSeqGene, dbSNP147 [14] and CADD version 1.3 [15].

2.3.3. Long-Term *Ne* and Divergence Time Analysis

Long-term effective population size (*Ne*) in the Lithuanian population and their ethnolinguistic groups was estimated using the R package NeON v1.0 based on LD patterns [16]. NeON v1.0 uses binary PLINK files as inputs and updates the genetic map information of the markers to calculate the *Ne* over time, exploiting the relationship between *Ne* and the average squared correlation

coefficient of LD (r^2_{LD}) within any predefined recombination distance categories.

The estimates of *Ne* for the comparison were obtained from a study by M. Mezzavilla and S. Ghirrotto (2015, University of Trieste, Italy), estimated in the HGDP- CEPH panel populations [17] with the R package NeON v1.0 [16].

We merged the Lithuanian population SNP genotyping data to those downloaded from the HGDP- CEPH dataset [17], generating a pooled dataset of 239 325 genome-wide distributed autosomal SNPs in a total of 1 234 individuals from 22 populations and 6 geographical regions: Africa, the Middle East, Central South Asia, Europe, East Asia and Maya. Having the estimates of *Ne* and knowing the population's differentiation, measured by F_{ST} with the software 4P [18], we have also estimated the time of divergence in generations between the Lithuanian and HGDP-CEPH populations and between the six ethnolinguistic groups of the Lithuanian population using the NeON R package [16].

To visualize the evolutionary relationships among the studied populations, a neighbor-joining (NJ) phylogenetic tree was calculated from the divergence time matrix using the R package Phangorn [19].

2.3.3. Recent *Ne* Estimation

To infer the history of the recent effective population size in the Lithuanian population and its ethnolinguistic groups, we used a non-parametric method based on the Wright-Fisher discrete-generation model, implemented in the open-source IBDNe v.04Sep15.e78 software package published by Browning and Browning (2015) [20]. This method is based on the identical by descent (IBD) segments that provide information about the *Ne* around 50 generations from the present one using SNP array data. The length filter used to detect IBD segments with the IBDseq v. r1206 software package was 7 cM.

3. RESULTS

3.1. Analysis of the Lithuanian Population's Genetic Structure

To investigate the genetic similarity between each pair of individuals in the Lithuanian population, we estimated the kinship [3] and inbreeding coefficients [5] using the generated genome-wide SNP data, which were, respectively to the coefficients, 0.00075 and 0.0022. Out of 424 individuals, four had F values higher than expected for offspring from second cousin mating (0.0156). Next, the genetic relationships of the six ethnolinguistic groups in Lithuania were explored at the regional level by performing a PCA using the 232 752 genome-wide-pruned SNPs successfully genotyped in the 424 Lithuanian samples. At that scale, the first two principal components (PC), explaining 22.36% of the variance, showed that the six ethnolinguistic groups formed a single cluster and displayed 8 outliers (Fig. 3.1).

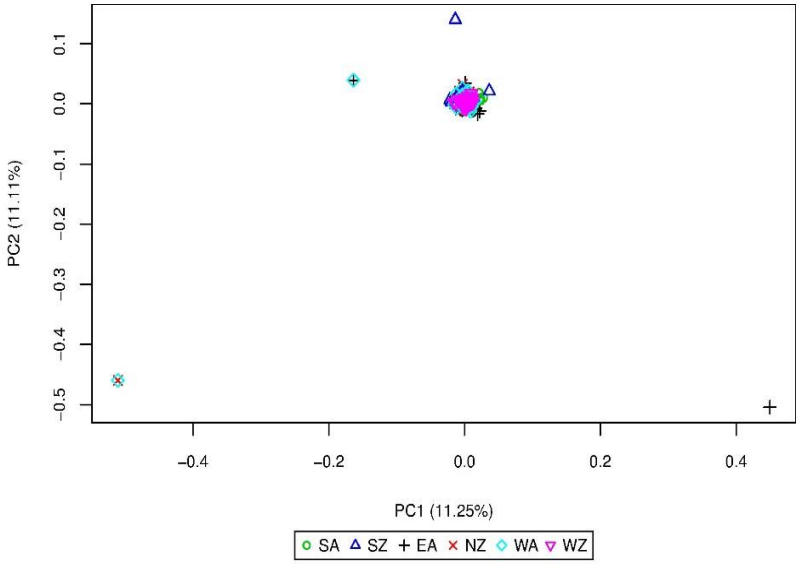


Figure 3.1. A principal component analysis of the six ethnolinguistic groups of the Lithuanian population. EA – Eastern Aukštaičiai, SA – Southern Aukštaičiai, WA – Western Aukštaičiai, NZ – Northern Žemaičiai, SZ – Southern Žemaičiai (SZ), WZ – Western Žemaičiai.

To obtain a European context for the genetic diversity of the Lithuanian population, we then performed a PCA using 158 633 SNPs on a merged dataset with four European populations (CEU, FIN, GBR and TSI) from the 1 000 Genomes Project Phase3 dataset [1] (Fig. 3.2). The first PC, explaining 27.11% of the genetic variance, separates Lithuanians from the four European populations included in the analysis, whereas the second PC, explaining 13.42% of the genetic variance, separates the FIN population, which appears to be more widely dispersed in the plot from the remaining populations. Notably, GBR clusters are grouped together with CEU, whereas the TSI is more closely related with CEU and GBR than with LT and FIN. Finally, Lithuanians lay closer to CEU and GBR than to FIN and TSY.

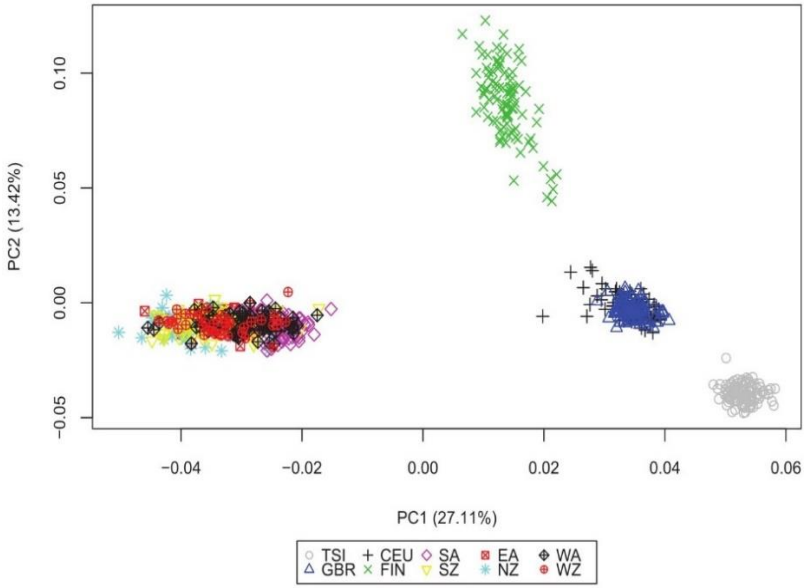


Figure 3.2. Figure 1. A principal component analysis of the six ethnolinguistic groups in Lithuania and the neighboring European populations. EA – Eastern Aukštaičiai, SA – Southern Aukštaičiai, WA – Western Aukštaičiai, NZ – Northern Žemaičiai, SZ – Southern Žemaičiai, WZ – Western Žemaičiai. CEU – Utah residents with ancestry from Northern and Western Europe, FIN – the Finnish in Finland, GBR – the British in Italy and Scotland and TSI – the Toscani in Italy.

Population differentiation was also assessed through a calculation of pairwise F_{ST}. The F_{ST} values between LT-CEU and LT-GBR were similar (0.006 and 0.007, respectively) and LT was most differentiated from TSI (F_{ST} = 0.011). These results correlate with the population substructure observed in the PCA plot of PC1 vs. PC2.

To verify the European context of the genetic diversity of the Lithuanian population, a PCA was then performed including 19 worldwide populations from the 1 000 Genomes Project Phase3 dataset (excluding those of the American origin) [1]. The first two PCs explained 52.7% and 32.3% of the variance, respectively, and showed a clear clustering of all populations according their continent. As expected, the Lithuanian population appeared within the European cluster.

To assess any potential genetic components and population structure within the Lithuanian population, a model-based ancestry analysis was subsequently performed with the ADMIXTURE software [4]. When analyzing the Lithuanian population with four European populations (CEU, FIN, GBR and TSI) from the 1 000 Genomes Project Phase3 dataset [1], the lowest cross-validation error was achieved with 3 ancestry components (Fig. 3.3). At K = 3, one main ancestry component (yellow) is distinguished in the Lithuanian population, which is found at similar proportions along the six ethnolinguistic groups. Notably, CEU, GBR and especially TSI exhibited high proportions of the second largest ancestry component (green), which has only a very low presence in Lithuanians. The FIN population presented the highest proportion of the third genetic component (brown), observed at K = 3, which was also found in considerable proportions in other European populations, such as the GBR and CEU, but at very low frequencies in Lithuanians.

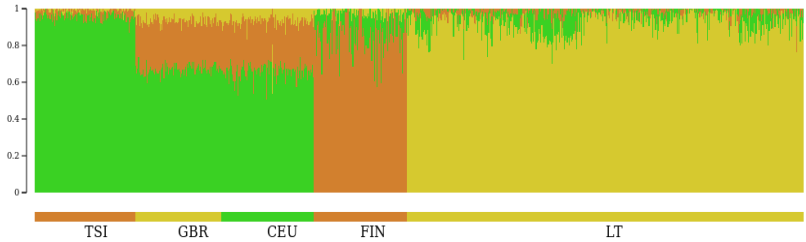


Figure 3.3. The ADMIXTURE plot at K = 3 of the individuals from the LT, CEU, FIN, GBR and TSI populations. Each individual is represented by a vertical colored bar, in which each segment of different color represents the proportion of an individual's ancestry derived from one of the K populations.

Global ancestry profiles in Lithuanians were next inferred including additional worldwide populations from the 1 000 Genomes Project Phase3 dataset [1] (Fig. 3.4). A total of 2 317 individuals from 20 populations and 187 447 SNPs were used in this global ADMIXTURE analysis. The lowest cross-validation error was achieved with eight ancestry components. At K = 2, all African populations (yellow) were distinguished from populations in East Asia, Europe and South Asia (brown). At K = 3, a new ancestry component (green) distinguished Europe and South Asia from East Asia (brown) and Africa (yellow). At K = 4, a new component (yellow) distinguished all the European populations (including LT) from South Asians, and it is not until K = 6 that an ancestry component appears specifically in high proportion in the Lithuanian population (fuchsia). At K = 8 (the lowest cross-validation error), the Lithuanians were characterized by a predominant ancestral genetic component (green) shared in low proportions with other neighboring Europeans (CEU, GBR and FIN) even if they also displayed small ancestry components belonging to South Asia and Africa.

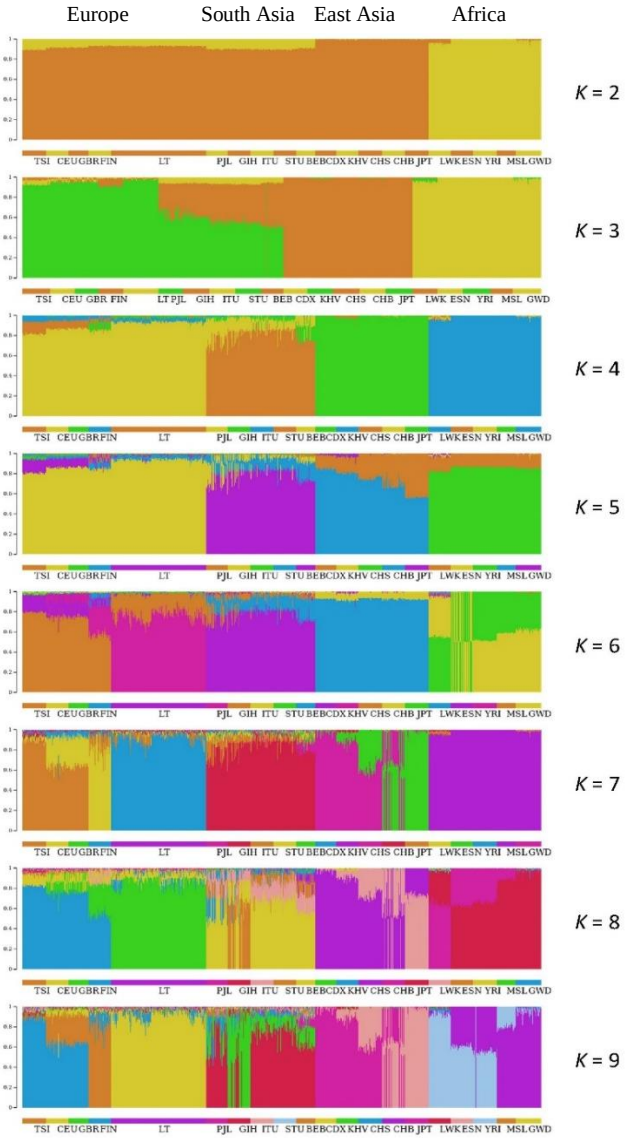
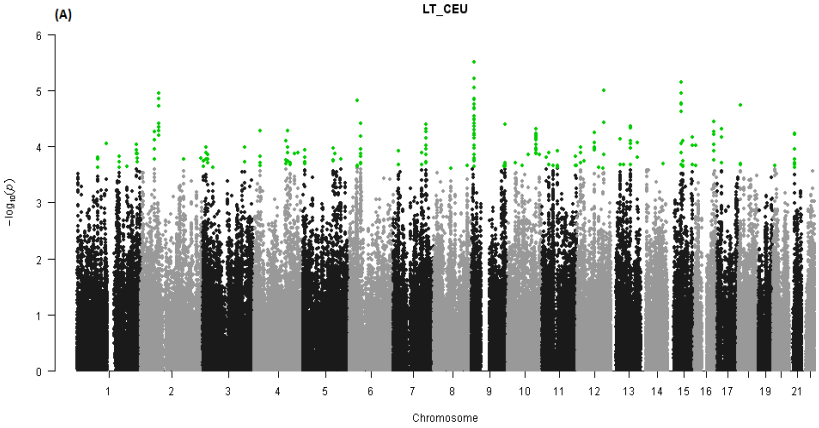
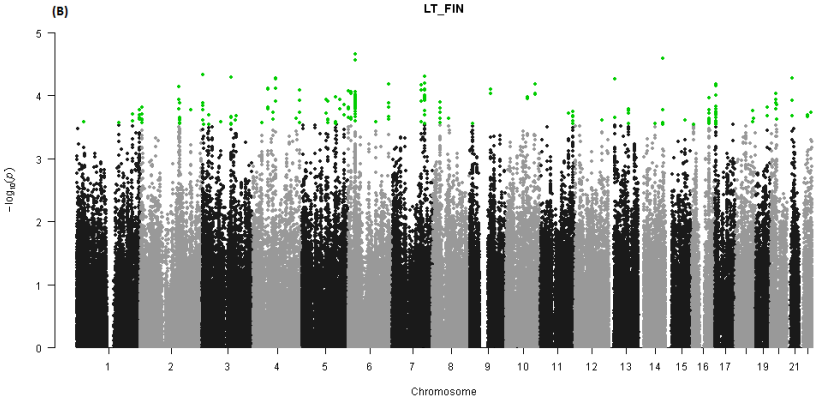


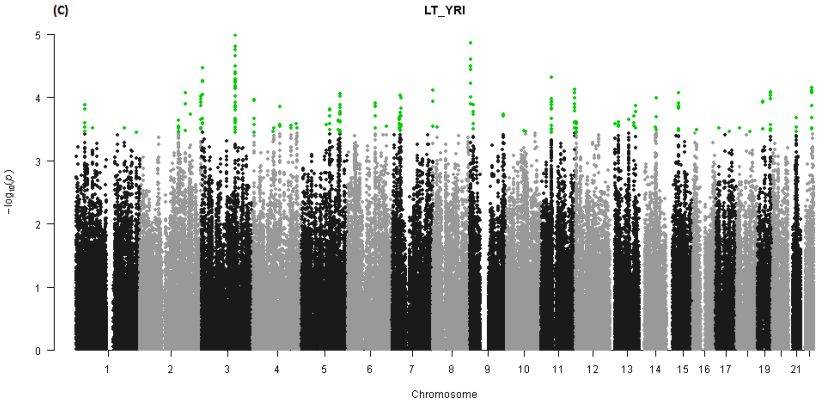
Figure 3.4. An ADMIXTURE analysis of Lithuanians and 19 external populations from the 1 000 Genomes Project Phase3 dataset [1]. ADMIXTURE plots from K = 2 to K = 9 are shown. Individuals are represented as vertical colored bars, in which each segment of different color represents the proportion of an individual's ancestry derived from one of the K populations.

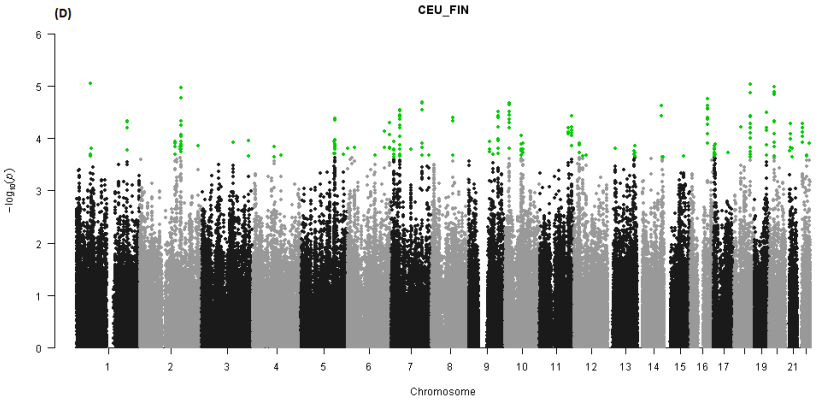
3.2. Signatures of Positive Selection in the Lithuanian Population

We carried out genome-wide scans for different signatures of positive selection. To detect local recent selective events, we calculated F_{ST} and XP–EHH between different pairs of populations (LT-CEU, LT-FIN, LT-YRI, CEU-FIN, CEU-YRI and FIN-YRI) and selected those specific in the Lithuanian population. Per each population comparison, the genome-wide distribution of signals detected with XP-EHH and F_{ST} is shown in Figures 3.5 and 3.6, respectively. We considered as top candidates for recent selection those genomic regions that had presented at least 2 SNPs over the top 0.1% of XP-EHH empirical values and a minimum of 1 SNP with an F_{ST} rank score p-value < 0.01. Out of the 32 signals of recent selection detected in the Lithuanian population, four were shared with other European populations: three between LT and CEU and one between LT and FIN.




(B)
LT_FIN
$-\log_{10}(p)$
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
21
Chromosome


(C)
LT_YRI
$-\log_{10}(p)$
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
21
Chromosome


(D)
CEU_FIN
$-\log_{10}(p)$
0
1
2
3
4
5
6
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
21
Chromosome

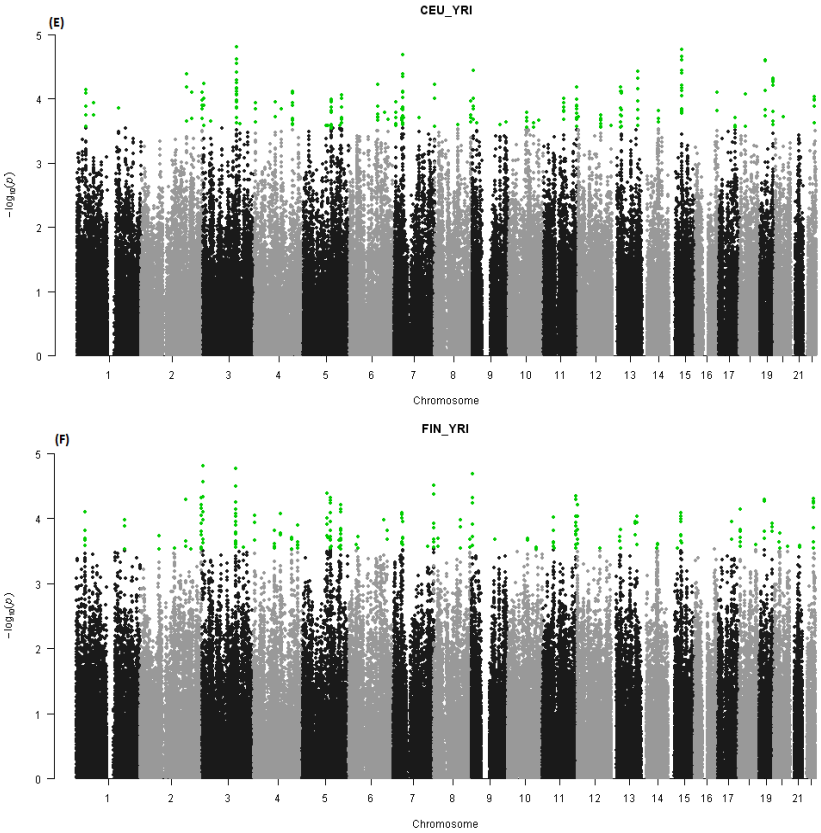


Figure 3.5. A Manhattan plot of XP-EHH signals across the autosomes. (a) XP-EHH in LT-CEU, (b) XP-EHH in LT-FIN, (c) XP-EHH in LT-YRI, (d) XP-EHH in CEU-FIN, (e) XP-EHH in CEU-YRI, (f) XP-EHH in FIN-YRI. In each plot, the green dots indicate 0.1% outlier regions.

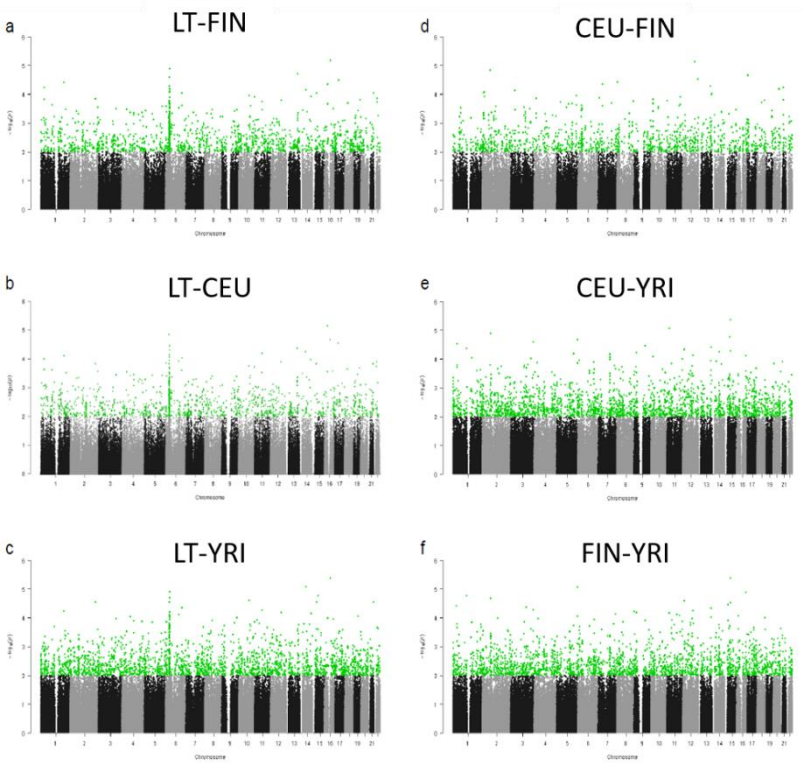


Figure 3.6. A Manhattan plot of F_{ST} p-values. The results are plotted as negative log-transformed empirical p-values of F_{ST} between (a) LT-CEU, (b) LT-FIN, (c) LT-YRI, (d) CEU-FIN, (e) CEU-YRI, (f) FIN-YRI. In each plot, the green dots indicate F_{ST} values with $p < 0.01$.

One of the strongest signals detected with XP-EHH and F_{ST} was found at an ~284 kb region in chr6:27811815–28096280, which comprises several members of the histone and olfactory receptor gene families. Functional variant annotation along the region revealed two non-synonymous SNPs in the *OR2B6* gene: rs7767176 (exon1:c.G349A) and rs9380030 (exon1:c.A809G). Another strong signal of recent positive selection for the LT-CEU comparison was found at a ~184 kb region in chromosome 9 comprising *PTPRD-AS2* and *TYRP1* genes. The *TYRP1* gene encodes a melanosomal enzyme

that participates in the melanin biosynthetic pathway, is involved in lighter skin pigmentation and has been described as a candidate for adaptive selection in European populations only [21]. Notably, the *SLC24A5* gene, which is a known target of recent positive selection related to light pigmentation in non-African populations, was detected as a shared signature in CEU and LT when compared to the Yoruba in our analysis [22]. Another strong signal detected in the LT-YRI comparison was found in a ~225 kb region in chromosome 3, which comprises the *COL6A5* and *COL6A6* genes encoding for the collagen type VI alpha 5 and alpha 6 chains, respectively. Interestingly, a non-synonymous variant in *COL6A5* (rs12488457) with a CADD value of 23.2 was found among the top XP-EHH and F_{ST} outliers along the region. Even if the signal was only significant in the LT-YRI comparison, all European populations present high frequencies (above 0.748) for the derived Pro allele at rs12488457, which is found at very low frequencies in YRI (~0.013). An additional signal in the LT-YRI comparison mapping on chromosome 1 includes two extreme XP-EHH and F_{ST} outliers (rs274750 and rs274752) at the 3' UTR of the *COL8A2* gene, which encodes for the collagen type VIII alpha 2 chain. Again, the signal was only significant in the LT-YRI comparison but the derived alleles at rs274750 and at rs274752, C and A respectively, are nearly fixed in the three European populations and found at intermediate frequencies in YRI. Interestingly, polymorphisms in *COL6A5* have been associated to body mass index [23] and dermal phenotypes, such as eczema and atopic dermatitis [24], while mutations at *COL8A2* have been linked to corneal endothelial dystrophies[25]. Moreover, two signals probably related to immunity have been identified, one in chromosome 6 comprising the *IL26* and *IL22* interleukine genes, and the other in chromosome 12 containing the *BRD2* and *HLA* genes. In particular, one variant downstream *IL22* (rs1182844) was identified as the most differentiated variant with significant XP-EHH values along the region.

By considering the empirical rank p-values that are less than 0.01, we detected up to 36 genomic regions with extreme negative Tajima's D values as potential candidates for older signals of positive selection in the Lithuanian population (Table 3.1). Notably, seven of the detected signals in the Lithuanians where shared with the two external European populations used (CEU and FIN), 5 were shared only with FIN and eight additional signals shared with CEU but not with FIN. The strongest signal specific for the Lithuanian population was identified in chr1:35818960-5948959 with 22 windows in the region and a p-value of 0.0008. Genes found in the region are *KIAA0319l* (Dyslexia-associated protein KIAA0319-like protein) and *ZMYM4* (Zinc finger MYM-type protein 4). Another strong signal with 20 windows in total was detected on chr8:48621077-8801076 with *CEBPD* (CCAAT/enhancer-binding protein delta), *PRKDC* (DNA-dependent protein kinase catalytic subunit) and *SPIDR* (DNA repair-scaffolding protein) genes in the region. Furthermore, a quite strong signal, specific to the Lithuanian population, was detected in chr7:30280729-30470728 with *NOD1* (Nucleotide-binding oligomerization domain-containing protein 1) and *ZNRF2* (E3 ubiquitin-protein ligase ZNRF2) genes.

Table 3.1. Candidate regions under positive selection in the Lithuanian population as detected with Tajima's D statistic.

Genome coordinates	Windows*	P-value	Region	Genes	Shared signal
chr1:35818960–35948959	4 (22)	0.0008	exonic	*KIAA0319L; ZMYM4*	LT
chr1:49988960–50728959	6 (10)	0.0003	exonic	*AGBL4; ELAVL4*	LT
chr1:188788960–188968959	9 (17)	0.0003	intergenic	*LINC01037; BRINP3*	LT; CEU
chr2:21728675–21888674	7 (7)	0.0004	intergenic	*TDRD15; LINC01822*	LT; FIN
chr2:179468675–179648674	9 (10)	0.0003	exonic	*TTN*	LT; CEU
chr3:50343412–51893411	4 (17)	0.0006	exonic	*C3orf18;CACNA2D2;CISH; CYB561D2;*	LT; CEU

Genome coordinates	Windows*	P-value	Region	Genes	Shared signal
				DCAF1; DOCK3;GRM2; HEMK1;HYAL2 ;IQCF3;IQCF6; MANF; MAPKAPK3; NPRL2; RAD54L2; RASSF1; RBM15B; TEX264; TMEM115; TUSC2; ZMYND10	
chr3:128763412–128903411	5 (13)	0.0006	exonic	*CNBP;GP9; ISY1; ISY1-RAB43; RAB43*	LT;CEU; FIN
chr3:143543412–143683411	5 (8)	0.0001	exonic	*SLC9A9*	LT;CEU; FIN
chr4:171930684–172360683	4 (9)	0.0002	ncRNA_exonic	*LINC02431; MIR6082*	LT
chr4:176190684–176400683	11 (15)	0.0005	intergenic	*ADAM29; GPM6A*	LT; FIN
chr5:50531164–50691163	6 (7)	0.0003	exonic	*ISL1*	LT
chr5:126311164–126441163	4 (7)	0.0005	exonic	*C5orf63*	LT
chr6:35265879–35395878	4 (7)	0.0005	exonic	*DEF6; PPARD*	LT; FIN
chr6:97855879–98005878	6 (6)	0.0007	ncRNA_exonic	*MIR548H3*	LT
chr7:30280729–30470728	8 (16)	0.0006	exonic	*NOD1; ZNRF2*	LT
chr7:151730729–151870728	5 (7)	0.0002	exonic	*GALNT11; KMT2C*	LT; CEU
chr8:48621077–48801076	9 (20)	0.0007	exonic	*CEBPD; PRKDC;SPIDR*	LT
chr8:93731077–93901076	8 (9)	0.0002	exonic	*TRIQK*	LT; CEU
chr9:38474202–38614201	5 (6)	0.0001	exonic	*ANKRD18A*	LT; FIN
chr9:125434202–125574201	5 (8)	0.0006	exonic	*OR1K1;OR1L3; OR1L4;OR1L6; OR5C1*	LT
chr10:66065709–	6 (7)	0.0002	intergenic	*REEP3;*	LT;

Genome coordinates	Windows*	P-value	Region	Genes	Shared signal
66215708				*ANXA2P3*	FIN
chr10:118195709–118325708	4 (11)	0.0008	exonic	*PNLIP; PNLIPRP3*	LT
chr11:71554229–71734228	9 (11)	0.0004	exonic	*DEFB131B; IL18BP; NUMA1; RNF121*	LT
chr12:1296235–1426234	4 (6)	0.0009	exonic	*ERC1*	LT; CEU; FIN
chr12:15736235–15886234	6 (11)	0.0003	exonic	*EPS8;PTPRO*	LT; CEU; FIN
chr13:34108565–34278564	8 (11)	0.0005	UTR5	*STARD13*	LT
chr14:64046743–64236742	10 (13)	0.0005	exonic	*SGPP1; WDR89*	LT
chr15:48357093–48477092	3 (4)	0.0005	exonic	*MYEF2; SLC24A5*	LT; CEU; FIN
chr15:69617093–69737092	3 (6)	0.0002	exonic	*KIF23; PAQR5*	LT; CEU
chr16:67321264–67501263	7 (20)	0.0005	exonic	*ATP6V0D1; HSD11B2; KCTD19; LRRC36; PLEKHG4; TPPP3; ZDHHC1*	LT; CEU
chr17:29252345–29392344	5 (5)	0.0002	exonic	*ADAP2; LOC107984974; RNF135*	LT; CEU; FIN
chr18:30389383–30559382	8 (11)	0.0001	exonic	*CCDC178*	LT; CEU
chr19:50580913–50700912	3 (4)	0.0005	exonic	*IZUMO2*	LT
chr20:58399095–58539094	3 (8)	0.0006	exonic	*CDH26; FAM217B; PHACTR3; PPP1R3D; SYCP2*	LT; FIN
chr21:44859932–44989931	3 (8)	0.0005	exonic	*HSF2BP*	LT

*Number of SNPs significant at the top 0.1% of the distribution. The total number of SNPs in the region is shown in brackets.

Further data analysis and studies on data sets are required to confirm such selection signatures.

3.3. Long-Term *Ne* and Divergence Time Analysis

To infer evolutionary relationships between Lithuania and other populations contained in the HGDP-CEPH panel, we reconstructed two human evolutionary forces – effective population size and the divergence time between populations by analyzing LD patterns in the SNP array data with the R package NeON [16]. We analyzed a total of 295 samples from unrelated Lithuanian individuals. The estimates of *Ne* for the populations contained in the HGDP-CEPH panel were obtained from a study by Mezzavilla and Ghirrotto (2015, University of Trieste, Italy), estimated with the R package NeON [16].

The *Ne* values for the Lithuanian population were obtained from 6 000 to 200 generations ago, assuming a generation time of 25 years. The estimated long-term *Ne*, calculated as the harmonic mean [26], is 5 404 for the Lithuanian population with a confidence interval (CI) [4 910; 5 643]. There is variation in *Ne* estimates through time for the Lithuanians (Fig. 3.7). Over the 150 000–25 000 YBP (years before present) period, the *Ne* of Lithuanians was in continuous reduction. The expansion is observed around 25 000 YBP.

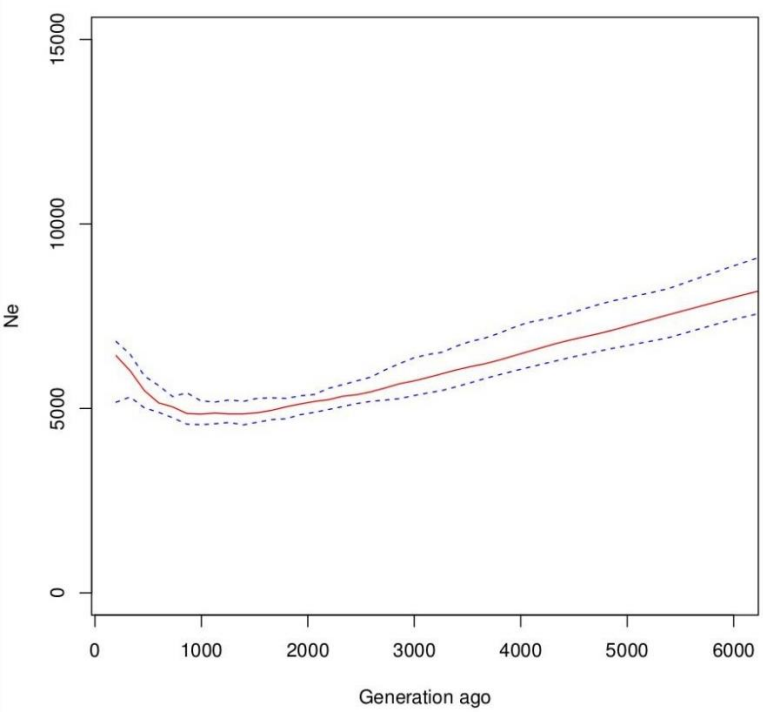


Figure 3.7. Lithuanian effective population size through time estimated from the LD analysis. The x-axis represents the time measured in generations; the y-axis represents the *Ne* values with the confidence interval (the 5th and 95th percentile) values in dashed lines.

The *Ne* estimates and the matrix of inter-population *Fst* values of 23 studied populations were used to reconstruct the divergence times summarized in a neighbor-joining phylogenetic tree (Fig. 3.8). In concordance with other authors, we observed three major groupings: Africans, East Asians and Europeans with Central South and Middle East Asians. The phylogenetic tree provides us with a clear picture saying that the most recent separations of populations and the geographical areas are related.

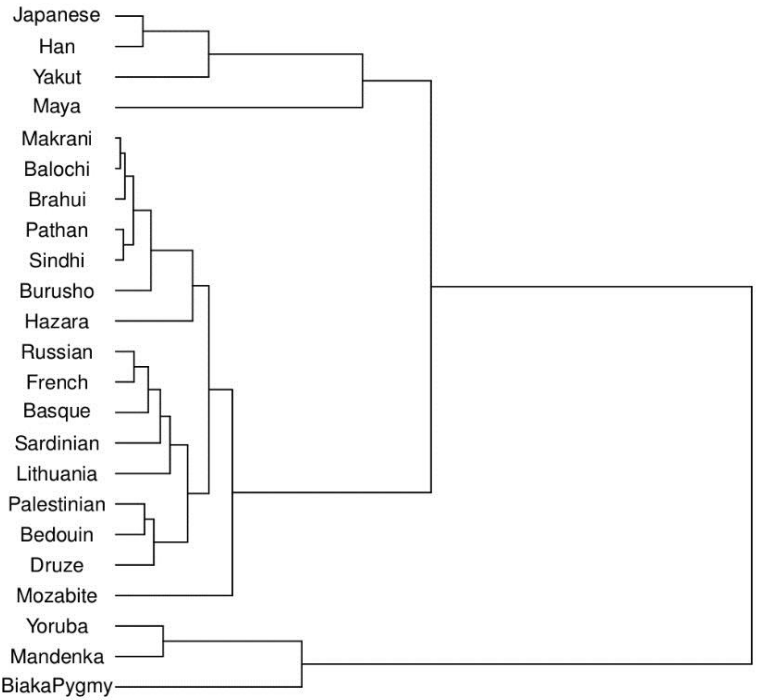


Figure 3.8. A neighbor-joining (NJ) clustering tree, based on the *Ne* and *Fst* estimates, shows the divergence between populations.

The oldest split is observed between African (Yoruba, Mandenka, Biaka Pygmy) and East Asian (Han, Yakut, Japanese) populations, 70 368 YBP, CI [64 036; 76 700]; another separation can be observed in 67 821 YBP, CI [50 989; 84 653], for Africans and Maya comparisons. The average divergence time between African and European populations occurred around 56 536 YBP, CI [54 000; 59 071] as well as between European and East Asian populations in 31 261 YBP, CI [27 089; 35 433]. The most recent separation occurred between European and Middle Eastern ancestors in 7 074 YBP, CI [5 973; 8 176], as well as between the European and Central South Asian ancestors in 8 970 YBP, CI [8 191; 9 750].

Considering the Lithuanian population, we observed that Lithuania first split from Africans in 52 160 YBP, CI [44 169; 60 151], and much later from East Asians – in 26 201 YBP, CI [12 272; 40 129]. The split from Central South and Middle East Asians happened around the same time – in 8 082, CI [7 198; 8 965], and 8 880 YBP, CI [6 179; 11 581], respectively. The most recent genetic separation happened with the Russians in 2 814 YBP and the French in 3 790 YBP. The results also showed that Lithuania was the first population, when compared with other studied European populations (French, Basque, Sardinian), to split from the Middle East Asian population.

In aiming to reconstruct past events between the ethnolinguistic groups of Lithuania, we analyzed the same 295 samples. The estimated long-term *Ne* for each ethnolinguistic group ranged from 4 940 [4 674; 5 304] in the West Žemaičiai (WZ) group to 5 314 [4 829; 5 490] in the West Aukštaičiai (WA) group (Figure 3.9). The difference in the estimated long-term *Ne* values between the two main ethnolinguistic groups (Žemaičiai and Aukštaičiai) of Lithuania was statistically significant ($p < \alpha$, $\alpha = 0.001$, Wilcoxon-Mann-Whitney-Test).

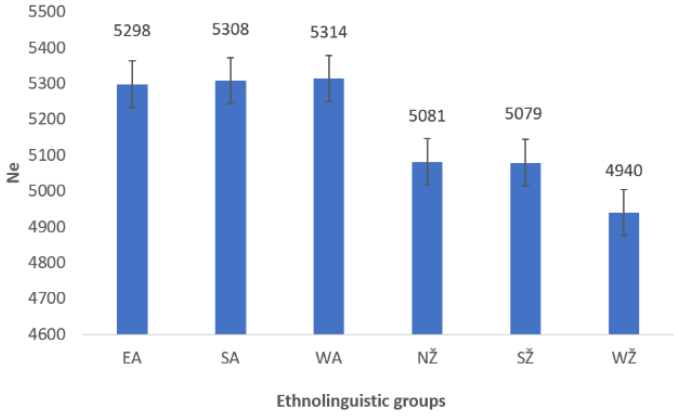


Figure 3.9. Long-term effective population size estimated in the Lithuanian population ethnolinguistic groups with 95% confidence intervals. EA – Eastern Aukštaičiai, SA – Southern Aukštaičiai, WA – Western Aukštaičiai, NŽ – Northern Žemaičiai, SŽ – Southern Žemaičiai, WŽ – Western Žemaičiai.

Considering the *Ne* values through the observed duration, we can conclude that all ethnolinguistic groups of the Lithuanian population suffered similar fluctuations in *Ne*. The *Ne* estimates between the two main ethnolinguistic groups (the Aukštaičiai and Žemaičiai) of Lithuania over the 150 000–37 500 YBP (6 000–1 500 generations ago) period show a constant reduction in *Ne*, possibly as a result of a series of founder effects of the migration of modern humans out of Africa. During the Neolithic Era, all ethnolinguistic groups started to expand; it is known that an intense cultural development occurred in Lithuania during this period as ceramics and farming had first appeared [27]. Distinct growth trajectories observed in WA show the strongest expansion in population size, and WŽ show the strongest reduction in *Ne* during the Neolithic period. Considering the recent *Ne* values, we observed that there were more recent reductions in *Ne* in the Žemaičiai group compared to the Aukštaičiai group.

The estimated times of divergence between the ethnolinguistic groups of Lithuania showed that WŽ is the oldest diverged group, 9 975 YBP (Table 3.2). As expected, the separations happened more recently for ethnolinguistic groups from the same geographical area. The most recent separation occurred between the Northern Žemaičiai (NŽ) and the Southern Žemaičiai (SŽ) groups (4 775 YBP).

Table 3.2. Calculated divergence times, in years, between six Lithuanian ethnolinguistic groups. EA – Eastern Aukštaičiai, SA – Southern Aukštaičiai, WA – Western Aukštaičiai, NŽ – Northern Žemaičiai, SŽ – Southern Žemaičiai, WŽ – Western Žemaičiai.

	SA	SŽ	EA	NŽ	WA	WŽ
SA	0					
SŽ	6 725	0				
EA	7 325	6 450	0			
NŽ	7 125	4 775	5 800	0		
WA	5 850	5 650	5 275	5 225	0	
WŽ	9 975	8 350	9 650	8 250	9 350	0

3.4. Recent *Ne* Analysis

We estimated the recent effective population size for 50 generations (*g*), or 1 250 years from the present, from IBD segments using the newest non-parametric approach [20] on the Lithuanian population (Figure 3.10). One generation is considered to be 25 years. Fifty generations ago, the effective population size in Lithuania was 11 900, whereas from 2015 to 1991, corresponding to generation 0, it was 41 7 000 (95% confidence interval, CI [218 000; 1 150 000]), and the mean census size was 3 373 154. The average estimated *Ne* for generations 30–50 was 16 228. An increased exponential growth is observed from generation 25.

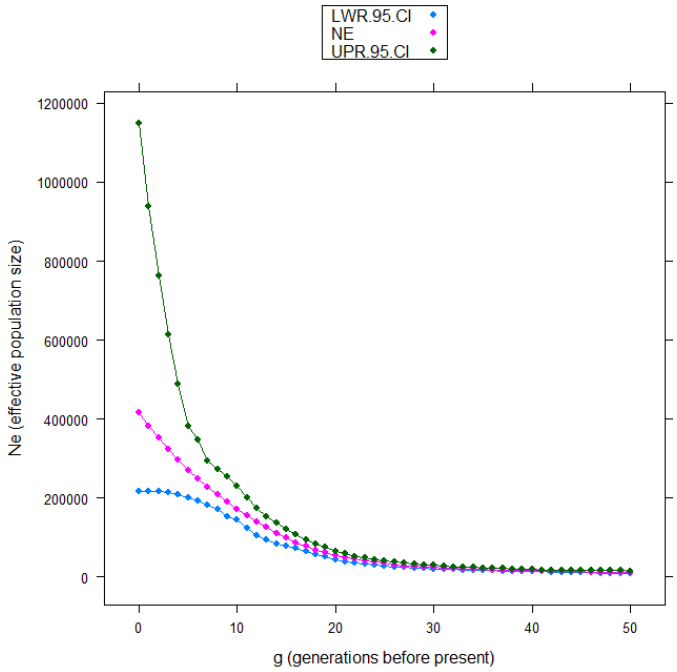


Figure 3.10. Recent effective population size estimated in the Lithuanian population for 50 generations with 95% confidence intervals.

We evaluated the ratio of effective population size to census size, *Ne/N*, in the Lithuanian population at selected time points. We obtained reliable census size data for only three generations of the Lithuanian population from Eurostat, the statistical yearbook of Lithuania (Official Statistics Portal: http://osp.stat.gov.lt/en/statistiko sleidiniu- katalogas?%20publication=1673) and from the e-book of the first Population Census of pre-war Lithuania (Stulginskis and Galvanauskas, 1923) [28]. For generation 1 (1990–1966), the *Ne* was 383 000 (95% confidence interval, CI [217 000; 938 000]), and the mean census size was 3 351 015. Furthermore, for generation 2 (1965–1941), the *Ne* was 352 000 (95% confidence interval, CI [217 000; 763 000]), and the mean census size was 2 816 800. The estimated ratio was 0.125 (95% CI [0.077; 0.271]) for $g = 2$ (corresponding to 1941), 0.114 (95% CI [0.065; 0.280]) for $g = 1$ (corresponding to 1966), and 0.124 (95% CI [0.065; 0.341]) for $g = 0$ (corresponding to 1991) (Figure 3.11). The estimates of *Ne* were approximately at one-tenth of the Lithuanian population size based on the census.

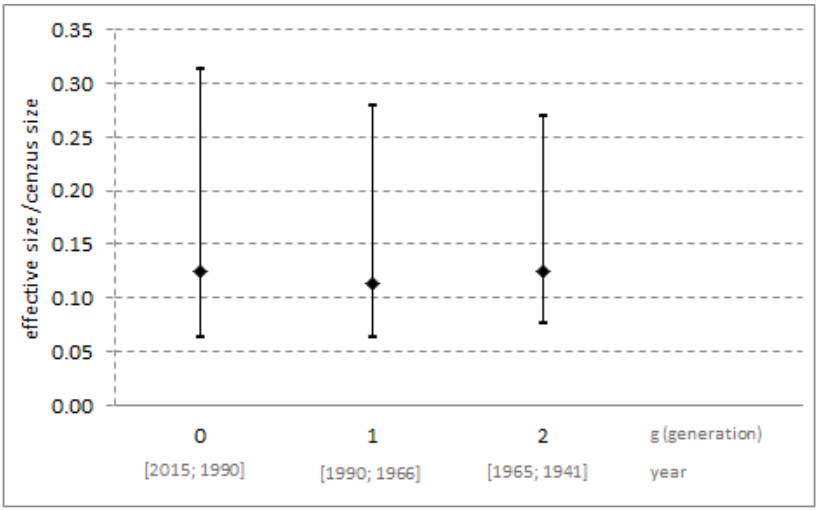


Figure 3.11. The ratio between *Ne/N* in the Lithuanian population for three generations with 95% confidence intervals.

4. DISCUSSION

In this study, we analyzed the local patterns of population structure, the signatures of adaptive positive selection and evolutionary demographic parameters from high-density SNP genotyping data generated in the Lithuanian population. After exploring the genetic relationships among the six ethnolinguistic groups present in Lithuania based on PCA, we found a clear homogeneous genetic landscape across them. The estimated small F_{ST} values among the six geographical regions confirmed their close genetic proximity. In turn, Lithuanians as a whole displayed high genetic similarity to CEU (F_{ST} = 0.006), and their genetic diversity was found in agreement with their European context. The global ancestry profiles obtained with ADMIXTURE when using several external worldwide populations from the 1 000 Genomes Project Phase3 dataset [1] revealed that at K = 8 (the lowest cross-validation error), the Lithuanians have a predominant ancestral genetic component shared at low proportions with other neighboring Europeans (CEU, GBR and FIN), even if they also displayed small ancestral components form South Asia and Africa. Any potential bias on the structure analysis due to the higher sample size of the Lithuanian population was rejected by twice subsampling 120 Lithuanian samples. Overall, these results indicated that Lithuanians are a homogenous population, genetically differentiated from their neighboring populations but within the expected general European context.

We next investigated whether the specific signals of positive selection could be identified in the Lithuanian population. A strong selection signal was identified in the intergenic region of chr9:2352971–12537279 comprising *PTPRD-AS2* and *TYRP1* genes. However, no obvious functional variant have been identified among the SNPs displaying the top outlier XP-EHH and F_{ST} values in that region. Interestingly, polymorphisms in the *TYRP1* gene have been associated with hair and iris colors in European populations [29].

Strong selection signals detected in chromosome 3, which comprises the *COL6A5* and *COL6A6* genes, and in chromosome 1 with *COL8A2* gene. Interestingly, polymorphisms in *COL6A5* have been associated to body mass index [23] and dermal phenotypes, such as eczema and atopic dermatitis [24], while mutations at *COL8A2* have been linked to corneal endothelial dystrophies[25]. Moreover, selection signals related to immunity have been identified (*IL26*, *IL22*, *BRD2* and *HLA* genes).

The results of the long-term *Ne* showed that the *Ne* of the Lithuanian population is quite low – 5 404 – likely the consequence of bottlenecks associated with the Last Glacial period of 25 000–12 000 YBP in Europe [27, 30]. The obtained divergence time estimates between study populations are in agreement with recent studies [16, 31, 32]. Our results support an initial migration from Africa to East Asia in 70 000 YBP and a later dispersal into Europe in around 56 000 YBP and another into the Middle East, Central South and North Asia in around 52 000 YBP. The divergence analyses showed that Lithuania was the first population, if compared with other studied European populations (the French, Basque, Sardinian, and Russian), to have split from the Middle East Asia in around 8 800 YBP. The reconstructed *Ne* between the two main ethnolinguistic groups (the Aukštaičiai and Žemaičiai) of Lithuania showed significant differences between the groups. Indo-Europeans, which had arrived in the Lithuanian territory during the Neolithic period, contributed to the formation of different Baltic tribes and may have had an important influence in the genetic variation and the differences of Lithuanians.

The obtained values of *Ne/N* ratios are small (0.1) compared with other genetics-based estimates of between 0.21 and 0.65 [33]. According to Nunney and Campbell (1994) and Nunney (1996) [34, 35], the *Ne/N* ratio is usually close to 0.5 and only rarely outside the range of 0.25–0.75. However, very low estimates of *Ne/N* (<0.1) raise the possibility that other factors acting to reduce *Ne* have been underestimated, e.g., the variation in female fecundity [34].

Furthermore, natural levels of fluctuations in population size, in the hypothetical absence of any other influence, are often sufficient to depress the *Ne/N* to small values [36]. A single factor is sufficient to produce very small *Ne/N* values, and additional factors tend to depress these values even further [36]. We conclude that natural levels of fluctuations, such as the variance in size, reproduction, sex ratio, as well as the degree to which generations overlap, have probably caused the small values of *Ne/N* in the Lithuanian population. The population of Lithuania is small and has historically suffered the effects of population bottlenecks (a rapid decrease in population size in generation 2) and expansions (a rapid increase in population size in generation 0), which might have produced very small *Ne/N* values.

Furthermore, the small sample size of this study and the Lithuanian population structure could introduce oscillations for the most recent generations. However, considering our results, we think that the true effective size is contained within the bootstrap confidence interval.

Further data analysis and studies on data sets are required to confirm the findings, as well as the selection signatures, of this study.

5. CONCLUSIONS

1. After exploring the genetic relationships among the six ethnolinguistic groups present in Lithuania based on PCA, we found a clear homogeneous genetic landscape across them. Lithuanians are a homogenous population, genetically differentiated from their neighboring populations but within the expected general European context.

2. Signatures of positive selection in the Lithuanian population were investigated over different time frames using three statistics: XP-EHH, F_{ST} and Tajima's D:
   2.1. Candidate regions for positive selection were identified in the Lithuanian population that are related with pigmentation (*SLC24A5*, *TYRP1*) immunity (*IL26, IL22, HLA, BRD2)* and other traits (*COL6A5, COL8A2).*

3. According to long-term effective population size and divergence time estimates:
   3.1. Lithuania was the first population, if compared with other studied European populations (the French, Basque, Sardinian, and Russian), to have split from the Middle East Asia in around 8 800 YBP.
   3.2. According to long-term *Ne* estimates between the two main ethnolinguistic groups of Lithuania (the Aukštaičiai and Žemaičiai), a statistically significant difference was determined.
   3.3. Indo-Europeans, which had arrived in the Lithuanian territory during the Neolithic period, contributed to the formation of different Baltic tribes and may have had an important influence on the genetic variation and differences of Lithuanians.

4. According to the estimated recent effective population size and to the evaluated effective/census size ratio in the Lithuanian population, a small *Ne*/*N* ratio was obtained (0.1).
    4.1 The population of Lithuania is small and has historically suffered the effects of population bottlenecks (a rapid decrease in population size in generation 2) and expansions (a rapid increase in population size in generation 0), which might have produced very small *Ne*/*N* values.

6. REFERENCES

1. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
2. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis**. *Plos Genet* 2006, **2**(12):e190.
3. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M A R, Bender D, Maller J, Sklar P, de Bakker P I W, Daly M J *et al*: **PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses**. In: *American journal of human genetics*. vol. 81; 2007: 559-575.
4. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals**. *Genome research* 2009, **19**(9):1655-1664.
5. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM: **Robust relationship inference in genome-wide association studies**. *Bioinformatics* 2010, **26**(22):2867-2873.
6. Excoffier L, Lischer HE: **Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows**. *Mol Ecol Resour* 2010, **10**(3):564-567.
7. Delaneau O, Zagury JF: **Haplotype inference**. *Methods Mol Biol* 2012, **888**:177-196.
8. Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population structure**. *Evolution* 1984, **38**(6):1358-1370.
9. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R *et al*: **Genome-wide detection and characterization of positive selection in human populations**. *Nature* 2007, **449**(7164):913-918.
10. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*: **The variant call format and VCFtools**. *Bioinformatics* 2011, **27**(15):2156-2158.
11. Szpiech ZA, Hernandez RD: **selscan: an efficient multithreaded program to perform EHH-based scans for positive selection**. *Mol Biol Evol* 2014, **31**(10):2824-2827.

12. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ: **PopGenome: an efficient Swiss army knife for population genomic analyses in R**. *Mol Biol Evol* 2014, **31**(7):1929-1936.
13. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**. *Nucleic acids research* 2010, **38**(16):e164.
14. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic acids research* 2007, **35**(Database issue):D61-65.
15. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nature genetics* 2014, **46**(3):310-315.
16. Mezzavilla M, Ghirotto S: **Neon: An R Package to Estimate Human Effective Population Size and Divergence Time from Patterns of Linkage Disequilibrium between SNPS**. *J Comput Sci Syst Biol* 2015, **8**:037-044.
17. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A *et al*: **A human genome diversity cell line panel**. *Science* 2002, **296**(5566):261-262.
18. Benazzo A, Panziera A, Bertorelle G: **4P: fast computing of population genetics statistics from large DNA polymorphism panels**. *Ecol Evol* 2015, **5**(1):172-175.
19. Team RC: **R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria**. 2014.
20. Browning SR, Browning BL: **Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent**. *American journal of human genetics* 2015, **97**(3):404-418.
21. Hider JL, Gittelman RM, Shah T, Edwards M, Rosenbloom A, Akey JM, Parra EJ: **Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry**. *BMC Evol Biol* 2013, **13**:150.
22. Basu Mallick C, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SY, Gallego Romero I, Crivellaro F *et al*: **The light skin allele of SLC24A5 in South Asians and**

Europeans shares identity by descent. *Plos Genet* 2013, **9**(11):e1003912.
23. Namjou B, Keddache M, Marsolo K, Wagner M, Lingren T, Cobb B, Perry C, Kennebeck S, Holm IA, Li R *et al*: **EMR-linked GWAS study: investigation of variation landscape of loci for body mass index in children**. *Front Genet* 2013, **4**:268.
24. Sabatelli P, Gara SK, Grumati P, Urciuolo A, Gualandi F, Curci R, Squarzoni S, Zamparelli A, Martoni E, Merlini L *et al*: **Expression of the collagen VI α5 and α6 chains in normal human skin and in skin of patients with collagen VI-related myopathies**. *J Invest Dermatol* 2011, **131**(1):99-107.
25. Iliff BW, Riazuddin SA, Gottsch JD: **The genetics of Fuchs' corneal dystrophy**. *Expert Rev Ophthalmol* 2012, **7**(4):363-375.
26. Wright S: **Evolution in Mendelian Populations**. *Genetics* 1931, **16**(2):97-159.
27. Rimantienė R: **Akmens amžius Lietuvoje**. Vilnius, Lithuania: Žiburys; 1966.
28. Stulginskis A. and Galvanauskas E.: **Population de la Lithuanie**. Kaunas: P. Sokolovskiens ir G. Lano Press; 1923.
29. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Jakobsdottir M, Steinberg S, Gudjonsson SA, Palsson A, Thorleifsson G *et al*: **Two newly identified genetic determinants of pigmentation in Europeans**. *Nature genetics* 2008, **40**(7):835-837.
30. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A *et al*: **The genetic history of Ice Age Europe**. *Nature* 2016, **534**(7606):200-205.
31. Tassi F, Ghirotto S, Mezzavilla M, Vilaça ST, De Santi L, Barbujani G: **Early modern human dispersal from Africa: genomic evidence for multiple waves of migration**. In: *Investig Genet*. vol. 6. London; 2015.
32. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A: **Bayesian inference of ancient human demography from individual genome sequences**. *Nat Genet* 2011, **43**(10):1031-1034.
33. Frankham R: **Conservation genetics**. *Annu Rev Genet* 1995, **29**:305-327.
34. Nunney L: **The influence of variation in female fecundity on effective population size**. *Biological Journal of the Linnean Society* 1996, **59**(4):411-425.

35. Nunney L, Elam DR: **Estimating the Effective Population Size of Conserved Populations**. *Conservation Biology* 1994, **8**(1):175-184.
36. Vucetich JA, Waite TA, Nunney L: **Fluctuating Population Size and the Ratio of Effective to Census Population Size**. *Evolution* 1997, **51**(6):2017-2021.

LIST OF PUBLICATIONS

Published articles**:**

1. A. Urnikytė, A. Molytė, V. Kučinskas. Recent effective population size estimated from segments of identity by descent in the Lithuanian population. *Anthropological Science*. 2017, 125(2): 53–58, 2017. DOI: 10.1537/ase.170125.
2. **A. Urnikyte**, I. Domarkiene, S. Stoma, L. Ambrozaityte, I. Uktveryte, R. Meskiene, V. Kasiulevičius. N. Burokiene. V. Kučinskas. CNV analysis in the Lithuanian population. *BMC Genetics* 2016 May 4;17(1):64. doi:10.1186/s12863-016-0373-6.
3. A. Molyte, **A. Urnikyte,** V. Kučinskas. A comparative analysis of mathematical methods for effective population size estimation. *Lietuvos matematikos rinkinys* 2016. Ser.A Vol. 57. p. 53–58. ISSN 0132-2818.

Poster presentations:

1. **A. Urnikyte**, A. Molyte, V. Kučinskas. Relationship between effective population size and inbreeding in the Lithuanian population. The European Human Genetics Conference. June 16–19, 2018, Milan, Italy.
2. **A. Urnikytė**, A. Molytė, V. Kučinskas, Z. A. Kučinskienė. Recent changes in contemporary effective population size from identical by descent segments. American Society of Human Genetics. October 17–21, 2017, Orlando, US.
3. **A. Urnikytė**, A. Molytė, V. Kučinskas. Recent effective population size estimated from segments of identity by descent in the Lithuanian population. The European Human Genetics Conference. May 27–30, 2017, Copenhagen, Denmark.
4. **A. Urnikytė**, A. Molytė, E. Pranckevičienė, V. Kučinskas. Inference of evolutionary relationships among human populations based on the estimates of effective population size.

Society of Human Genetics 66th Annual Meeting. October 18–22, 2016, Vancouver, Canada. *Poster abstracts*, Vancouver: ASHG. 2016. p. 491.

5. **A. Urnikytė**, A. Molytė, E. Pranckevičienė, V. Kučinskas. Demographic inference of the Lithuanian population. The European Human Genetics Conference. May 21–24, 2016, Barcelona, Spain (abstracts book).
6. **A. Urnikytė**, I. Domarkiene, I. Uktveryte, L. Ambrozaityte, R. Meskiene, V. Kučinskas. Genomic diversity and distribuĕon of CNVs in Lithuanian population.European Human Genetics Conference. June 6–9, 2015, Glasgow, Scotland, United Kingdom. *European Journal of Human Genetics*. London: Nature Publishing Group. 2015. Vol. 23. Supplement 1. p. 332–333.

Oral presentations:

1. A. Molytė (presenter), **A. Urnikytė**, V. Kučinskas, A comparative Analysis of Mathematical Methods for homogeneity estimation of the Lithuanian population. Lietuvos matematikų draugijos konferencija. 2018, June 18–19, Kaunas, Lithuania.
2. **A. Urnikyte** (presenter), M. Mondal, E. Bosch, A. Molyte, V. Kučinskas. Detecting signatures of adaptive positive selection from high–density genotyping data in the Lithuanian population. 56th Polish and 14th International Conference Juvenes Pro Medicina. 2018, May 25-26, Lodz, Poland. 3rd place award**.**
3. A. Molyte (presenter), **A. Urnikyte**, V. Kučinskas. A Comparative Analysis of Effective Population Size in Six Ethnolinguistic Groups of the Lithuania Population. XIV Baltic Congress of Laboratory Medicine. 2018, May 10–12, Vilnius, Lithuania.
4. A. Molytė, **A. Urnikytė** (presenter), V. Kučinskas. „Efektyvaus populiacijos dydžio lyginamoji analizė tarp Lietuvos etnolingvistinių grupių“ Lietuvos matematikos draugijos 58-oji

konferencijoje. June 21–22, 2017, Vilnius, Lithuania.

5. **A. Urnikytė** (presenter). Lietuvos populiacijos evoliucinių ryšių analizė (3rd place award). Lietuvos mokslų akademijos Biologijos. medicinos ir geomokslų skyriaus jaunųjų mokslininkų konferencija BIOATEITIS: gamtos ir gyvybės mokslų perspektyvos. December 7, 2016, Vilnius, Lithuania.
6. A. Molytė, **A. Urnikytė** (presenter), V. Kučinskas. A comparative analysis of mathematical methods for effective population size estimation. Lietuvos matematikų draugijos LVII konferencija. June 20 – 21, 2016, Vilniaus Gedimino technikos universitetas, Vilnius, Lithuania.
7. **A. Urnikytė** (presenter), A. Molytė, V. Kučinskas. Recent effective population size estimated from segments of identity by descent in the Lithuanian population Evoliucinė medicina: šiuolakinių sveikatos problemų evoliuciniai mechanizmai ir dėsningumai (Evolutionary Medicine: Pre- Existing Mechanisms and the Patterns of Current Health Issues), trečioji tarptautinė konferencija. June 14–19, 2016, Vilniaus universiteto Medicinos fakultetas; Lietuvos mokslų akademija, Vilnius, Lithuania (published 2016, p. 77, ISBN: 9786094597206).
8. **A. Urnikytė** (presenter). DNR kopijų skaičiaus pokyčių įvairovės ir pasiskirstymo analizė Lietuvos populiacijoje. LITGEN: Lietuvos populiacijos genetinė įvairovė ir sandaros kitimai. susiję su evoliucija ir dažniausiai paplitusiomis ligomis: konferencija. March 6, 2015, Vilnius, Lithuania.

Traineeship:

Evolutionary Population Genetics Lab of the Department of Experimental and Health Sciences at the Universitat Pompeu Fabra in Barcelona, Spain, from 11-02-2017 to 11-05-2018.

ABOUT THE AUTHOR

Name: Alina
Surname: Urnikytė
Present Position: Junior research associate, PhD student at the Department of Human and Medical Genetics, Biomedical Science Institute, Faculty of Medicine, Vilnius University
Address: 2 Santariškių Str., LT-08661, Vilnius, Lithuania
Phone number: (+370 5) 2501788
Email: alina.urnikyte@mf.vu.lt

Education:

2018 PhD studies at the Department of Human and Medical Genetics of the Faculty of Medicine, Vilnius University.
2014 Master's degree in medical genetics at the Department of Human and Medical Genetics of the Faculty of Medicine, Vilnius University (diploma Cum Laude)
2009 Bachelor's degree in Biotechnology, University of Vic (Universidad de Vic), Vic, Spain.
2005 Passed the PAU (university access exams), Spain.
2000 Vincentas Borisevičius Secondary School, Telšiai, Lithuania.

Languages: Lithuanian (mother tongue), Spanish, Catalan, English, Russian.

Membership in societies:

Member of the European Society of Human Genetics.
Member of the American Society of Human Genetics.

Vilniaus universiteto leidykla
Universiteto g. 1. LT-01513 Vilnius
El. p. info@leidykla.vu.lt.
www.leidykla.vu.lt
Tiražas 35 egz.