SIGNATURES OF POSITIVE NATURAL SELECTION IN THE LITHUANIAN ETHNOLINGUISTIC GROUPS FROM HIGH-DENSITY SNP DATA

Master’s Thesis

Systems Biology Master’s Program

Vilnius University

STUDENT NAME:	Karolina Mukauskaitė
STUDENT NUMBER:	1916241
SUPERVISOR:	dr. Alina Urnikytė
SUPERVISOR DECISION:	**………………………………**
FINAL GRADE	**………………………**
DATE OF SUBMISSION:	12 May 2021

CONTENTS

LIST OF ABBREVIATIONS

AK – Aukštaičiai

BC - before Christ (used to indicate that date is before the Christian era)

BMI – body mass index

CEU – Utah Residents with Northern and Western European Ancestry

CHD - Coronary heart disease

DNA - Deoxyribonucleic acid

EHH - extended haplotype homozygosity

FADS - fatty acid desaturases

FIN – Finnish in Finland)

Fst - fixation index

his - integrated haplotype score

HLA - Human Leukocyte Antigen

IBD - identical by descent

LD - LINKAGE disequilibrium

MAF - minor allele frequency

MDS - multidimensional scaling analysis

mtDNA - Mitochondrial DNA (also seen as mDNA)

RFLP - Restriction Fragment Length Polymorphism

SFS - site frequency spectrum

SNP - Single nucleotide polymorphism

STR - short tandem repeat

WHR – waist-to-hip ratio

XP-EHH - cross-population extended haplotype homozygosity

YRI – Yoruba in Ibadan, Nigeria

ZM – Žemaičiai

1KGP – 1000 genome project

INTRODUCTION

The last decade has seen a dramatic increase in the number of human genomic data that has allowed us to study the adaptive history of human populations. Notably, identifying variants, that are subject to selection and are likely to be of functional value, can lead to insights into how genes influence phenotypic variation in humans. Such variants include those that predispose or protect individuals from disease and therefore give information about the development of therapeutic and preventive strategies and support biologically informed drug discovery (Benton et al., 2021).

Local adaptation occurs when individuals in a population have higher average fitness in their local environment than individuals from other populations of the same species due to genetic variation. It is driven by natural selection that differs among populations and leads to genetic and phenotypic population differentiation over time. Differences in genetic diversity are not limited to populations with different backgrounds or the same continent but can also be found within a population (Jakkula et al., 2008). Linguistic, cultural, or geographical differences determine unequal genetic diversity between and within populations. Geographical barriers are considered to be one of the leading causes of genetic diversity in populations (Ashraf and Galor, 2013).

Natural selection has been little studied at the micro-geographic level through dense sampling. Even though exploring how genetic diversity is distributed in geographically close populations, we may complement studies that aim to identify patterns of variation on a larger scale (macro-regional, continental or global) (Esoh et al., 2021).

Lithuania consists of two main ethnolinguistic groups: Aukštaičiai and Žemaičiai. No studies concerning the internal heterogeneity or the potential presence of footprints of natural selection in genomes of inhabitants of these regions have been reported, as lithuanians are considered a highly homogeneous population (Kasperavičiūtė and Kučinskas, 2002). In the research conducted by A. Urnikytė (2019), quantitively slight differences between ethnolinguistic groups were detected, and weak signals of genetic structure were found. However, how these groups differ is yet to find out. In this work, high-density single-nucleotide polymorphisms (SNP) genotyping data of 424 Lithuanian individuals was applied to detect specific genome regions affected by positive natural selection in the main ethnolinguistic groups of Lithuania. The characterization of genetic diversity especially analysing geographically specific regions, may aid in a much better understanding of the microevolutionary processes affecting local human populations (Urnikytė et al., 2019).

The work is relevant because such research expands knowledge of basic science and helps to understand links between natural selection and diseases and the evolutionary mechanisms that lead to the fact that predisposition to certain diseases is different at the individual and the population level.

AIM AND TASKS

Research Question/Hypothesis:

The main objective of this research is to analyze signatures of positive natural selection of two main ethnolinguistic groups, Žemaičiai and Aukštaičiai, from genome-wide SNP genotyping data. The main question is whether natural selection is acting equally on genomic groups under study.

Tasks:

1. To perform population genetic structure analysis of the Lithuanian population.
2. To infer relatedness from genomic data in the Lithuanian population.
3. To investigate signals of positive natural selection in two main ethnolinguistic groups of the Lithuanian population.
4. To identify top significant candidate genomic regions and pathways enriched for signals of positive selection in the Lithuanian population.

1. LITERATURE REVIEW

1.1. The genetic prehistory of Balts

The archaeological records of The eastern Baltics populations show that the first settlement in contemporary Lithuania was founded relatively late. Early inhabitants in this area were reindeer hunters who came to the Eastern Baltic region after the retreat of the continental ice sheet ~12 000 years before today, at the end of the last glaciation As climate conditions improved around 7,500-10,000 years before present in the Eastern Baltic region, the first stable settlements appeared (Gimbutienė, 1985). The oldest settlements in southern Lithuania are about 2,000 years older than the first settlements in Latvia and Estonia (Rimantienė, 1996)

Genomic ancestry studies show that Present-day Europeans have predecessors in three profoundly separated source populations: European hunter-gatherers who inhabited the European continent during the Upper Paleolithic period, Europe's first farmers who migrated from Anatolia across Europe at the beginning of the Neolithic Age around 8,000 BC, and populations from the Pontic Steppe that reached continental Europe in the final Neolithic and early Bronze Age around 4,500 years ago (Lamnidis et al., 2018). Subsequently, most present-day Europeans can be considered a mixture of these three ancestral populations (Patterson et al., 2012).

However, this particular model is inadequate for Baltic populations, from whom lithuanians originated. Anthropological and genetic studies proved that early farmers expanded from Anatolia and mostly replaced hunter-gatherer populations in Europe in the Neolithic period's final years (Lazaridis et al., 2014). The study published in 2017 by Jones et al. provides information that farmers did not mix with hunters-gatherers in the Baltic as they did in Western and central Europe. Therefore, until the great migrations during the Bronze Age, the Baltic hunter-gatherer genome remains exceptionally unchanged (Jones et al., 2017; Mittnik et al., 2018).

After the Neolithic period, historical records reveal that Finno-Ugric tribes enter the Baltic region from the East around 6000–5000 BC. Finno-Ugric-speaking ethnic group is viewed as the ancestors of modern Estonians (Laitinen et al., 2002). The migration wave of steppe pastoralists considered the Indo-European languages in Europe came from the South around 1000 years later (Haak et al., 2015). The genetic component brought by the steppe nomads spread and can be detected in modern-day European populations in a decreasing northeast to the southwest gradient (Lazaridis et al., 2014). It can be found in the genome of present-day latvians and lithuanians (Česnys and Balčiūnienė, 1988).

However, local hunter-gatherers societies were not replaced entirely, left a lasting genetic impact on the ancestry of Eastern Baltic occupants, and can explain why modern-day Eastern Baltic populations' genetic signature is remarkably unique (Haak et al., 2015).

This theory was confirmed by A. Urnikytė (2019). In her research, she included ancient individuals from different periods throughout Western Eurasia to distinguish the genetic signature prehistoric sources that shape the specificity of present-day lithuanians (Urnikytė et al., 2019). Three main different archaic genetic signatures were found: the Early to Middle Bronze Age Steppe pastoralists, pre-Neolithic Hunter-Gatherer groups, and Late Neolithic Bronze Age Europeans. It confirmed the theory that Neolithic movements from Anatolia, which contributed to genetically differentiated populations in Europe, are not particularly prevalent in Lithuania. This research confirmed the unique genetic distinctiveness of lithuanians when compared to European populations (Mittnik et al., 2018).

The Baltic tribes were completely differentiated around 2000–1500 years before the present. According to the collected data, around 500 BC, Finno-Ugric tribes were still living in the northern Curonian lands and Latvia's entire territory to the north of the Daugava River. When the tribes of Semigalian, Selonian, Curonian and Latgalian people and their languages began to blend into the one population in the north, the formation of the Latvian nation began. East Baltic tribes remained in the South and constituted a rise to the Lithuanian population (Kučinskas, 2001).

The Lithuanian population was divided into two large groups. One of them was the group who lived around the mouth of Nemunas river called Žemaičiai ("Lowlanders"), and another group was the Aukštaičiai ("Highlanders") who inhabited lands further upriver to the east. Both of these groups were composed of several tribal territories or lands. The territory of Lithuania was also surrounded by other Baltic tribes that were closely related to the lithuanians. The Scalvians, Yotvingians were living in the west and southwest. The Prussians were inhabiting the territory of today's north-eastern Poland and the Kaliningrad. The Curonians occupied the lands along the western coast of present-day Latvia and Lithuania (Kasekamp, 2018).

The formation and relationships of Baltic tribes led to the development of the different Lithuanian dialects and, eventually, to the regional linguistic differentiation. There are six different dialects distinguishable in present-day Lithuania: three groups from Žemaičiai (north, west, and South) and three groups from Aukštaičiai (west, south, and east (Urnikytė et al., 2019).

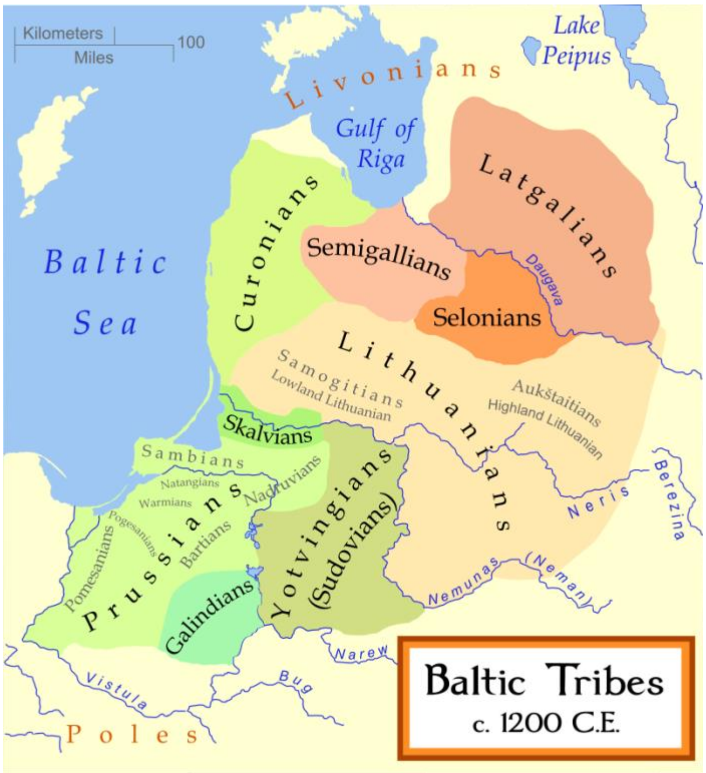


Figure 1.1: The map of the Baltic tribes. The Eastern Balts are shown in brown tones, and the Western Balts are shown in green. The boundaries are approximate based on a map by Marija Gimbutas, published in *The Balts* (1963).

1.2. Structure of the Lithuanian population

The influence of the different tribes of Balts suggested a theory that lithuanians could be a heterogeneous population. All ethnolinguistic regions in the state have distinct dialects and differences in cultures and traditions. However, collected anthropological data of the Lithuanian population led to the conclusion that differences between Aukštaičiai and Žemaičiai disappeared in the Middle Ages. Onwards, the Lithuanian nation was viewed as a very homogeneous population (Česnys and Balčiūnienė, 1988).

To investigate whether the influence of the different tribes can be demonstrated in the gene pools of current Lithuanian ethnolinguistic groups, individuals from six different ethnolinguistic groups were analyzed. Physical features (dermatoglyphics), serological markers (12 blood group systems, 3 systems of serum proteins), and modern DNA markers (mtDNA, genomic Y chromosome markers, and whole-genome markers: STR, SNP) were studied.

The first research that detected differences between occupants of the separate geographic regions in Lithuania was conducted by N. Klevcova and V. Kučinskas (1987) and focused on phenetic distances - dermatoglyphic features of fingers and palms. Lithuanian regions were found to differ significantly. It was noticed that according to dermatoglyphic features, maximum genetic distances were found between south Aukštaičiai and other Lithuanian ethnolinguistic groups (Kučinskas, 2004).

Differences between ethnolinguistic groups were also determined by studying the distribution of AB0 and Rh (D) blood groups in the Lithuanian population. Based on the obtained results, Žemaičiai is the most homogeneous as an ethnolinguistic group and as a region (Kučinskas et al., 1994). These results confirm that the Žemaičiai are autochthons and that Aukštaičiai is eastern Balts that immigrated later (Kučinskas, 2004).

By analyzing Alu sequences, statistically, significant differences were determined between Northern Žemaitija and Southern Aukštaitija, and results of the first mtDNA restriction fragment length polymorphism studies (RFLP) showed minimal differences between Aukštaičiai and Žemaičiai (Kučinskas, 2001, 1994).

Ethnolinguistic groups were studied by mtDNA and Y chromosome, and statistically significant differences between groups were not identified (Kasperavičiūtė and Kučinskas, 2002).

Domarkienė I. (2014) conducted exciting research to find clinically significant and specific for the Lithuanian population genomic markers of Coronary heart disease (CHD). No statistically significant differences between Aukštaičiai and Žemaičiai were found. Consequently, it can be said that according to the number of tested risk alleles per person, the Lithuanian population is homogeneous.

A group of 253 individuals was also studied using whole-genome SNP analysis. A dendrogram of six Lithuanian ethnolinguistic groups was formed using a hierarchical clustering method based on genetic distances (Fig. 1.2). Three groups were formed: Northern, Western, Southern Žemaitija, Eastern and Western Aukštaitija, and Southern Aukštaitija. The last group (Southern Aukštaitija-Dzūkija) has the most considerable distance from the other two groups (Uktverytė, 2014).

Summarising the obtained results from all of the researches, it was stated that the southern Aukštaičiai group stood out from all Lithuanian ethnolinguistic groups (Kučinskas, 2017). Knowing the tribes that were inhabiting Lithuania, the differences could be explained by the fact that the Northern Žemaičiai and the inhabitants of the Curonian Spit of Latvia have taken over a part of the ancient Curonian gene pool. The data of the southern Aukštaičiai suggest that they inherited part of the Jotvingian gene pool (Kučinskas, 1994).

The most recent study was conducted by A. Urnikytė (2019), and results display lithuanians as a homogeneous population. She also concluded that the genetic differentiation between the Lithuanian ethnolinguistic groups is quantitatively small, even if statistically significant.

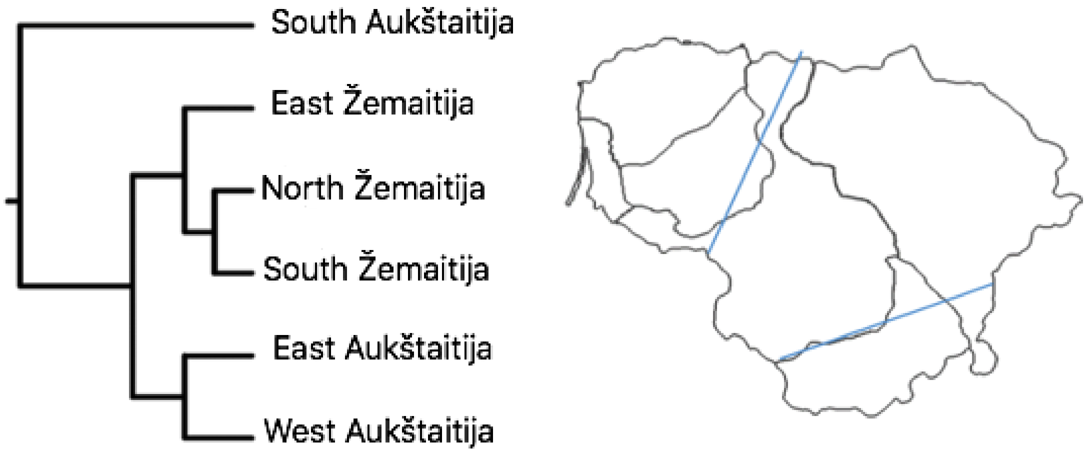


Figure 1.2: Dendrogram of Lithuanian ethnolinguistic groups, composed based on autosomal SNP data. Ethnolinguistic groups are separated into clusters by the lines in the picture on the right (Uktverytė 2014).

1.3. Natural selection

Natural selection is the primary source of evolution. Individuals who are better adapted to their environment tend to survive and reproduce more successfully, changing the distribution of heritable biological traits in the population. Other factors, for example, mutation and genetic drift, are also critical evolutionary factors. However, selection plays a unique role in driving adaptation, the evolutionary changes in response to environmental stimuli. Therefore, studying natural selection is critical in understanding how populations adapt to the environment (Balding et al., 2019). For example, the indigenous Bajau ("Sea Nomads") people of Southeast Asia practice a traditional marine hunter-gatherer lifestyle and are the first known people genetically adapted to diving (Ilardo et al., 2018).

Detection of selection signals also is very significant in the medical field. As selection acts on the phenotype, segments subject to selection are often of functional importance. For example, it can be associated with resistance to pathogens or genetic diseases (Balding et al., 2019). Pathogens have been shown to harbour genes that are continually evolving. As immune systems and host genes themselves adapt over time, evolution gives them the ability to remain virulent. Pathogens indeed are thought to be the primary selective pressure that has left its mark on the human genome. It is suggested that the mortality of diseases such as plague explains the presence of widespread, unidentified selection signals in the human genome (Corona et al., 2018).

1.4. The different types of natural selection

The types of natural selection have been clearly described by Kimura (1983). The removal of deleterious mutations is called negative or purifying selection because if a mutation is deleterious, it soon disappears from the population. In the same way, the selection of favorable mutations is called positive selection. However, the most common instance in evolution is so-called neutral selection. Typically, selection could be classified into these types (Loewe, 2008):

- Stabilizing selection - the stabilization of the most common variant in the population. This selection occurs when the population stabilizes on a variant that is not an extreme case. It is one of the most common natural selection types because most variants do not change drastically in frequency.
- Directional selection - selection in which one of the alleles and its variants are favored, resulting in a shift toward that allelic variant over time.
- Disruptive selection - this selection occurs when a change in the environment increases the rare variant's fitness. The most common variant becomes relatively more minor compared to the formerly rare variants.
- Balancing selection - in this selection, all alleles in the population remain stable because they are approximately equal in fitness, so all alleles remain in the population, increasing variability.

Although many different types of selection processes are recognized, most research has focused on developing methods to detect instances of positive selection (Fan et al., 2016). One reason for this is practical, as detecting positive selection may be more accessible due to the more conspicuous signature it leaves in the genome (Vitti et al., 2013). A positive selection might be the main focus of many research pieces because it could be considered a primary driver of local adaptations, especially in human history (Figure 1.3) (Fan et al., 2016; Vitti et al., 2013).

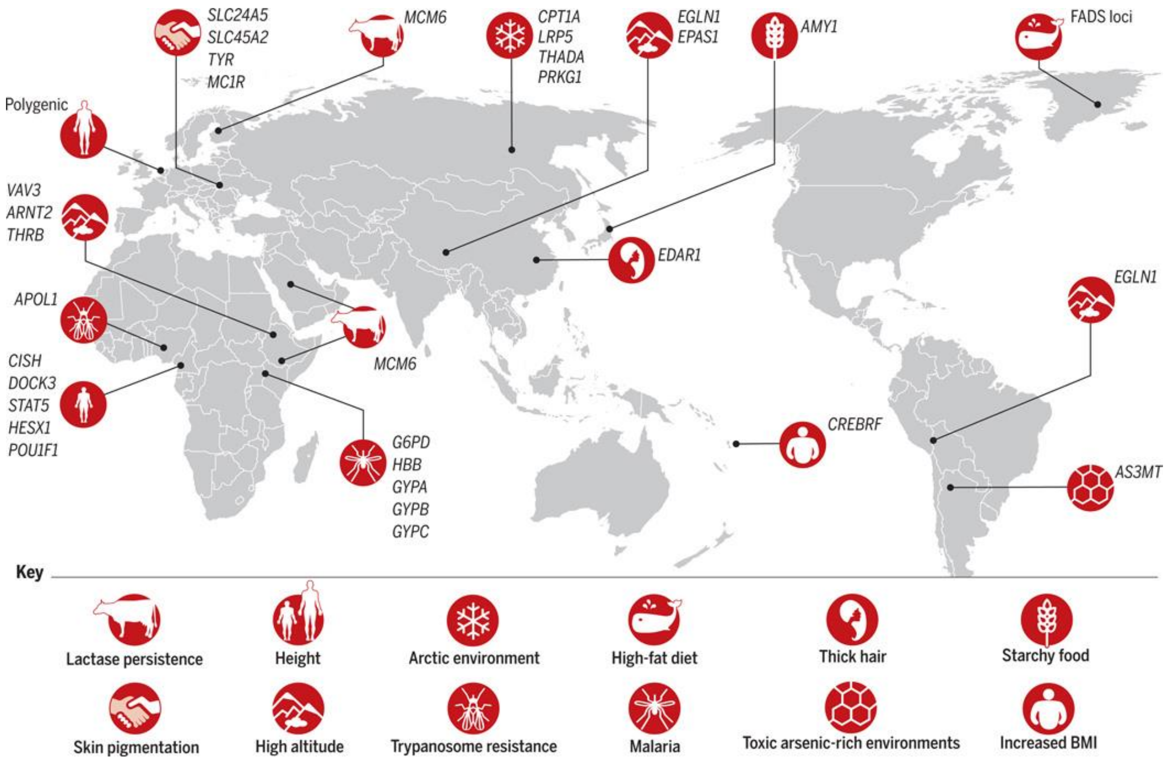


Figure 1.3**:** Examples of recent human local adaptation. Each example includes the candidate gene under selection, the phenotype, and/or the selective pressure (Fan et al., 2016).

Positive natural selection can lead to a smaller genomic diversity at the selected locus and loci linked to it. That produces a characteristic signature of increased expected haplotype homozygosity. Detection of indication of positive selection in genomes is accomplished by searching for signatures caused by selective sweeps. These selective sweeps are commonly classified into two types - soft and hard. Classification is based on their influence on the population. Hard selective sweeps influence the population and leave the selected region only in the population of interest. That is called fixating. On the contrary, soft selective sweeps fixate only partially. Hard selective sweeps can be recognized as an excessive number of variants in low frequencies (Pavlidis and Alachiotis, 2017).

Selective sweeps are significant in the modern world, especially as they are considered critical factors in disease-causing bacteria and viruses' ability to affect their hosts and remain despite the medicines used to treat them. Bacteria, and especially pathogenic bacteria, live very shortly, meaning that natural selection will cause beneficial allele, which makes bacteria more pathogenic, disperse rapidly, and the less dominant haplotypes will be diminished. The promise of very fast change combined with intense selection pressure from external forces such as antibiotics or antivirals results in many selective spreads. Natural selection and the brief life spans of the pathogens have caused selective sweeps in influenza and Toxoplasma gondii (Kirkpatric, 2016).

The selective sweeps in bacteria have proven to be harmful to humans, but selective sweeps can help humanity likewise. For example, there is evidence that the selective sweeps should be held accountable for uniting a differing population into what we know as modern-day corn. Farmers selected only optimal offspring's applying artificial pressure and forced a rapid evolution of corn. The optimal traits are chosen, and the intense evolutionary pressure produced a selective sweep. The selective sweep, in this case, proved most beneficial to humans (Kirkpatric, 2016).

1.5. Methods to detect positive natural selection

Selective sweeps alter many measures of genetic variation. Most commonly, methods that focus on three different measurements are used in the analysis: linkage disequilibrium (LD), site frequency spectrum (SFS,) and population differentiation-based tests.

1.5.1. Linkage disequilibrium-based methods

Most of the LD-based methods focus on finding a specific signature of a selective sweep of a particular LD pattern, which occurs between SNPs in the neighbourhood of the target site for positive selection. It focuses on the loci of beneficial mutation. When mutation rises in frequency, LD levels rise on each side of the particular area. In contrast, decreased LD levels are spotted between sites located on different sides of the selected region. The high LD levels on different sides of the selected locus can be detected because of the recombination in the genome. It lets existing polymorphisms on the same side of the sweep avoid the sweep. However, polymorphisms located on the different sides of the chosen locus require a minimum of two recombination occurrences to avoid the sweep. Since the recombination is happening independently, the observed value of LD is reduced between SNPs located on different sides of the positively selected mutation (Pavlidis and Alachiotis, 2017).

LD-based methods to detect natural selection are very handy for identifying variants with a partial or incomplete selective sweep. A new mutation rises to a reasonable frequency in the population rather than achieving fixation. One of the developed tests is extended haplotype homozygosity (EHH). It estimates the probability that two randomly selected haplotypes are identical up to a distance x around a particular candidate single nucleotide polymorphism (SNP) called the core SNP. The rate at which EHH decays to 0 with respect to x reflects the core SNP's age, and that depends on selection strength (Sabeti et al., 2002). The slow decay of EHH shows the recent selection. So, the region surrounding the selected allele often is more depleted of variation and recombination, and in that case, EHH decays more slowly than under selective neutrality. Therefore, a raised level of EHH around a core SNP helps detect loci under positive selection (Balding et al., 2019).

The method widely used to detect natural selection is the cross-population extended haplotype homozygosity (XP-EHH) statistic. This method takes into comparison lengths of haplotype between populations to control local variation in recombination rates (Vitti et al., 2013).

Another analysis based on EHH is the Rsb test, used to detect such signatures of recently or almost completed selective sweeps. This test finds haplotypes that are selected positively in one population by estimating the length of haplotypes around each allele at a core SNP, then comparing these lengths between populations. The Rsb test can detect genetically differentiated regions across the two populations (Pavlidis and Alachiotis, 2017).

1.5.2. Site frequency spectrum-based methods

Site frequency spectrum (SFS) based methods rely on the belief that frequency of variants is affected by selective sweeps in predictable ways, i.e., a higher proportion of high and low-frequency variants and a smaller proportion of intermediate-frequency variants. Many tests compare different estimates of scaled measures of genetic diversity (θ) depending on the different SFS aspects. The estimates of θ should result in similar values when neutral because if selective sweeps occur, the estimates are unequal. Other tests that compare the SFS with neutral assumptions of specific genomic regions based on the entire data set or data from a second population can be used. Parts that experienced natural selection will deviate strongly from neutral assumptions. However, the frequency spectrum gets back to baseline gradually. The alterations, however, can be distinguished even after several hundred thousand years in human populations, meaning that it stays for thousands of generations (Pavlidis and Alachiotis, 2017).

The pioneer method to detect a signal of natural selection was Tajima's D statistics, and it is the most commonly used test. Tajima's D test depends on analyzing the number of pairwise differences between individuals compared to the total number of segregating polymorphisms (Tajima, 1989). When studying a sample, low-frequency alleles contribute less to the number of pairwise differences than do alleles of intermediate frequency. An excess of rare alleles inflates the value disproportionately to the former. In such a way, smaller (more negative) Tajima's D values indicate an excess of rare alleles, indicating a positive selection or population expansion. Several variations of this method have been developed to account for each allele's polarity and to measure the frequency of rare alleles in different ways (Vitti et al., 2013).

1.5.3. Population differentiation-based methods

Tests on population differentiation are based on the assumption that populations occur in diverse environments and different selection regimes. When positive selection is suspected, advantageous alleles and variants linked to it are at very high frequencies in applicable populations, while intermediate to low frequencies are expected in contrast. Thus, disproportionately high population differentiation is expected for these loci compared to neutral loci. The simplest method is to perform a comparative analysis of nucleotide diversity between regions of genomic populations. If the incidence of SNP in a single population is very high compared to the other group of people, it can be assumed that the reason is natural selection (Weigand and Leese, 2018).

Wright's fixation index (F_{ST}) is the most advanced population differentiation-based method used to detect the signature of selection using genome data. This method is based on the variance of allele frequencies and can be used to compare those frequencies within or between the populations. Large values of F_{ST} statistic signify strong differentiation between populations and suggest the directional selection. Comparatively small values indicate that the compared populations are homogeneous and indicate balancing or directional selection in both (Vitti et al., 2013).

Table 1. Overview of different methods commonly applied to detect micro-evolutionary patterns of positive selection (based on Weigand and Leese, 2018).

Method	**Type of data**	**Method**	**Type of data**	**Method**	**Type of data**
Linkage disequilibrium-based methods		**Site frequency spectrum-based methods**		**Population differentiation-based methods**	
LRH	Haplotypes	**Tajima's *D***	Genotypes or allele frequencies	**FDist**	Allele frequencies
iHS	Haplotypes	**Fu and Li's tests**	Genotypes or allele frequencies	**BayeScan**	Allele frequencies
XP-EHH	Haplotypes	**Fay and Wu's *H***	Genotypes or allele frequencies	**FLK**	Allele frequencies
Rsb	Genotypes or haplotypes	**CLR**	Allele frequencies		
H12	Genotypes	**XP-CLR**	Allele frequencies		
ω statistic	Haplotypes	**Pool-HMM**	Pooled sequence data with the quality score		
HapFLK	Genotypes				

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="6">CLR, composite likelihood ratio test; FLK, extended Lewontin and Krakauer test; iHS, integrated haplotype score; LRH, long-range haplotype test; XP-CLR, cross-population composite likelihood ratio test; XP-EHH, cross-population extended haplotype homozygosity.</td></tr>
</table>

1.6. Challenges in detecting natural selection

While each method of natural selection detection has its specific strengths and limits, there are many challenges that these tests share, particularly in interpreting their results. There might be many other probable explanations other than natural selection for the observed genomic outcomes. For instance, demographic events (migration, expansions, and bottlenecks) can frequently produce selection-mimicking signals(Vitti et al., 2013). In the past, many researchers have attempted to reduce the chances of this risk by comparing locus-specific data with genome-wide data because demographic events are thought to influence the genome as a whole. In contrast, selection usually works in a more targeted approach. However, in recent years, some have challenged this outlier approach, arguing that if a selection is ubiquitous, then distributed patterns of genetic hitchhiking would be misinterpreted as reflecting demographic events (Vitti et al., 2013).

Another recurring challenge that researchers have to deal with when detecting natural selection is systematic biases that might be present in the study's data. Most selection studies are known to use SNP data (Vitti et al., 2013). The data is collected by the usage of genotyping arrays specifically created to detect known polymorphisms. Based on the used SNP detection protocols, some of the low-frequency alleles might remain undetected, and they end up excluded from datasets. That means generated data do not represent the full degree of genetic diversity. This disadvantage is recognized as ascertainment bias. To reduce the weight of ascertainment bias, it is essential that the SNP detection protocol is acknowledged and that correct statistical measures would be taken (Balding et al., 2019).

For the results to be correct and significant, it is crucial that the specific time at which a selective sweep occurred is considered because the different characteristics analyzed in the tests show other detection rates for younger and older events of positive selection. For example, LD based EHH method can provide information on a natural selection that began <30 thousand years ago, while Tajima's D statistics can identify signals that are as old as 250,000 years (Vatsiou et al., 2016). Figure 1.4 shows the time-scale during which different classes of tests are best able to detect selection.

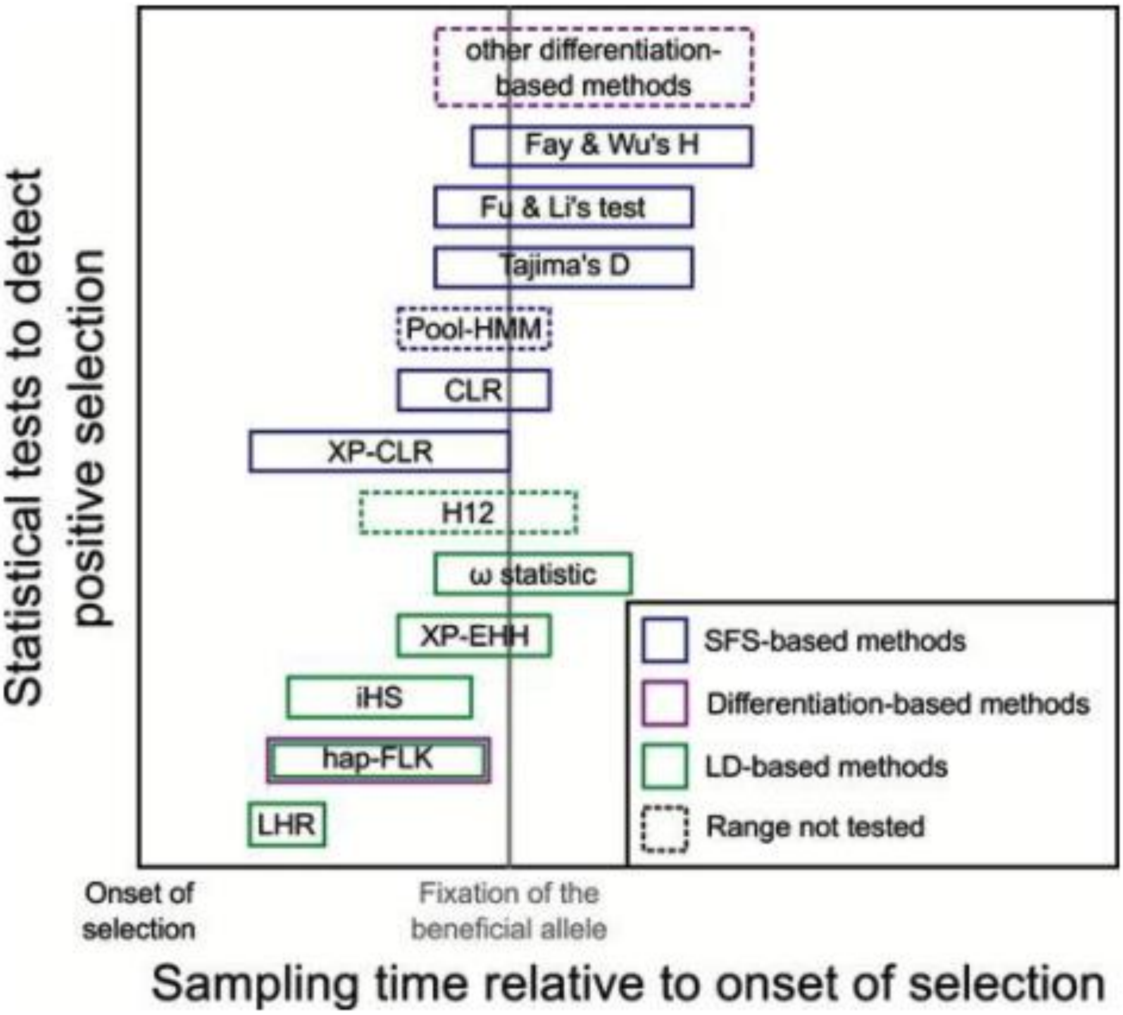


Figure 1.4: Relevant time frames for analyzing signatures of positive selection with the different methods. The frequency of the beneficial allele scales the time axis. The same range of beneficial allele frequencies with optimal performance depends on the simulation settings for the different methods. Thus the shown ranges are not exact thresholds but only guidelines. CLR, composite likelihood ratio test; FLK, extended Lewontin and Krakauer test; iHS, integrated haplotype score; LD, linkage disequilibrium; LRH, long-range haplotype test; SFS, site frequency spectrum; XP-CLR, cross-population composite likelihood ratio test; XP-EHH, cross-population extended haplotype homozygosity (Weigand and Leese, 2018).

1.7. Review of the relevant studies

Investigating the role of natural selection in influencing adaptations to the environment, diet, and diseases in humans is a prevalent issue. Populations must adjust to new environmental conditions with varying climatic conditions and food resources. This type of environmental and nutritional stress typically induces genetic adaptation (Vasseur and Quintana-Murci, 2013). An illustration of such a mechanism is the evolution of fatty acid desaturases (FADS) genes. Because of exposure to a protein-rich diet and the low temperature of the Greenland Arctic region, these genes are impacted and produce positive selection signals in Inuit populations (Fumagalli et al., 2015).

The characterization of these adaptive events represents a highly anticipated opportunity to explore the genetic basis of human adaptation and its crucial medical implications, turning out to be considerably necessary for evaluating the genetic causes behind human diseases (Piras et al., 2012).

However, most of the studies focusing on natural selection have targeted the adaptations that evolve throughout distances that extend hundreds to thousands of kilometers. Just a few of these studies have explored variation at microgeographic scales as small as tens of kilometers as it has been assumed that excessive rates of gene flow will limit adaptive divergence at fine spatial scales (Richardson et al., 2014). With an increasing amount of genetic data collected from hundreds and thousands of individuals, it can be stated that most if not all human populations display at least some degree of genetic population structure, even at small scales (Pankratov et al., 2020). Actually, many studies have accentuated that the genetic structure of the population should always be considered when investigating genetic associations and natural selection signals because if undetected, it may give false-positive results, even in datasets representing one nation or ethnic group (Kerminen et al., 2019).

However, more information about the effects of structure on population genetic analyses is needed. Further research may help answer many questions, such as whether local groups within a country may actually be different in their evolutionary histories, particularly in more recent times, and would help determine if analyzing such groups separately might offer more information into the population's past (Pankratov et al., 2020).

Several studies can already be found that demonstrate the importance of sampling strategies that consider the substructure of relatively homogeneous groups of individuals. The population substructures were detected in geographically and culturally homogeneous population isolates such as Iceland and Finland. It confirmed that even inhabitants of the isolated region should not be considered a single population (Pistis et al., 2009).

Studies of natural selection also benefit from the clustering of human populations into subgroups. For example, one study examined genetic structure within Cameroon that might determine differences among Cameroon ethnic groups in genome-wide malaria association studies in Saharan Africa. Three regions of Cameroon and three ethnic groups were included in the study population. Model-based clustering revealed differential ancestry proportions among ethnic groups. Strong selection signatures were found in the Human Leukocyte Antigen (*HLA*) - malaria-resistant gene region. It confirmed the long-standing knowledge that natural selection affected African populations in this genomic region due to immense disease pressure. The results also showed that regions affected by natural selection were, in fact, different between the different ethnic groups. These selection signatures implicated genes associated with disease response (Esoh et al., 2021).

Modes and targets of positive selection were also investigated in five groups that were living in Ethiopia. Group-specific and standard signals of selection were found. It was associated with vitamin B's metabolism from foods, ultraviolet response, and skin pigmentation standing out as a common pathway, probably in response to high ultraviolet irradiation. In genes such as *IFNA*, *MRC1*, immunoglobulins, and T-cell receptors, strong selection signals were identified, as they contribute to pathogen defence (Walsh et al., 2020).

The research for signatures of natural selection was also conducted in Estonia. They concluded that different regions of Estonia differ in both adequate population size and signatures of natural selection. Performed analysis showed that South-East Estonia strongly genetically differentiates from the other parts of the country. Researchers found that expression levels of the *EPM2A* gene, which is associated with Lafora disease, were elevated in this region, as well as levels of the *GRM5* gene, which has been shown previously to be a potential target of natural selection for the pigmentation phenotype (Pankratov et al., 2020).

A precise understanding of fine-scale genetic population structure is also essential in the medical field as better results in diagnostic and treatment of the disease might be more successful when individuals with a disease can be compared to healthy individuals with the same genetic background. Finnish population considered a homogeneous population, can be divided into small genetic populations that are geographically clustered. Finnish population was divided into 52 subgroups, and the groups followed the old dialect areas of Finland. Geographical isolations of Finland due to the country's very Nordic location and cultural isolation because of religious and linguistic boundaries led to the enrichment of disease-causing genes and the loss of others. (Kerminen et al., 2019).

To conclude, the observations presented in these studies suggest that population stratification should be taken into account while conducting genetic research. Even small populations that occupy the relatively small area with no solid geographic barriers might be genetically structured and exhibit interregional differences such as potential action of natural selection. Consequently, detailed knowledge of the regional genetic differentiation in such populations can be crucial to minimize the risk of false-positive results in genetic studies (Pistis et al., 2009).

2. METHODS

2.1. Study samples

The analysis was conducted with samples of 425 individuals from six ethnolinguistic groups collected in the Lithuanian population. The samples were grouped by dividing Lithuania into two main regions (Aukštaitija and Žemaitija) and then subdivide each region into smaller groups by dialect. Samples were collected from three groups of Aukštaitija (western (n = 79), southern (n = 67), and eastern (n = 79)) and three groups of Žemaitija (northern (n = 79), western (n = 43), and southern (n = 78). Participants of the study had to confirm affiliations with a particular ethnolinguistic region for at least two generations in the past. Genomic DNA for this study was extracted from venous blood. Genotyping was executed with the Illumina HumanOmniExpress-12 v1.1 and the Infinium OmniExpress-24 at the Department of Human and Medical Genetics, Institute of Biomedical Sciences, Faculty of Medicine, Vilnius University, Lithuania. The individuals all contributed DNA samples voluntarily and provided written informed consent. The ethical approval of this study was provided by the Vilnius Regional Research Ethics Committee No. 158200-05-329-79.

The data used in this research were collected in 2011–2013, during the LITGEN project (VP1-3.1-ŠMM-07-K-01-013).

All the genotyping data used in this study have been downloaded and are freely available at https://figshare.com/articles/Patterns_of_genetic_structure_and_adaptive_positive_selection_in_the_Lithuanian_population_from_high-density_SNP_data/7964159.

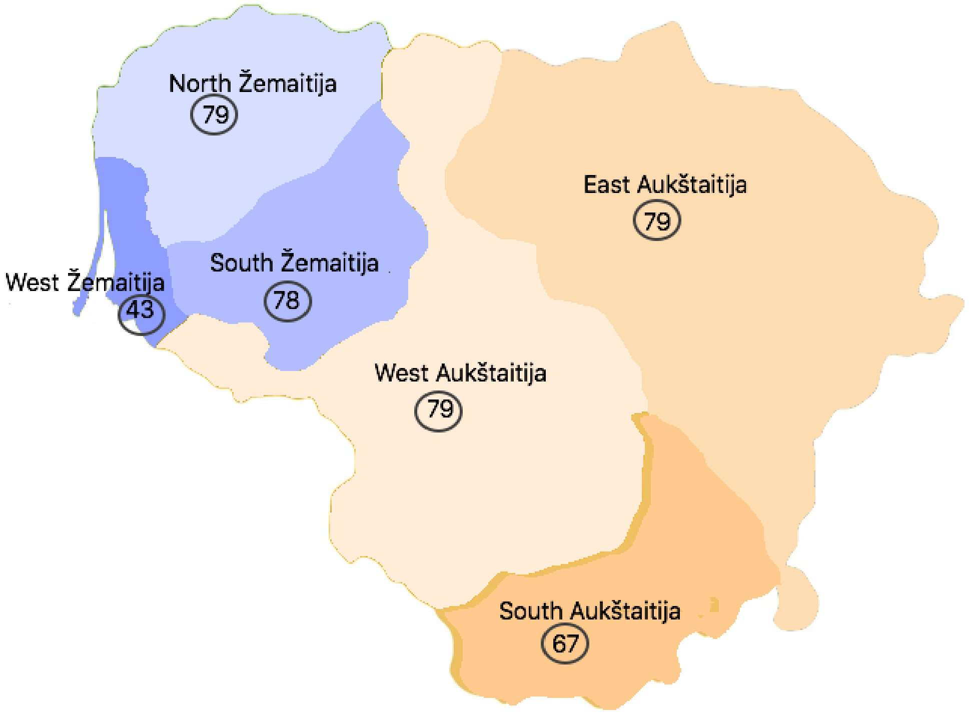


Figure 2.1. Map of the sampling sites. The number of samples per region is indicated in circles

2.2. Quality control of data

Quality control of the dataset was performed using *plink* v1.09 software (Chang et al., 2015) based on the standard manufacturer recommendations. The criteria for the exclusion process was: 1) individuals with a missing rate >10%; 2) SNPs with a missing rate >10%; 3) SNPs with minor allele frequency (MAF) <0.01; 4) SNPs deviating from Hardy–Weinberg equilibrium ($P < 0.0001$). After quality control, one person was removed, 424 individuals and 589 752 SNPs remained in the dataset of the study.

Multidimensional scaling analysis (MDS) was executed to test population stratification between different ethnolinguistic groups. Only SNPs with pairwise $r^2 < 0.5$ were used in this analysis. Pruning analysis was performed using *plink* (v1.09) command: *plink --file data --indep-pairwise* 50 5 0.5. After pruning, 300129 SNPs remained for subsequent analysis. MDS was also performed using *plink* (v1.09) *–mds-plot* and *–cluster* options, visualized by R programming language (R Core Team, 2018). Individuals outside of their expected group clusters were excluded from further analysis. Eight individuals were identified as outliers and removed, leaving the dataset with 416 samples.

2.3. Relatedness

The kinship and the inbreeding (F) coefficients were calculated using the SEEKIN (SEquence-based Estimation of KINship) (Dou et al., 2017) and *plink* (v1.09) (Chang et al., 2015) software, respectively, to ensure that individuals participating in the study are unrelated. Based on the results of this analysis, individuals with inbreeding coefficients higher than that expected for second cousin mating offspring (F values ≥ 0.0156)) were removed for all subsequent analyses. In pairs where values of kinship coefficient > 0.0084 (2nd-degree relative) were detected, one individual from each such pair was removed to minimize sample loss.

2.4. Detection of Natural Positive Selection

The original Lithuanian SNP genotyping data were merged with the 1000 Genomes Project (1KGP) phase 3 data set (Auton et al., 2015). Three populations were selected in particular: Africa, including the Yoruba in Ibadan, Nigeria (YRI), Finnish in Finland (FIN), Europe, including Utah residents with ancestry from northern and western Europe (CEU). SNP data sets were merged for analysis using only a common subset of SNPs, and quality control was repeated using the same criteria as in the analysis above. The generated dataset consisted of 249,794 distributed genome-wide in a total of 712 individuals. For this purpose, the *plink (*v1.90) command –merge was used. For analyses that require phased data, phasing was done using Beagle software with default parameters (Browning and Browning, 2007).

For finding signatures of recent positive selection, the two EHH-based metrics were calculated using the *REHH v3.0.1* package in R (Gautier et al., 2017). The *Rsb* and XP-EHH scores were computed. Both methods are based on the differentiation between the population of interest and reference populations. The analysis was carried out between all possible pairs of populations among Aukštaičiai (AK), Žemaičiai (ZM), CEU, FIN, and YRI. To identify differentiated regions of the genome, an exclusion method was chosen based on the empirical distribution of the data. P values were calculated according to the Rank scores method, proposed to provide a better comparison to the rest of the genome (Pybus et al., 2014). For identification of SNPs under selection in Aukštaičiai and Žemaičiai, while comparing it to the populations of 1KGP, only positive *Rsb* and XP-EHH values were considered. While comparing ethnolinguistic groups Aukštaičiai and Žemaičiai directly, negative values were also included. Genomic regions that carry an unusually high density of SNPs with highly positive values suggest that a selection event is likely to have occurred in one population but not the other.

In contrast, negative values suggest a selection event in the latter population, but not the former (Liu et al., 2013). Regions of interest that were detected by *Rsb* analysis were defined as those containing at least four neighbouring SNPs exceeding the threshold of *Rsb* ≥ 4, as suggested by authors of the software. Regions found by performing XP-EHH analysis were considered as candidate region if it contains at least four SNPs located at the 0.1% top extreme of the XP-EHH genome-wide empirical distribution. Genomic regions detected by both methods were selected as candidate regions and interrogated for genes.

Older natural selection signals were determined by Tajima's D neutrality statistical method. Tajima's D estimates were calculated using *VCF-kit* software, selecting a 100 kb sliding window across all autosomes with a 10 kb step size (Cook and Andersen, 2017). Outliers of the results were determined from the empirical distribution of Tajima's D p values calculated according to the Rank scores method. Only negative Tajima's D values were considered to represent signatures of positive selection (Vitti, Grossman, and Sabeti, 2013). Results with p values <0.01 are included in the further analysis. Genes were annotated using SNPnexus web-based software (Oscanoa et al., 2020).

3. RESULTS

3.1. Population Structure and relatedness within the population

The presence of related individuals in a dataset can influence genetic research results (Cadzow et al., 2014). First, an analysis of genetic relatedness between pairs of individuals in the Lithuanian population was performed by calculating kinship and inbreeding (F) coefficients.

A total of 4 individuals were identified with an F greater than that typical of second-generation cousin marriages (0.0156). After investigating kinship coefficients, 14 pairs of individuals appeared to be related (kinship coefficient > 0.0084). The estimates of F and kinship were sometimes negative but were increased to zero. Often such negative values can merely reflect random sampling error (Li et al., 2011). The average kinship coefficient among the studied persons of the Lithuanian population was set at 0.00074, the average F - 0.0022 (Table 2). The coefficients calculated within the ethnolinguistic groups showed that Žemaičiai are more related to each other than individuals in the Aukštaičiai group. A total of 4 individuals were identified with an F greater than that typical of second-generation cousin marriages (0.0156). After investigating kinship coefficients, 14 pairs of individuals appeared to be related (kinship coefficient > 0.0084). To minimize sample loss, only one individual from each pair was removed.

To characterize the ethnolinguistic groups in Lithuania, multidimensional scaling (MDS) analysis was applied (Figure 3.1). Eight outliers were identified in this step and removed from subsequent analysis. As expected, MDS analysis showed that ethnolinguistic groups in Lithuania form a single cluster. That shows that the Lithuanian population is homogeneous. The list of individuals identified as outliers and removed from the data set can be seen in Appendix 1.

Table 2. Average kinship and inbreeding coefficients determined within the populations of interest

Population	Number of individuals	Average inbreeding coefficient F	Average kinship coefficient	Individuals F > 0,0156
Aukštaičiai	223	0.0012	0.00096	1
Žemaičiai	200	0.0033	0.00087	3
Lietuviai	423	0.0022	0.00074	4

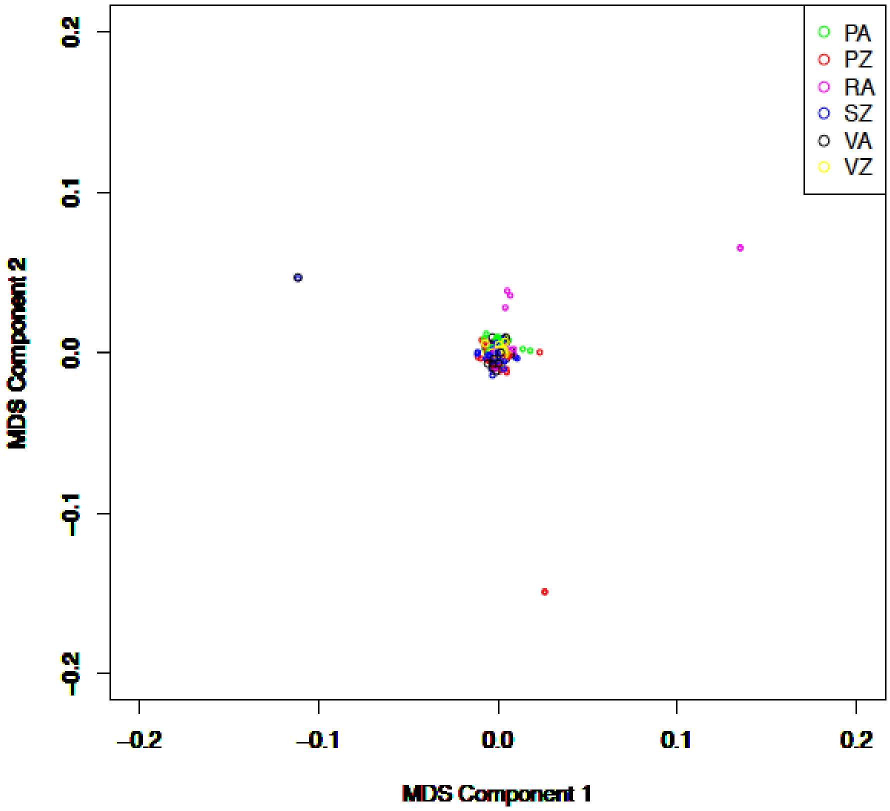


Figure 3.1. MDS analysis of the Lithuanian ethnolinguistic groups. Individuals from each group are depicted by the different colored dots: PA – south Aukštaičiai, PZ – south Žemaičiai, RA – east Aukštaičiai, SZ – north Žemaičiai; VA – west Aukštaičiai, VZ – west Žemaičiai

3.2. Candidate Regions of positive selection between the ethnolinguistic groups and 1KGP populations

In the studied populations of the Lithuanian ethnolinguistic groups, areas of the genome affected by positive natural selection were identified based on large-scale SNP genotyping data. *Rsb* and XP–EHH were calculated between all possible pairs of populations under study to detect putative recent selective events unique to each ethnolinguistic group in the context of 1KGP populations: CEU, FIN, YRI. The distribution of natural selection signals detected by pairwise genome-wide XP-EHH and *Rsb* analysis is shown in the *Manhattan* plots in Figures 3.2 and 3.3, respectively. In the analysis of the identified candidate regions for natural selection, more attention was given to regions containing genes that are involved in human adaptation, such as pathogen resistance, diet, climate, or pigmentation.

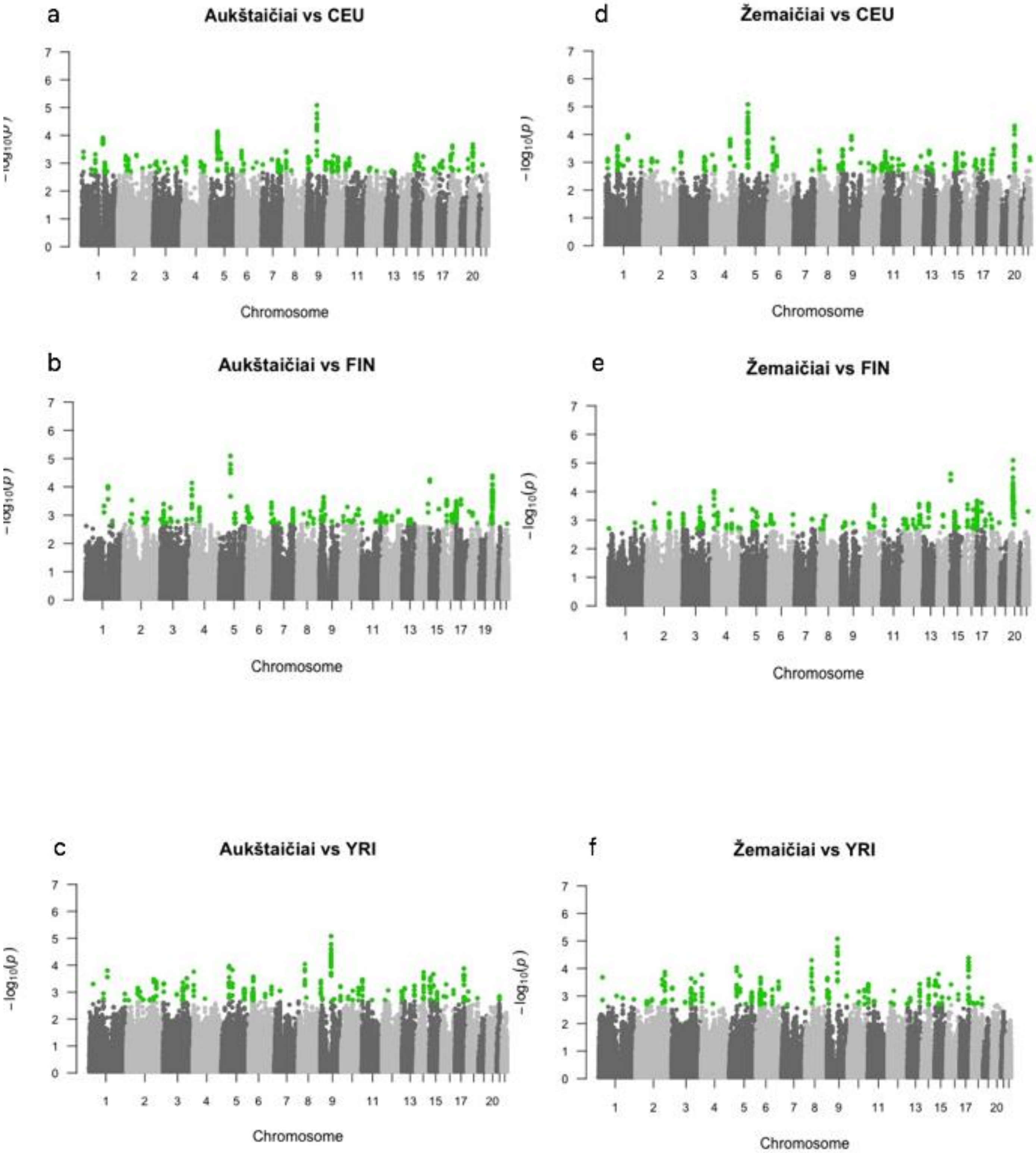


Figure 3.2. Manhattan plots of -log10 transformed XP-EHH p-values across the autosomes. (a) XP-EHH in AK-CEU, (b) XP-EHH in AK-FIN, (c) XP-EHH in AK-YRI, (d) XP-EHH in ZM-CEU, (e) XP-EHH in ZM-FIN, (f) XP-EHH in ZM-YRI. In each plot, green dots indicate 0.1% outlier regions

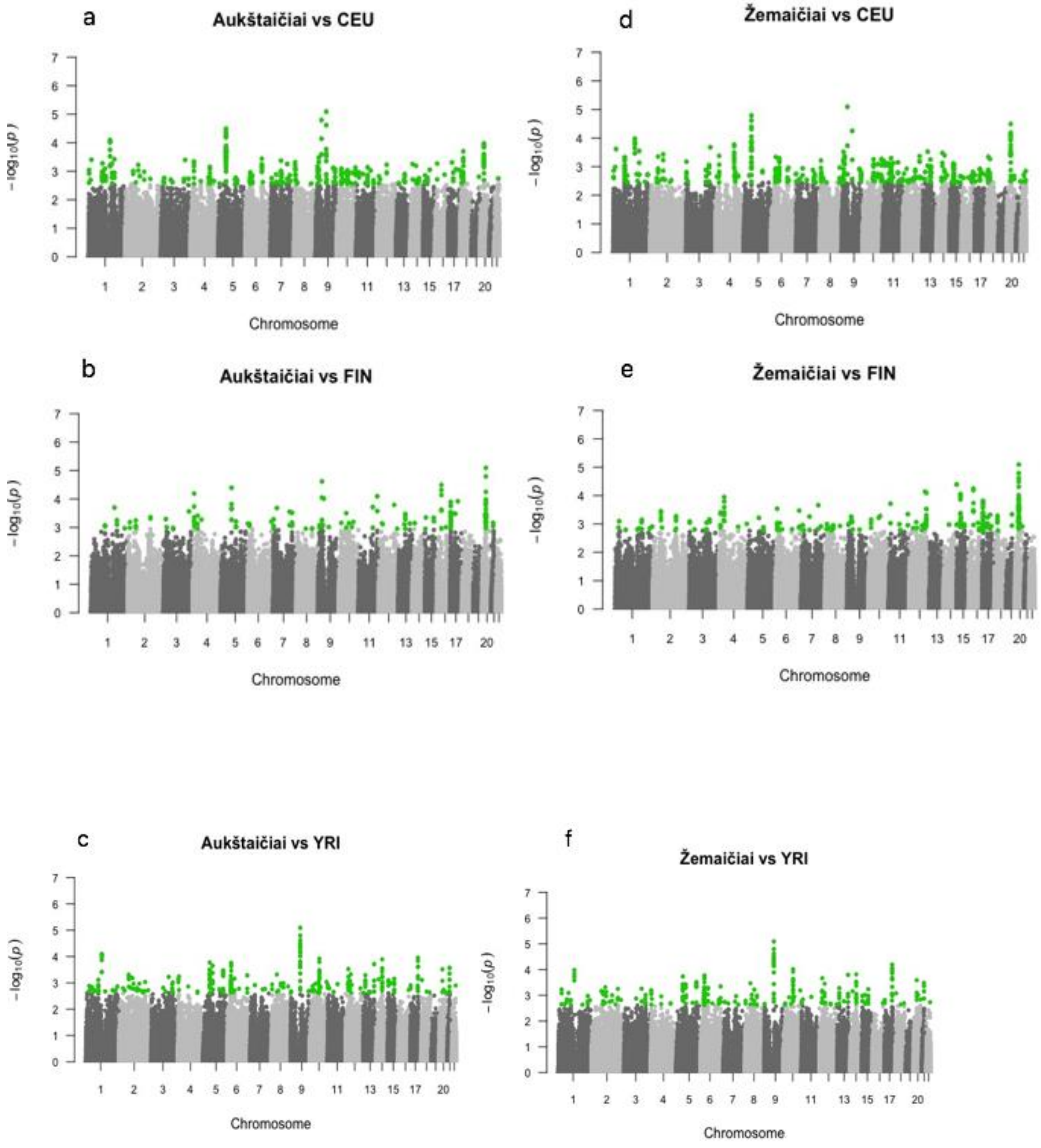
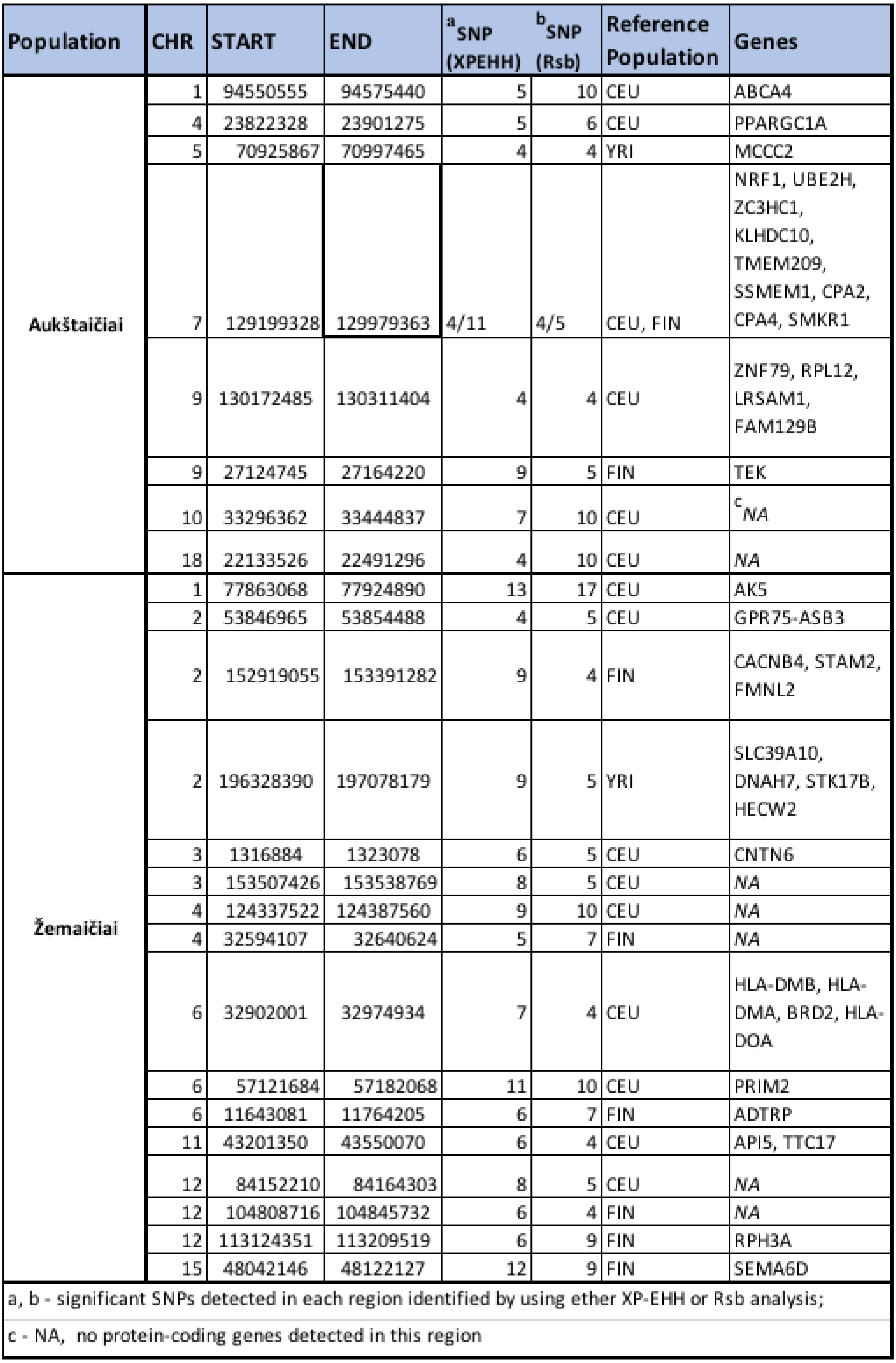


Figure 3.3. Manhattan plots of -log10 transformed *Rsb* p-values across the autosomes. (a) *Rsb* in AK-CEU, (b) *RSB* in AK-FIN, (c) *RSB* in AK-YRI, (d) *RSB* in ZM-CEU, (e) *RSB* in ZM-FIN, (f) *RSB* in ZM-YRI. In each plot, green dots indicate SNPs exceeding the threshold

XP-EHH analysis aided in the identification of 64 candidate regions possibly affected by recent positive natural selection in populations of interest. The exact number (64) of the regions was detected in both ethnolinguistic groups, Aukštaičiai and Žemaičiai. Using Rsb analysis, 54 candidate regions were detected in the Aukštaičiai population and 70 in the Žemaičiai population. The genomic regions used in the subsequent analysis were the ones detected by both EHH-based methods. In conclusion, 34 genomic regions putatively under selection were found in the ethnolinguistic group of Aukštaičiai, and 38 regions in the group of Žemaičiai (Appendix 2). Some of these overlap in both groups meaning that they are not population-specific. Unique signals detected in each ethnolinguistic group and significant SNPs detected using each method are summarized in Table 3.

The strongest signal in Aukštaičiai ethnolinguistic group was detected in chromosome 10 at a ~148 kb region. No protein-coding genes were identified in this specific region. However, this region reported being associated with body fat distribution, measured by waist-to-hip ratio (WHR) adjusted for body mass index (BMI), as well as elevated levels of high-density lipoprotein levels (Pulit et al., 2019; Richardson et al., 2020). Another top significant signal was spotted in chromosome 1, region 94550555 – 94575440, size ~24 kb. Here, SNP variants rs4147825 (NC_000001.10:g.94560938G>A) and rs4147823 (NC_000001.10:g.94561272A>C) was identified in gene *ABCA4*. This gene provides instructions for producing a protein found in the retina, the specialized light-sensitive tissue that lines the back of the eye (Tracewska et al., 2019). The region in the 7th chromosome detected in this study harbours many genes. The one gene *SMKR1* is associated with a prudent diet consisting of fruits, vegetables, whole grains, legumes, nuts, fish, and low-fat dairy products (Guénard et al., 2017). Another gene in the same region that was found analyzing the Aukštaičiai population was the gene *NRF1* (variant rs754386, NC_000007.13:g.129214174G>A), and it can be linked to BMI. Also, region 27124745 - 27164220 in chromosome 9 harbours gene *TEK*, where Upstream transcript variants were identified rs2232419 (NC_000006.11:g.28367768G>A), provides instructions for making a protein called *TEK* receptor tyrosine kinase (Limaye et al., 2009).

Table 3. Non-overlapping genomic regions under positive selection identified with two EHH-based approaches in two ethnolinguistic groups of the Lithuanian population



Population	CHR	START	END	[a]SNP (XPEHH)	[b]SNP (Rsb)	Reference Population	Genes
Aukštaičiai	1	94550555	94575440	5	10	CEU	ABCA4
	4	23822328	23901275	5	6	CEU	PPARGC1A
	5	70925867	70997465	4	4	YRI	MCCC2
	7	129199328	129979363	4/11	4/5	CEU, FIN	NRF1, UBE2H, ZC3HC1, KLHDC10, TMEM209, SSMEM1, CPA2, CPA4, SMKR1
	9	130172485	130311404	4	4	CEU	ZNF79, RPL12, LRSAM1, FAM129B
	9	27124745	27164220	9	5	FIN	TEK
	10	33296362	33444837	7	10	CEU	[c]*NA*
	18	22133526	22491296	4	10	CEU	*NA*
Žemaičiai	1	77863068	77924890	13	17	CEU	AK5
	2	53846965	53854488	4	5	CEU	GPR75-ASB3
	2	152919055	153391282	9	4	FIN	CACNB4, STAM2, FMNL2
	2	196328390	197078179	9	5	YRI	SLC39A10, DNAH7, STK17B, HECW2
	3	1316884	1323078	6	5	CEU	CNTN6
	3	153507426	153538769	8	5	CEU	*NA*
	4	124337522	124387560	9	10	CEU	*NA*
	4	32594107	32640624	5	7	FIN	*NA*
	6	32902001	32974934	7	4	CEU	HLA-DMB, HLA-DMA, BRD2, HLA-DOA
	6	57121684	57182068	11	10	CEU	PRIM2
	6	11643081	11764205	6	7	FIN	ADTRP
	11	43201350	43550070	6	4	CEU	API5, TTC17
	12	84152210	84164303	8	5	CEU	*NA*
	12	104808716	104845732	6	4	FIN	*NA*
	12	113124351	113209519	6	9	FIN	RPH3A
	15	48042146	48122127	12	9	FIN	SEMA6D

a, b - significant SNPs detected in each region identified by using ether XP-EHH or Rsb analysis;

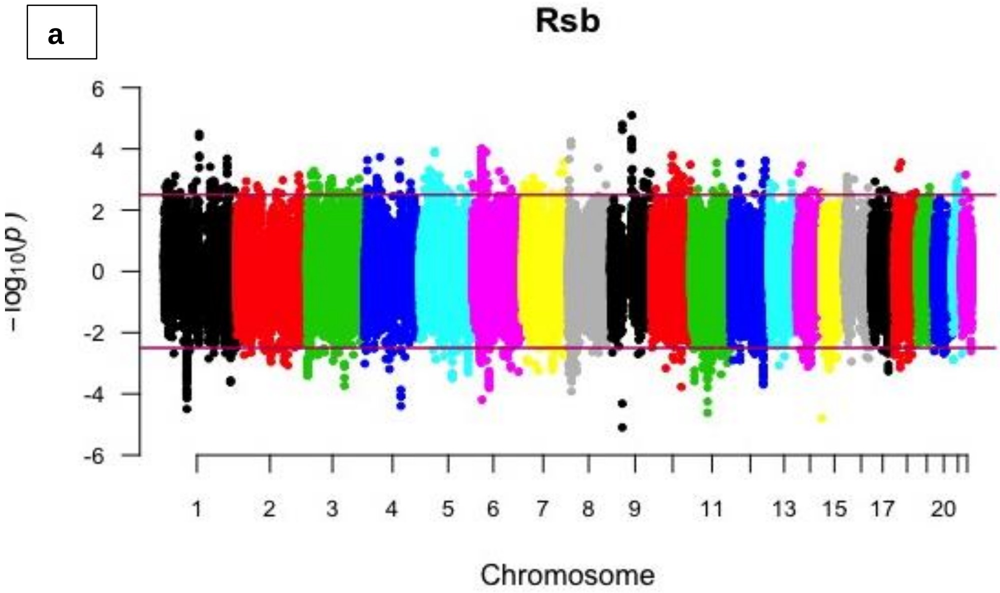
c - NA, no protein-coding genes detected in this region

In the ethnolinguistic group of Žemaičiai, the strongest signal was detected in chromosome 1, region 77863068 - 77924890 (size ~61 kb). This site harbors the gene *AK5*, and four transcript variants were identified here (rs12034899 (NC_000001.10:g.77903611A>G), rs6658302 (NC_000001.10:g.77908985A>G), rs10873941 (NC_000001.10:g.77902981G>A), rs12407481 (NC_000001.10:g.77904451G>A)). This gene encodes a member of the adenylate kinase family, which regulates the adenine nucleotide composition within a cell by catalysing the reversible transfer of phosphate groups among adenine nucleotides. The *AK5* can be associated with BMI (Pulit et al., 2019). The second strongest signal was observed in chromosome 15. Here the gene *SEMA6D* can be found (non-synonymous variant rs532598 (NG_029119.2:g.586669G>A)). The mutations in this specific gene are associated with traits such as pork consumption, fish and plant-related diet, oily fish consumption (Niarchou et al., 2020). Many genes in the group of Žemaičiai were linked to the BMI and body composition (*DNAH7, HLA-DMA, STAM2, STK17B*), however non-synonymous variants were not detected.

3.3. Adaptation within the Lithuanian population

Using the same two XP-EHH and *Rsb* methods of analysis, candidate regions putatively affected by natural selection were identified, comparing two ethnolinguistic groups of Lithuania, Aukštaičiai, and Žemaičiai. Positive scores suggest that selection is likely to have happened in population Aukštaičiai. In contrast, negative scores suggest the same about population Žemaičiai (Figure 3.4) (Sabeti et al., 2002). Outliers were detected using the same thresholds as in analysis using 1KGP populations.

Seventeen candidate regions possibly under the positive selection were found - 7 in the Aukštaičiai group and 10 in Žemaičiai. All genomic regions detected in the Aukštaičiai population were unique in comparison with the results collected in the previous step, while 5 regions in the Žemaičiai population were already detected (Table 4).



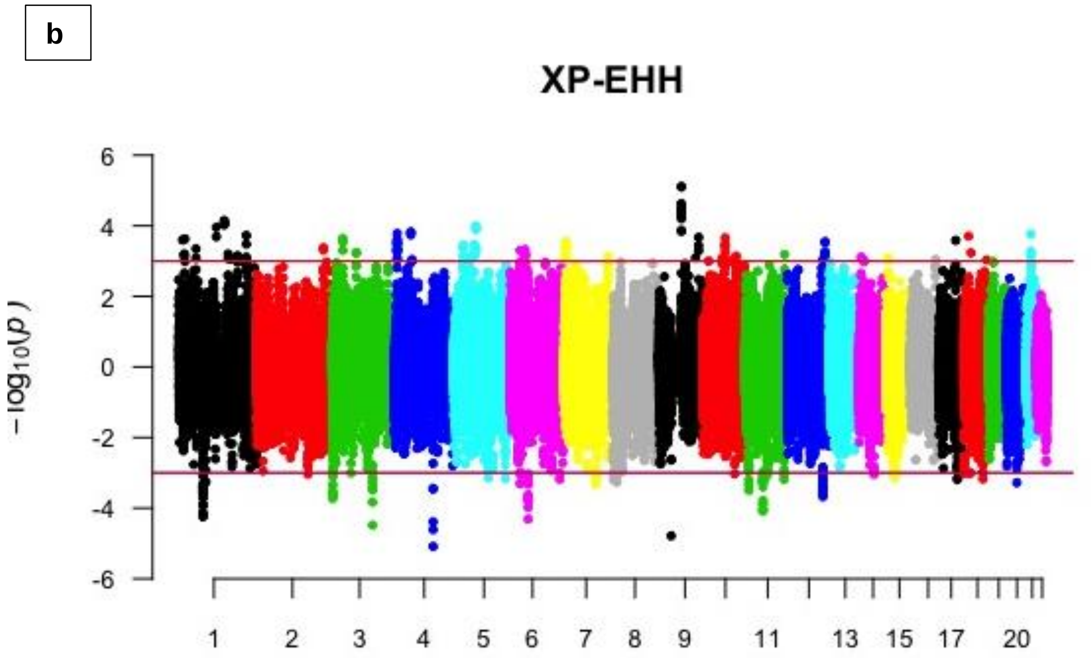
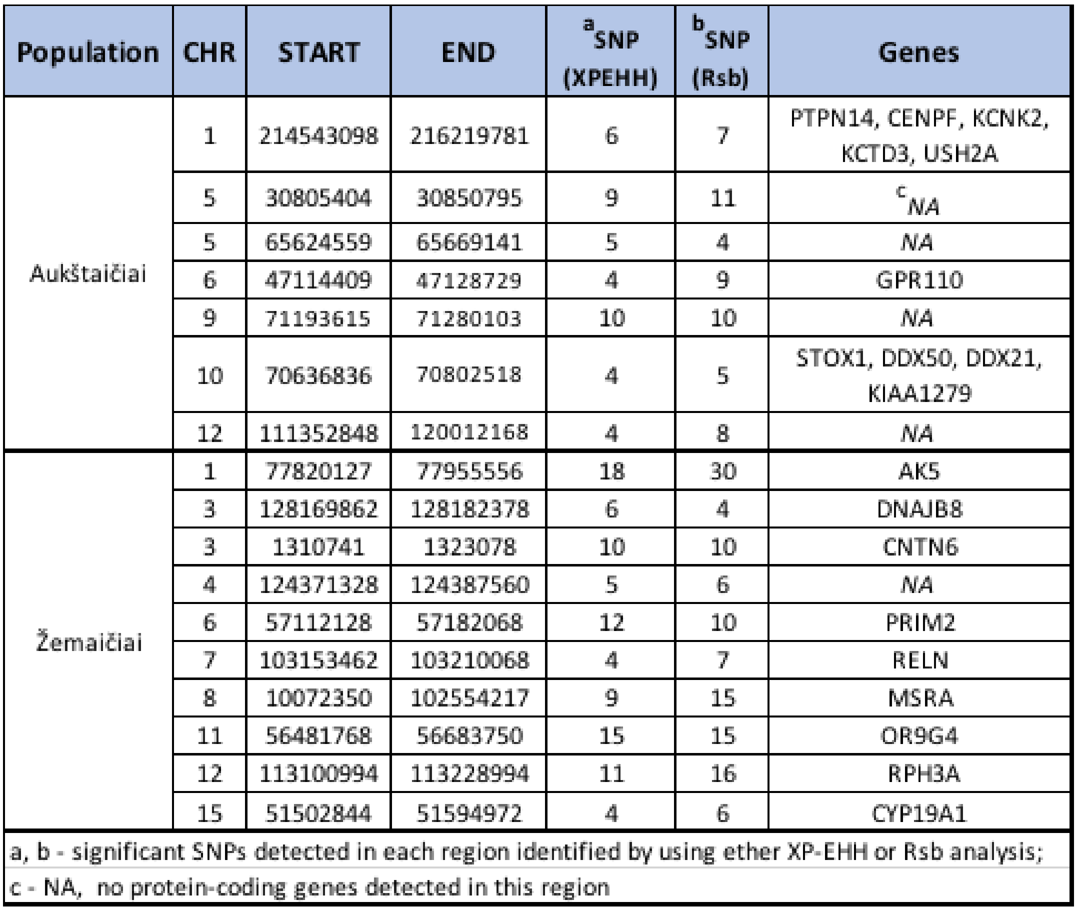


Figure 3.4. Manhattan plots of selection signatures determined by comparing AK and ZM groups using (a) Rsb and (b) XP-EHH approaches

The strongest signal in the Aukštaičiai population, while comparing it to Žemaičiai, was identified in chromosome 9 (size ~86 kb). This region does not harbour any protein-coding genes, but three intron variants were detected here (rs4271029, rs10869119, rs7851511). They do not have any reported clinical significance. The same situation is in another significant region, 30805404 - 30850795 in chromosome 5. In the chromosome 1, variant rs12029094 (NC_000001.10:g.215889410G>A) were found in gene *USH2A*, which is known to be associated with vision problems (Huang et al., 2018).

In the Žemaičiai ethnolinguistic group, the strongest signal was observed in chromosome 1, region 77820127 – 77955556. Here the gene *AK5* can be found. The same putatively under natural selection region was recognized while comparing Žemaičiai to 1KGP populations. A significant signal was also detected in chromosome 11, where olfactory receptor family member gene *OR9G4* (with coding sequence variants rs513873 (NM_001005284.1:c.128T>C) and rs11228763 (NM_001005284.1:c.128T>C)) is located. The chromosome 12 harbours the gene *RPH3A*, which can be linked to the protein and fat intake (also detected in the previous analysis) (Liu et al., 2020). For the gene *DNAJB8* in chromosome 3, two non-synonymous variants rs2981026 and rs13071640 were found. Connection of this gene and BMI were reported (Pulit et al., 2019). Chromosome 8 in the region 10072350 – 102554217 have the gene *MSRA*. Non-synonymous variants in this gene are reportedly linked to the fish-based diet (Niarchou et al., 2020).

Table 4. Genomic regions under positive selection identified with two EHH-based approaches within the Lithuanian population



Population	CHR	START	END	[a]SNP (XPEHH)	[b]SNP (Rsb)	Genes
Aukštaičiai	1	214543098	216219781	6	7	PTPN14, CENPF, KCNK2, KCTD3, USH2A
	5	30805404	30850795	9	11	[c]*NA*
	5	65624559	65669141	5	4	*NA*
	6	47114409	47128729	4	9	GPR110
	9	71193615	71280103	10	10	*NA*
	10	70636836	70802518	4	5	STOX1, DDX50, DDX21, KIAA1279
	12	111352848	120012168	4	8	*NA*
Žemaičiai	1	77820127	77955556	18	30	AK5
	3	128169862	128182378	6	4	DNAJB8
	3	1310741	1323078	10	10	CNTN6
	4	124371328	124387560	5	6	*NA*
	6	57112128	57182068	12	10	PRIM2
	7	103153462	103210068	4	7	RELN
	8	10072350	102554217	9	15	MSRA
	11	56481768	56683750	15	15	OR9G4
	12	113100994	113228994	11	16	RPH3A
	15	51502844	51594972	4	6	CYP19A1

a, b - significant SNPs detected in each region identified by using ether XP-EHH or Rsb analysis;

c - NA, no protein-coding genes detected in this region

3.4. Genetic hitchhiking by Tajima's D statistics

Tajima's D statistics analysis was performed to find older signals of natural selection. This method is appropriate for detecting evidence of positive selection in human populations occurring within the past 250,000 years or roughly 10,000 generations and runs by identifying a surplus of low-to-intermediate frequency variants (Cadzow et al., 2014).

Taking into account the empirical distribution of p values less than 0.01, 25 regions of the genome with extremely negative Tajima's D estimates were identified as potential sites affected by natural selection areas for the Lithuanian population (see Table 5). 14 regions were identified in the Aukštaičiai population and 19 in the Žemaičiai group. Tajima's D analysis was also performed for the CEU and FIN populations in order to compare and identify specific areas for the ethnolinguistic groups (see Figure 3.5). While comparing the Aukštaičiai and Žemaičiai populations, seven genomic regions were recognized as overlapping, and many detected sites were also found in 1KGP populations. A total of 7 specific genome regions were identified, 5 in the Aukštaičiai and 2 in the Žemaičiai population.

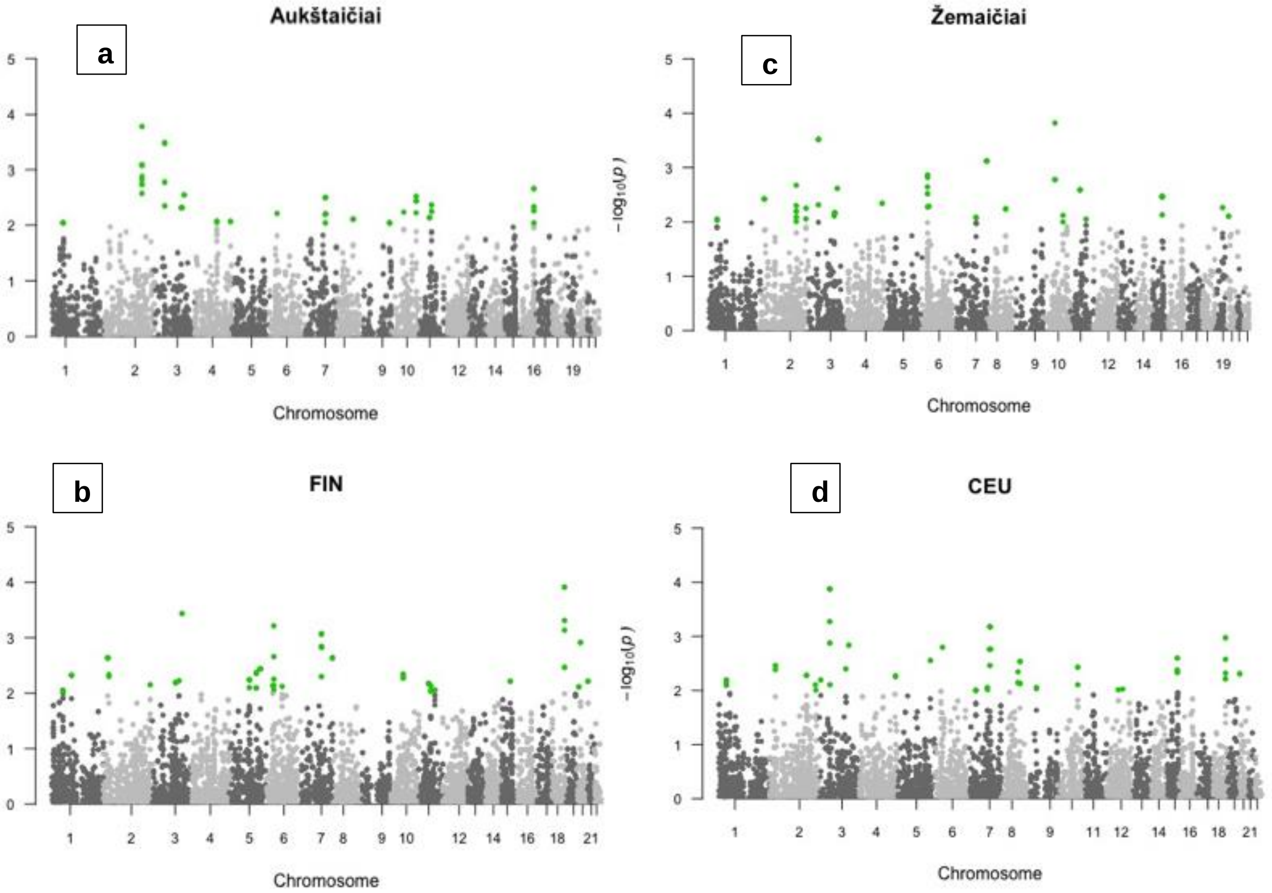
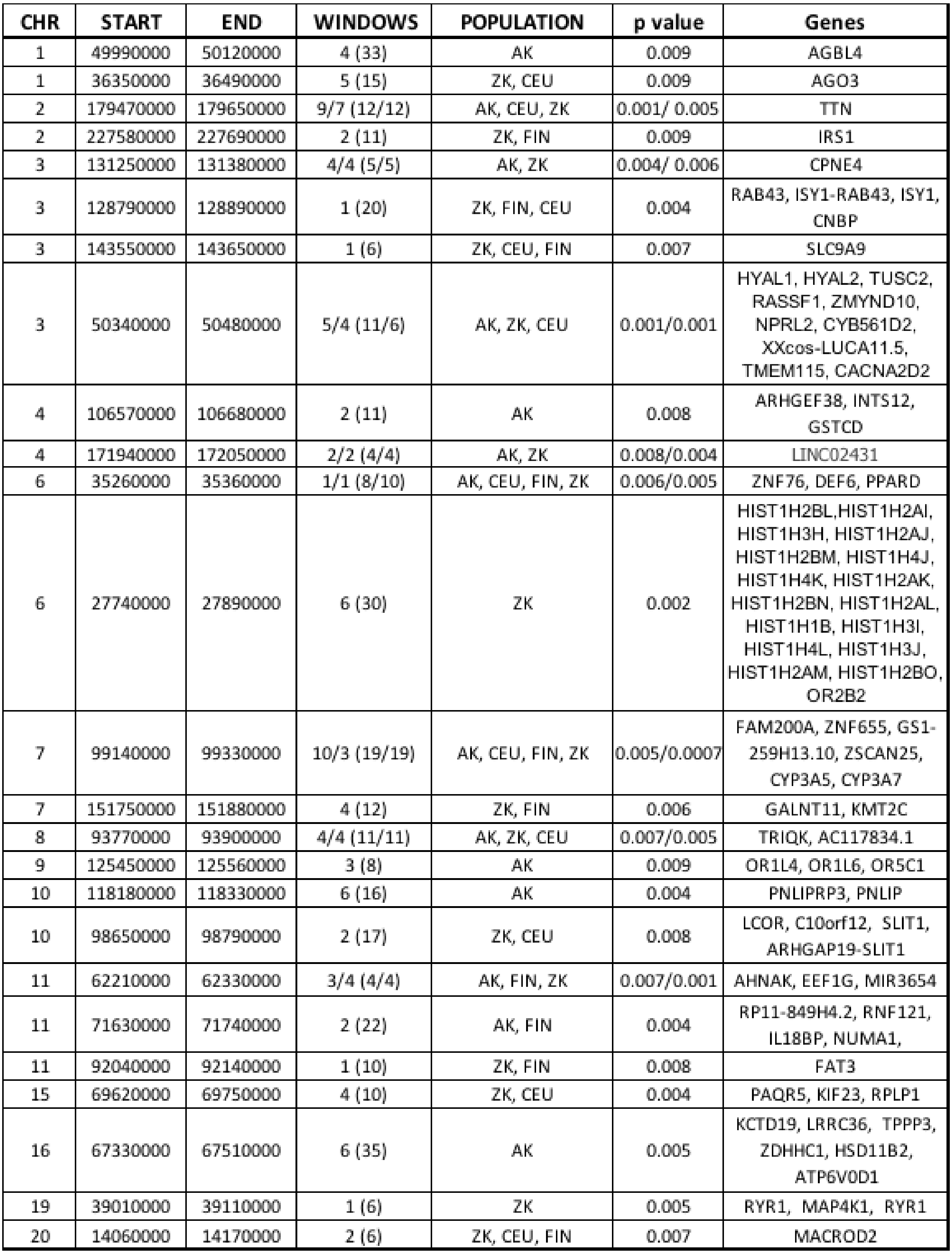


Figure 3.5. Manhattan plots of the log-transformed Tajima's D p-values (a) In AK, (b) In FIN, (c) In ZM, (d) In CEU. In each plot, green dots indicate 0.1% outlier regions

After the genomic annotation, more than one gene was identified in many genetic domains. The strongest signals specific to the Aukštaičiai population were found on the 16th chromosome (67330000-67510000), which had 35 windows in the region, the value of p was 0.0005, and it is harbouring genes such as *LRRC36, ZDHHC1, HSD11B2*. Based on the previously conducted researches, those genes can be associated with BMI and WHR (Ng et al., 2017). *KCTD19* is another gene in the same region, and some non-synonymous variants in this gene are known to cause hypothyroidism (Kichaev et al., 2019). Another strong signal with 30 windows was found in chromosome 10 with *PNLIPRP3, PNLIP* genes, also linked to BMI. A strong signal was also identified on chromosome 1 (49990000- 50120000) on the *AGBL4* gene.

The two unique regions in the Žemaičiai population were at chromosomes 6 (27740000 – 27890000) and 19 (39010000 - 39110000). The site in the 6th chromosome, with 30 windows, includes many genes, and based on gene enrichment, they might be linked to skin and connective tissue diseases (Johannesdottir et al., 2012).

Table 5. Genomic regions exhibiting positive selection and identified by Tajima's D method. The windows indicate the number of Tajima's D p values in the top 0.1 %. The number of windows in that genetic domain is indicated in parentheses



CHR	START	END	WINDOWS	POPULATION	p value	Genes
1	49990000	50120000	4 (33)	AK	0.009	AGBL4
1	36350000	36490000	5 (15)	ZK, CEU	0.009	AGO3
2	179470000	179650000	9/7 (12/12)	AK, CEU, ZK	0.001/ 0.005	TTN
2	227580000	227690000	2 (11)	ZK, FIN	0.009	IRS1
3	131250000	131380000	4/4 (5/5)	AK, ZK	0.004/ 0.006	CPNE4
3	128790000	128890000	1 (20)	ZK, FIN, CEU	0.004	RAB43, ISY1-RAB43, ISY1, CNBP
3	143550000	143650000	1 (6)	ZK, CEU, FIN	0.007	SLC9A9
3	50340000	50480000	5/4 (11/6)	AK, ZK, CEU	0.001/0.001	HYAL1, HYAL2, TUSC2, RASSF1, ZMYND10, NPRL2, CYB561D2, XXcos-LUCA11.5, TMEM115, CACNA2D2
4	106570000	106680000	2 (11)	AK	0.008	ARHGEF38, INTS12, GSTCD
4	171940000	172050000	2/2 (4/4)	AK, ZK	0.008/0.004	LINC02431
6	35260000	35360000	1/1 (8/10)	AK, CEU, FIN, ZK	0.006/0.005	ZNF76, DEF6, PPARD
6	27740000	27890000	6 (30)	ZK	0.002	HIST1H2BL,HIST1H2AI, HIST1H3H, HIST1H2AJ, HIST1H2BM, HIST1H4J, HIST1H4K, HIST1H2AK, HIST1H2BN, HIST1H2AL, HIST1H1B, HIST1H3I, HIST1H4L, HIST1H3J, HIST1H2AM, HIST1H2BO, OR2B2
7	99140000	99330000	10/3 (19/19)	AK, CEU, FIN, ZK	0.005/0.0007	FAM200A, ZNF655, GS1-259H13.10, ZSCAN25, CYP3A5, CYP3A7
7	151750000	151880000	4 (12)	ZK, FIN	0.006	GALNT11, KMT2C
8	93770000	93900000	4/4 (11/11)	AK, ZK, CEU	0.007/0.005	TRIQK, AC117834.1
9	125450000	125560000	3 (8)	AK	0.009	OR1L4, OR1L6, OR5C1
10	118180000	118330000	6 (16)	AK	0.004	PNLIPRP3, PNLIP
10	98650000	98790000	2 (17)	ZK, CEU	0.008	LCOR, C10orf12, SLIT1, ARHGAP19-SLIT1
11	62210000	62330000	3/4 (4/4)	AK, FIN, ZK	0.007/0.001	AHNAK, EEF1G, MIR3654
11	71630000	71740000	2 (22)	AK, FIN	0.004	RP11-849H4.2, RNF121, IL18BP, NUMA1,
11	92040000	92140000	1 (10)	ZK, FIN	0.008	FAT3
15	69620000	69750000	4 (10)	ZK, CEU	0.004	PAQR5, KIF23, RPLP1
16	67330000	67510000	6 (35)	AK	0.005	KCTD19, LRRC36, TPPP3, ZDHHC1, HSD11B2, ATP6V0D1
19	39010000	39110000	1 (6)	ZK	0.005	RYR1, MAP4K1, RYR1
20	14060000	14170000	2 (6)	ZK, CEU, FIN	0.007	MACROD2

DISCUSSION

The aim of this study was to analyze signatures of positive natural selection of two main ethnolinguistic groups, Žemaičiai and Aukštaičiai, from genome-wide SNP genotyping data.

A similar analysis was already performed by A. Urnikytė in 2019. However, in that study, the Lithuanian population was considered homogeneous, and the ethnolinguistic division of Lithuania was not considered. In this study, the MDS analysis of six Lithuanian ethnolinguistic groups according to the wide-scale SNP markers confirmed the homogeneity of the studied Lithuanian population.

A. Urnikytė, in her dissertation, by increasing the sample size to 399 individuals and covering the whole genome, were able to detect weak signals of genetic structure in the Lithuanian population, providing new insights to any genetic analysis in the future because population structure analysis is crucial to the design, for proper analysis, and interpretation of genetic association and natural selection studies (Esoh et al., 2021).

In the analysis of positive natural selection, using large-scale SNP genotyping data and three statistical methods, specific areas of the genome affected by positive natural selection candidates in the two different ethnolinguistic groups were identified. The work identified candidate genes involved in human adaptation: diet, body mass index, and other traits.

In the population of the Aukštaičiai, gene *SMKR1* were detected as putatively under natural selection, using EHH-based methods. This particular gene might be associated with the prudent dietary pattern, meaning that more fruits, vegetables, whole grains, legumes, nuts, fish, and low-fat dairy products are consumed compared with other foods (Guénard et al., 2017). On the other hand, in the ethnolinguistic group of Žemaičiai, genes located in the candidate regions that seem to be linked to the diet were also found (*SEMA6D* and *MSRA*). However, these genes are associated with more fish, meat, particularly pork, related diets (Niarchou et al., 2020). This scenario is highly possible because, based on the researches of Lithuanian ethnologists, it is highly inevitable that the basis of people's diet consisted of food products and dishes determined by natural and economic conditions. People ate what they produced on their farms or got from nature. Moreover, as Žemaitija is next to the sea is highly possible that more fish were consumed in this region (Senvaitytė, 2014).

A strong significant signal was observed in chromosome 9 in the gene *TEK* with three SNP variants while comparing Aukštaičiai with 1KGP populations. This gene encodes a receptor that belongs to the protein tyrosine kinase Tie2 family. Mutations in this gene are associated with inherited venous malformations of the skin and mucous membranes (Limaye et al., 2009). The different variant SNP actually can be associated with trans-fatty acid levels. This is correlated with the diet patterns in a population (Mozaffarian et al., 2015).

In the ethnolinguistic group of Žemaičiai, the top significant signal was detected in chromosome 1, in the region harbouring gene *AK5*. This gene is highly associated with obesity and BMI (Pulit et al., 2019). Actually, in both ethnolinguistic groups, many genes associated with weight were found (*NRF1, DNAH7, HLA-DMA, STAM2, STK17B, DNAJB8*), and even some genes were detected using Tajima's D (*LRRC36, ZDHHC1, HSD11B2, PNLIPRP3, PNLIP*). The hypothesis suggests that frequency in those genes actually might rise because our ancestors have undergone a positive selection for genes that favored energy storage as a consequence of the cyclical episodes of famine and surplus (Sellayah et al., 2014).

SNP variants in the AK5 gene are also possibly connected to the circadian rhythms in the population, particularly morningness. Two more regions in chromosome 2 were also harbouring genes related to this trait (*GPR75-ASB3, HECW2*) (Hu et al., 2016). What caused the variants in those genes to rise in frequency cannot be determined. One of the environmental conditions affecting circadian rhythms is temperature. Higher temperatures are usually associated with morningness, and based on the meteorological conditions of Lithuania, the average temperature in Žemaitija is noticeably higher than in Aukštaitija (Bukantis and Kažys, 2020).

On chromosome 11, the olfactory receptor gene *OR9G4* was found, with two SNP variants, while comparing groups of Aukštaičiai and Žemaičiai directly. The effect of natural selection has only been identified for specific olfactory receptor genes. Why natural selection affects some olfactory receptors is not fully elucidated but is related to nutrition and health (Gilad et al., 2003).

In the 12th chromosome region 113124351- 113209519, the gene *RPH3A* was identified in the Žemaičiai group, and it is connected to that can be linked to protein and fat intake (Liu et al., 2020). The region of Žemaitija, located along the Baltic sea, probably was rich in the presence of marine food resources (Kučinskas, 2017). As fish is high in fat and protein, the diet based on it might be the reason why this region is putatively under natural selection.

The signals of natural selection were already identified in the Lithuanian population, where all ethnolinguistic groups were treated as a single group (Urnikytė et al., 2019), a comparison of the results of that study and the analysis conducted in this thesis were available. The genes detected by both analysis were *OR2B6* (olfactory receptor gene), *MINK1* and *ENO3* (associated with diseases). *SLC24A5* gene is known to be linked with the pigmentation of the skin and the *TYRP1* gene associated with hair and iris color in European populations (Hider et al., 2013). Even more, overlapping gene regions were found when comparing Tajima's D statistic results for evidence of older positive selection between populations. The fact that these two pieces of research complement each other increases the reliability of the results. Nevertheless, the fact that new genes under selection were found when analyzing ethnolinguistic groups instead of the lithuanians as a single population confirms that population structure should always be considered when working with genetic data.

This study only focused only on the positive selection, but other forms of adaptation may provide additional insights into the ethnolinguistic groups in Lithuania. The main disadvantage of this study is the inability to identify rare variants based on large-scale SNP genotyping data. The use of next-generation sequencing data would address this shortcoming. Functional validation of the identified variants is also required.

CONCLUSIONS

1. After analyzing the genetic structure of the Lithuanian population using high-density SNP genotyping data and performing MDS analysis to investigate their genetic similarity, the evident homogeneous genetic landscape across the six ethnolinguistic groups was detected.

2. The calculations of kinship and inbreeding coefficients within the ethnolinguistic groups detected that individuals in the group of Žemaičiai are more closely related than Aukštaičiai.

3. Unique signatures of positive selection in the Lithuanian ethnolinguistic groups were investigated over different time frames using three statistics: XP-EHH, Rsb and Tajima's D.

4. Candidate regions for positive selection in the ethnolinguistic group of Aukštaičiai were identified that are related to BMI (*NRF1*, *LRRC36, ZDHHC1, HSD11B2*, *PNLIPRP3, PNLIP*), diet (*SMKR1*) and other traits (*TEK*, *USH2A*, *KCTD19*).

5. In the ethnolinguistic group of Žemaičiai candidate regions harboured genes associated with fish-rich diet (*SEMA6D* and *MSRA*), circadian rhythms (*AK5*, *GPR75-ASB3, HECW2*) and genes linked to BMI as well as in Aukštaičiai population (*DNAH7, HLA-DMA, STAM2, STK17B*, *DNAJB8*).

REFERENCES

1. Ashraf, Q., Galor, O., 2013. The “Out of Africa” Hypothesis, Human Genetic Diversity, and Comparative Economic Development. Am Econ Rev 103, 1–46. https://doi.org/10.1257/aer.103.1.1
2. Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., Gabriel, S.B., Gibbs, R.A., Green, E.D., Hurles, M.E., Knoppers, B.M., Korbel, J.O., Lander, E.S., Lee, C., Lehrach, H., Mardis, E.R., Marth, G.T., McVean, G.A., Nickerson, D.A., Schmidt, J.P., Sherry, S.T., Wang, J., Wilson, R.K., Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., Lan, T., Li, G., Li, J., Li, Y., Liu, S., Liu, X., Lu, Y., Ma, X., Tang, M., Wang, B., Wang, G., Wu, H., Wu, R., Xu, X., Yin, Y., Zhang, D., Zhang, W., Zhao, J., Zhao, M., Zheng, X., Lander, E.S., Altshuler, D.M., Gabriel, S.B., Gupta, N., Gharani, N., Toji, L.H., Gerry, N.P., Resch, A.M., Flicek, P., Barker, J., Clarke, L., Gil, L., Hunt, S.E., Kelman, G., Kulesha, E., Leinonen, R., McLaren, W.M., Radhakrishnan, R., Roa, A., Smirnov, D., Smith, R.E., Streeter, I., Thormann, A., Toneva, I., Vaughan, B., Zheng-Bradley, X., Bentley, D.R., Grocock, R., Humphray, S., James, T., Kingsbury, Z., Lehrach, H., Sudbrak, R., Albrecht, M.W., Amstislavskiy, V.S., Borodina, T.A., Lienhard, M., Mertes, F., Sultan, M., Timmermann, B., Yaspo, M.-L., Mardis, E.R., Wilson, R.K., Fulton, L., Fulton, R., Sherry, S.T., Ananiev, V., Belaia, Z., Beloslyudtsev, D., Bouk, N., Chen, C., Church, D., Cohen, R., Cook, C., Garner, J., Hefferon, T., Kimelman, M., Liu, C., Lopez, J., Meric, P., O’Sullivan, C., Ostapchuk, Y., Phan, L., Ponomarov, S., Schneider, V., Shekhtman, E., Sirotkin, K., Slotta, D., Zhang, H., McVean, G.A., Durbin, R.M., Balasubramaniam, S., Burton, J., Danecek, P., Keane, T.M., Kolb-Kokocinski, A., McCarthy, S., Stalker, J., Quail, M., Schmidt, J.P., Davies, C.J., Gollub, J., Webster, T., Wong, B., Zhan, Y., Auton, A., Campbell, C.L., Kong, Y., Marcketta, A., Gibbs, R.A., Yu, F., Antunes, L., Bainbridge, M., Muzny, D., Sabo, A., Huang, Z., Wang, J., Coin, L.J.M., Fang, L., Guo, X., Jin, X., Li, G., Li, Q., Li, Y., Li, Z., Lin, H., Liu, B., Luo, R., Shao, H., Xie, Y., Ye, C., Yu, C., Zhang, F., Zheng, H., Zhu, H., Alkan, C., Dal, E., Kahveci, F., Marth, G.T., Garrison, E.P., Kural, D., Lee, W.-P., Fung Leong, W., Stromberg, M., Ward, A.N., Wu, J., Zhang, M., Daly, M.J., DePristo, M.A., Handsaker, R.E., Altshuler, D.M., Banks, E., Bhatia, G., del Angel, G., Gabriel, S.B., Genovese, G., Gupta, N., Li, H., Kashin, S., Lander, E.S., McCarroll, S.A., Nemesh, J.C., Poplin, R.E., Yoon, S.C., Lihm, J., Makarov, V., Clark, A.G., Gottipati, S., Keinan, A., Rodriguez-Flores, J.L., Korbel, J.O., Rausch, T., Fritz, M.H., Stütz, A.M., Flicek, P., Beal, K., Clarke, L., Datta, A., Herrero, J., McLaren, W.M., Ritchie, G.R.S., Smith, R.E., Zerbino, D., Zheng-Bradley, X., Sabeti, P.C., Shlyakhter, I., Schaffner, S.F., Vitti, J., Cooper, D.N., Ball, E.V., Stenson, P.D., Bentley, D.R., Barnes, B., Bauer, M., Keira Cheetham, R., Cox, A., Eberle, M., Humphray, S., Kahn, S., Murray, L., Peden, J., Shaw, R., Kenny, E.E., Batzer, M.A., Konkel, M.K., Walker, J.A., MacArthur, D.G., Lek, M., Sudbrak, R., Amstislavskiy, V.S., Herwig, R., Mardis, E.R., Ding, L., Koboldt, D.C., Larson, D., Ye, K., Gravel, S., The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Molecular Biology Laboratory, E.B.I., Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, National Eye Institute, N., 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393
3. Balding, D., Moltke, I., Marioni, J. (Eds.), 2019. Handbook of statistical genomics, 4th ed. Hoboken, NJ : Wiley.

4. Benton, M.L., Abraham, A., LaBella, A.L., Abbot, P., Rokas, A., Capra, J.A., 2021. The influence of evolutionary history on human health and disease. Nature Reviews Genetics 22, 269–283. https://doi.org/10.1038/s41576-020-00305-9
5. Browning, S.R., Browning, B.L., 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics 81, 1084–1097. https://doi.org/10.1086/521987
6. Bukantis, A., Kažys, J., 2020. 330 Years of Vilnius' Climate: History and Future. PROC 10, 10. https://doi.org/10.15388/Klimatokaita.2020.3
7. Cadzow, M., Boocock, J., Nguyen, H.T., Wilcox, P., Merriman, T.R., Black, M.A., 2014. A bioinformatics workflow for detecting signatures of selection in genomic data. Front. Genet. 5. https://doi.org/10.3389/fgene.2014.00293
8. Česnys, G., Balčiūnienė, I., 1988. Senųjų Lietuvos gyventojų antropologija. Mokslas, Vilnius.
9. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7. https://doi.org/10.1186/s13742-015-0047-8
10. Cook, D.E., Andersen, E.C., 2017. VCF-kit: Assorted utilities for the variant call format. Bioinformatics btx011. https://doi.org/10.1093/bioinformatics/btx011
11. Corona, E., Wang, L., Ko, D., Patel, C.J., 2018. Systematic detection of positive selection in the human-pathogen interactome and lasting effects on infectious disease susceptibility. PLOS ONE 13, e0196676. https://doi.org/10.1371/journal.pone.0196676
12. Domarkienė, I., 2014. Lithuanian population genetic structure definition on the basis of analysis of disease associated single nucleotide polymorphisms (PhD Dissertation). Vilnius university.
13. Esoh, K.K., Apinjoh, T.O., Nyanjom, S.G., Wonkam, A., Chimusa, E.R., Amenga-Etego, L., Amambua-Ngwa, A., Achidi, E.A., 2021. Fine scale human genetic structure in three regions of Cameroon reveals episodic diversifying selection. Scientific Reports 11, 1039. https://doi.org/10.1038/s41598-020-79124-1
14. Fan, S., Hansen, M.E.B., Lo, Y., Tishkoff, S.A., 2016. Going global by adapting local: A review of recent human adaptation. Science 354, 54–59. https://doi.org/10.1126/science.aaf5098
15. Gautier, M., Klassmann, A., Vitalis, R., 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. Mol Ecol Resour 17, 78–90. https://doi.org/10.1111/1755-0998.12634
16. Gilad, Y., Bustamante, C.D., Lancet, D., Pääbo, S., 2003. Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73, 489–501. https://doi.org/10.1086/378132
17. Gimbutienė, M., 1985. Baltai priešistoriniais laikais: etnogenezė, materialinė kultūra ir mitologija. Mokslas, Vilnius.
18. Guénard, F., Bouchard-Mercier, A., Rudkowska, I., Lemieux, S., Couture, P., Vohl, M.-C., 2017. Genome-Wide Association Study of Dietary Pattern Scores. Nutrients 9. https://doi.org/10.3390/nu9070649
19. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R.G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S.L., Risch, R., Rojo Guerra, M.A., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K.W., Reich, D., 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522, 207–211. https://doi.org/10.1038/nature14317
20. Hider, J.L., Gittelman, R.M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J.M., Parra, E.J., 2013. Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. BMC Evolutionary Biology 13, 150. https://doi.org/10.1186/1471-2148-13-150
21. Hu, Y., Shmygelska, A., Tran, D., Eriksson, N., Tung, J.Y., Hinds, D.A., 2016. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nature Communications 7, 10448. https://doi.org/10.1038/ncomms10448

22. Huang, L., Mao, Y., Yang, J., Li, Yuanfeng, Li, Yang, Yang, Z., 2018. Mutation screening of the USH2A gene in retinitis pigmentosa and USHER patients in a Han Chinese population. Eye 32, 1608–1614. https://doi.org/10.1038/s41433-018-0130-3
23. Ilardo, M.A., Moltke, I., Korneliussen, T.S., Cheng, J., Stern, A.J., Racimo, F., de Barros Damgaard, P., Sikora, M., Seguin-Orlando, A., Rasmussen, S., van den Munckhof, I.C.L., ter Horst, R., Joosten, L.A.B., Netea, M.G., Salingkat, S., Nielsen, R., Willerslev, E., 2018. Physiological and Genetic Adaptations to Diving in Sea Nomads. Cell 173, 569-580.e15. https://doi.org/10.1016/j.cell.2018.03.054
24. Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O.P.H., Paunio, T., Pedersen, N.L., deFaire, U., Järvelin, M.-R., Saharinen, J., Freimer, N., Ripatti, S., Purcell, S., Collins, A., Daly, M.J., Palotie, A., Peltonen, L., 2008. The genome-wide patterns of variation expose significant substructure in a founder population. Am J Hum Genet 83, 787–794. https://doi.org/10.1016/j.ajhg.2008.11.005
25. Johannesdottir, S.A., Schmidt, M., Horváth-Puhó, E., Sørensen, H.T., 2012. Autoimmune skin and connective tissue diseases and risk of venous thromboembolism: a population-based case-control study. Journal of Thrombosis and Haemostasis 10, 815–821. https://doi.org/10.1111/j.1538-7836.2012.04666.x
26. Jones, E.R., Zarina, G., Moiseyev, V., Lightfoot, E., Nigst, P.R., Manica, A., Pinhasi, R., Bradley, D.G., 2017. The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers. Curr Biol 27, 576–582. https://doi.org/10.1016/j.cub.2016.12.060
27. Karlsson, E.K., Kwiatkowski, D.P., Sabeti, P.C., 2014. Natural selection and infectious disease in human populations. Nature Reviews Genetics 15, 379–393. https://doi.org/10.1038/nrg3734
28. Kasekamp, A., 2018. A History of the Baltic States, 2nd ed. Red Globe Press, London.
29. Kasperavičiūtė, D., Kučinskas, V., 2002. Variability of the human mitochondrial DNA control region sequences in the Lithuanian population. J Appl Genet 43, 255–260.
30. Kerminen, S., Martin, A.R., Koskela, J., Ruotsalainen, S.E., Havulinna, A.S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M.J., Ripatti, S., Pirinen, M., 2019. Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. The American Journal of Human Genetics 104, 1169–1181. https://doi.org/10.1016/j.ajhg.2019.05.001
31. Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., Price, A.L., 2019. Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet 104, 65–75. https://doi.org/10.1016/j.ajhg.2018.11.008
32. Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511623486
33. Kirkpatric, 2016. It's a (Selective) Sweep for the Good Genes! | BEACON. URL https://www3.beacon-center.org/blog/2016/06/27/its-a-selective-sweep-for-the-good-genes/ (accessed 2.12.21).
34. Kučinskas, V., 2017. Jotvingių genomo šaknų beieškant, in: TERRA JATWEZENORUM: Jotvingių Krašto Istorijos Paveldo Metraštis. pp. 53–65.
35. Kučinskas, V., 2004. Genomo įvairovė: lietuviai Europoje. Spalvų Šalis, Vilnius.
36. Kučinskas, V., 2001. Population genetics of Lithuanians. Ann Hum Biol 28, 1–14. https://doi.org/10.1080/03014460150201832
37. Kučinskas, V., 1994. Human mitochondrial DNA variation in Lithuania. Anthropol Anz 52, 289–295.
38. Kučinskas, V., Radikas, J., Rasmuson, M., 1994. Genetic Diversity in the Lithuanian Rural Population as Illustrated by Variation in the ABO and Rh(D) Blood Groups. HHE 44, 344–349. https://doi.org/10.1159/000154242
39. Laitinen, V., Lahermo, P., Sistonen, P., Savontaus, M.-L., 2002. Y-Chromosomal Diversity Suggests that Baltic Males Share Common Finno-Ugric-Speaking Forefathers. Human Heredity 53, 68–78.
40. Lamnidis, T.C., Majander, K., Jeong, C., Salmela, E., Wessman, A., Moiseyev, V., Khartanovich, V., Balanovsky, O., Ongyerth, M., Weihmann, A., Sajantila, A., Kelso, J., Pääbo, S., Onkamo, P., Haak, W., Krause, J., Schiffels, S., 2018. Ancient Fennoscandian

genomes reveal origin and spread of Siberian ancestry in Europe. Nature Communications 9, 5018. https://doi.org/10.1038/s41467-018-07483-5

41. Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P.H., Schraiber, J.G., Castellano, S., Lipson, M., Berger, B., Economou, C., Bollongino, R., Fu, Q., Bos, K.I., Nordenfelt, S., Li, H., de Filippo, C., Prüfer, K., Sawyer, S., Posth, C., Haak, W., Hallgren, F., Fornander, E., Rohland, N., Delsate, D., Francken, M., Guinet, J.-M., Wahl, J., Ayodo, G., Babiker, H.A., Bailliet, G., Balanovska, E., Balanovsky, O., Barrantes, R., Bedoya, G., Ben-Ami, H., Bene, J., Berrada, F., Bravi, C.M., Brisighelli, F., Busby, G.B.J., Cali, F., Churnosov, M., Cole, D.E.C., Corach, D., Damba, L., van Driem, G., Dryomov, S., Dugoujon, J.-M., Fedorova, S.A., Gallego Romero, I., Gubina, M., Hammer, M., Henn, B.M., Hervig, T., Hodoglugil, U., Jha, A.R., Karachanak-Yankova, S., Khusainova, R., Khusnutdinova, E., Kittles, R., Kivisild, T., Klitz, W., Kučinskas, V., Kushniarevich, A., Laredj, L., Litvinov, S., Loukidis, T., Mahley, R.W., Melegh, B., Metspalu, E., Molina, J., Mountain, J., Näkkäläjärvi, K., Nesheva, D., Nyambo, T., Osipova, L., Parik, J., Platonov, F., Posukh, O., Romano, V., Rothhammer, F., Rudan, I., Ruizbakiev, R., Sahakyan, H., Sajantila, A., Salas, A., Starikovskaya, E.B., Tarekegn, A., Toncheva, D., Turdikulova, S., Uktveryte, I., Utevska, O., Vasquez, R., Villena, M., Voevoda, M., Winkler, C.A., Yepiskoposyan, L., Zalloua, P., Zemunik, T., Cooper, A., Capelli, C., Thomas, M.G., Ruiz-Linares, A., Tishkoff, S.A., Singh, L., Thangaraj, K., Villems, R., Comas, D., Sukernik, R., Metspalu, M., Meyer, M., Eichler, E.E., Burger, J., Slatkin, M., Pääbo, S., Kelso, J., Reich, D., Krause, J., 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413. https://doi.org/10.1038/nature13673
42. Li, M.-H., Strandén, I., Tiirikka, T., Sevón-Aimonen, M.-L., Kantanen, J., 2011. A Comparison of Approaches to Estimate the Inbreeding Coefficient and Pairwise Relatedness Using Genomic and Pedigree Data in a Sheep Population. PLoS One 6. https://doi.org/10.1371/journal.pone.0026256
43. Limaye, N., Boon, L.M., Vikkula, M., 2009. From germline towards somatic mutations in the pathophysiology of vascular anomalies. Human Molecular Genetics 18, R65–R74. https://doi.org/10.1093/hmg/ddp002
44. Liu, M., Jin, H.S., Park, S., 2020. Protein and fat intake interacts with the haplotype of PTPN11_rs11066325, RPH3A_rs886477, and OAS3_rs2072134 to modulate serum HDL concentrations in middle-aged people. Clin Nutr 39, 942–949. https://doi.org/10.1016/j.clnu.2019.03.039
45. Liu, X., Ong, R.T.-H., Pillai, E.N., Elzein, A.M., Small, K.S., Clark, T.G., Kwiatkowski, D.P., Teo, Y.-Y., 2013. Detecting and Characterizing Genomic Signatures of Positive Selection in Global Populations. Am J Hum Genet 92, 866–881. https://doi.org/10.1016/j.ajhg.2013.04.021
46. Loewe, L., 2008. Negative selection. Nature Education 1.
47. Mittnik, A., Wang, C.-C., Pfrengle, S., Daubaras, M., Zariņa, G., Hallgren, F., Allmäe, R., Khartanovich, V., Moiseyev, V., Tõrv, M., Furtwängler, A., Andrades Valtueña, A., Feldman, M., Economou, C., Oinonen, M., Vasks, A., Balanovska, E., Reich, D., Jankauskas, R., Haak, W., Schiffels, S., Krause, J., 2018. The genetic prehistory of the Baltic Sea region. Nature Communications 9, 442. https://doi.org/10.1038/s41467-018-02825-9
48. Mozaffarian, D., Kabagambe, E.K., Johnson, C.O., Lemaitre, R.N., Manichaikul, A., Sun, Q., Foy, M., Wang, L., Wiener, H., Irvin, M.R., Rich, S.S., Wu, H., Jensen, M.K., Chasman, D.I., Chu, A.Y., Fornage, M., Steffen, L., King, I.B., McKnight, B., Psaty, B.M., Djoussé, L., Chen, I.Y.-D., Wu, J.H.Y., Siscovick, D.S., Ridker, P.M., Tsai, M.Y., Rimm, E.B., Hu, F.B., Arnett, D.K., 2015. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. Am J Clin Nutr 101, 398–406. https://doi.org/10.3945/ajcn.114.094557
49. Ng, M.C.Y., Graff, M., Lu, Y., Justice, A.E., Mudgal, P., Liu, C.-T., Young, K., Yanek, L.R., Feitosa, M.F., Wojczynski, M.K., Rand, K., Brody, J.A., Cade, B.E., Dimitrov, L., Duan, Q., Guo, X., Lange, L.A., Nalls, M.A., Okut, H., Tajuddin, S.M., Tayo, B.O., Vedantam, S., Bradfield, J.P., Chen, G., Chen, W.-M., Chesi, A., Irvin, M.R., Padhukasahasram, B., Smith, J.A., Zheng, W., Allison, M.A., Ambrosone, C.B., Bandera, E.V., Bartz, T.M., Berndt, S.I., Bernstein, L., Blot, W.J., Bottinger, E.P., Carpten, J., Chanock, S.J., Chen, Y.-D.I., Conti,

D.V., Cooper, R.S., Fornage, M., Freedman, B.I., Garcia, M., Goodman, P.J., Hsu, Y.-H.H., Hu, J., Huff, C.D., Ingles, S.A., John, E.M., Kittles, R., Klein, E., Li, J., McKnight, B., Nayak, U., Nemesure, B., Ogunniyi, A., Olshan, A., Press, M.F., Rohde, R., Rybicki, B.A., Salako, B., Sanderson, M., Shao, Y., Siscovick, D.S., Stanford, J.L., Stevens, V.L., Stram, A., Strom, S.S., Vaidya, D., Witte, J.S., Yao, J., Zhu, X., Ziegler, R.G., Zonderman, A.B., Adeyemo, A., Ambs, S., Cushman, M., Faul, J.D., Hakonarson, H., Levin, A.M., Nathanson, K.L., Ware, E.B., Weir, D.R., Zhao, W., Zhi, D., Group, T.B.M.D. in C.S. (BMDCS), Arnett, D.K., Grant, S.F.A., Kardia, S.L.R., Oloapde, O.I., Rao, D.C., Rotimi, C.N., Sale, M.M., Williams, L.K., Zemel, B.S., Becker, D.M., Borecki, I.B., Evans, M.K., Harris, T.B., Hirschhorn, J.N., Li, Y., Patel, S.R., Psaty, B.M., Rotter, J.I., Wilson, J.G., Bowden, D.W., Cupples, L.A., Haiman, C.A., Loos, R.J.F., North, K.E., 2017. Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. PLOS Genetics 13, e1006719. https://doi.org/10.1371/journal.pgen.1006719

50. Niarchou, M., Byrne, E.M., Trzaskowski, M., Sidorenko, J., Kemper, K.E., McGrath, J.J., O' Donovan, M.C., Owen, M.J., Wray, N.R., 2020. Genome-wide association study of dietary intake in the UK biobank study and its associations with schizophrenia and other traits. Translational Psychiatry 10, 1–11. https://doi.org/10.1038/s41398-020-0688-y
51. Oscanoa, J., Sivapalan, L., Gadaleta, E., Dayem Ullah, A.Z., Lemoine, N.R., Chelala, C., 2020. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). Nucleic Acids Research 48, W185–W192. https://doi.org/10.1093/nar/gkaa420
52. Pankratov, V., Montinaro, F., Kushniarevich, A., Hudjashov, G., Jay, F., Saag, L., Flores, R., Marnetto, D., Seppel, M., Kals, M., Võsa, U., Taccioli, C., Möls, M., Milani, L., Aasa, A., Lawson, D.J., Esko, T., Mägi, R., Pagani, L., Metspalu, A., Metspalu, M., 2020. Differences in local population history at the finest level: the case of the Estonian population. European Journal of Human Genetics 28, 1580–1591. https://doi.org/10.1038/s41431-020-0699-4
53. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient admixture in human history. Genetics 192, 1065–1093. https://doi.org/10.1534/genetics.112.145037
54. Pavlidis, P., Alachiotis, N., 2017. A survey of methods and tools to detect recent and strong positive selection. Journal of Biological Research-Thessaloniki 24, 7. https://doi.org/10.1186/s40709-017-0064-0
55. Pemberton, T.J., Wang, C., Li, J.Z., Rosenberg, N.A., 2010. Inference of Unexpected Genetic Relatedness among Individuals in HapMap Phase III. Am J Hum Genet 87, 457–464. https://doi.org/10.1016/j.ajhg.2010.08.014
56. Piras, I.S., De Montis, A., Calò, C.M., Marini, M., Atzori, M., Corrias, L., Sazzini, M., Boattini, A., Vona, G., Contu, L., 2012. Genome-wide scan with nearly 700 000 SNPs in two Sardinian sub-populations suggests some regions as candidate targets for positive selection. Eur J Hum Genet 20, 1155–1161. https://doi.org/10.1038/ejhg.2012.65
57. Pistis, G., Piras, I., Pirastu, N., Persico, I., Sassu, A., Picciau, A., Prodi, D., Fraumene, C., Mocci, E., Manias, M.T., Atzeni, R., Cosso, M., Pirastu, M., Angius, A., 2009. High Differentiation among Eight Villages in a Secluded Area of Sardinia Revealed by Genome-Wide High Density SNPs Analysis. PLOS ONE 4, e4654. https://doi.org/10.1371/journal.pone.0004654
58. Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., Yang, J., Jones, S., Beaumont, R., Croteau-Chonka, D.C., Winkler, T.W., GIANT Consortium, Hattersley, A.T., Loos, R.J.F., Hirschhorn, J.N., Visscher, P.M., Frayling, T.M., Yaghootkar, H., Lindgren, C.M., 2019. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum Mol Genet 28, 166–174. https://doi.org/10.1093/hmg/ddy327
59. Pybus, M., Dall'Olio, G.M., Luisi, P., Uzkudun, M., Carreño-Torres, A., Pavlidis, P., Laayouni, H., Bertranpetit, J., Engelken, J., 2014. 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. Nucleic Acids Research 42, D903–D909. https://doi.org/10.1093/nar/gkt1188
60. R Core Team, 2018. R: The R Project for Statistical Computing. Austria, Vienna.

61. Richardson, J.L., Urban, M.C., Bolnick, D.I., Skelly, D.K., 2014. Microgeographic adaptation and the spatial scale of evolution. Trends in Ecology & Evolution 29, 165–176. https://doi.org/10.1016/j.tree.2014.01.002
62. Richardson, T.G., Sanderson, E., Palmer, T.M., Ala-Korpela, M., Ference, B.A., Smith, G.D., Holmes, M.V., 2020. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. PLOS Medicine 17, e1003062. https://doi.org/10.1371/journal.pmed.1003062
63. Rimantienė, R., 1996. Akmens amžius Lietuvoje. Žiburio leidykla, Vilnius.
64. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832–837. https://doi.org/10.1038/nature01140
65. Sellayah, D., Cagampang, F.R., Cox, R.D., 2014. On the Evolutionary Origins of Obesity: A New Hypothesis. Endocrinology 155, 1573–1588. https://doi.org/10.1210/en.2013-2103
66. Senvaitytė, D., 2014. Lithuanian Ethnic Culture, in: History of Lithuanian Culture. Versus Aureus, Vytautas Magnus University, pp. 9–50.
67. Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.
68. Tang, K., Thornton, K.R., Stoneking, M., 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol 5. https://doi.org/10.1371/journal.pbio.0050171
69. Tracewska, A.M., Kocyła-Karczmarewicz, B., Rafalska, A., Murawska, J., Jakubaszko-Jablonska, J., Rydzanicz, M., Stawiński, P., Ciara, E., Khan, M.I., Henkes, A., Hoischen, A., Gilissen, C., van de Vorst, M., Cremers, F.P.M., Płoski, R., Chrzanowska, K.H., 2019. Genetic Spectrum of ABCA4-Associated Retinal Degeneration in Poland. Genes (Basel) 10. https://doi.org/10.3390/genes10120959
70. Uktverytė, I., 2014. Lietuvos etnolingvistinių grupių genetinės struktūros analizė remiantis informatyviais genomo žymenimis (PhD Dissertation). Vilnius university.
71. Urnikytė, A., Flores-Bello, A., Mondal, M., Molyte, A., Comas, D., Calafell, F., Bosch, E., Kučinskas, V., 2019. Patterns of genetic structure and adaptive positive selection in the Lithuanian population from high-density SNP data. Scientific Reports 9, 9163. https://doi.org/10.1038/s41598-019-45746-3
72. Vatsiou, A.I., Bazin, E., Gaggiotti, O.E., 2016. Detection of selective sweeps in structured populations: a comparison of recent methods. Mol Ecol 25, 89–103. https://doi.org/10.1111/mec.13360
73. Vitti, J.J., Grossman, S.R., Sabeti, P.C., 2013. Detecting natural selection in genomic data. Annu Rev Genet 47, 97–120. https://doi.org/10.1146/annurev-genet-111212-133526
74. Walsh, S., Pagani, L., Xue, Y., Laayouni, H., Tyler-Smith, C., Bertranpetit, J., 2020. Positive selection in admixed populations from Ethiopia. BMC Genetics 21, 108. https://doi.org/10.1186/s12863-020-00908-5
75. Weigand, H., Leese, F., 2018. Detecting signatures of positive selection in non-model species using genomic data. Zoological Journal of the Linnean Society 184, 528–583. https://doi.org/10.1093/zoolinnean/zly007

VILNIUS UNIVERSITY

Systems Biology Master ‘s Program

Karolina Mukauskaitė

Master‘s Thesis

SIGNATURES OF POSITIVE NATURAL SELECTION IN THE LITHUANIAN ETHNOLINGUISTIC GROUPS FROM HIGH-DENSITY SNP DATA

SUMMARY

The main purpose of this master thesis is to analyze signatures of positive natural selection of two main ethnolinguistic groups, Žemaičiai and Aukštaičiai, from genome-wide SNP genotyping data. The main question is whether natural selection is acting equally on genomic groups under study.

Identifying positive natural selection in the human population provides an opportunity to understand the phenotypes of modern humans and their ability to adapt to changing environmental conditions. The genome areas affected by the natural selection identified in this research fill the knowledge gaps in this area and help determine their effect on the phenotype and identify adaptive alleles.

Genome-wide high-density SNP genotype data in 425 individuals from six geographical regions in Lithuania were used to find signatures of natural positive selection in the Lithuanian population and check for population structure. The results show that Lithuania is a homogeneous population and that all ethnolinguistic groups form one cluster. Different genomic regions under natural positive selection were detected in Aukštaičiai and Žemaičiai ethnolinguistic groups. Among the top signatures of positive selection detected in the ethnolinguistic populations, we identified several candidate genes related to diet (*SMKR1*, *SEMA6D* and *MRA*), body mass index BMI (*NRF1*, *LRRC36, ZDHHC1, HSD11B2*, *PNLIPRP3, PNLIP*, *DNAH7, HLA-DMA, STAM2, STK17B*, *DNAJB8*).

The results support the theory that positive natural selection can occur even at the micro-geographical levels. Dependency on an ethnolinguistic region might be considered an important factor for the genetic data analysis.

SUMMARY IN LITHUANIAN

VILNIAUS UNIVERSITETAS
Systemų Biologijos Magistro Programa
Karolina Mukauskaitė

Magistrins darbas

Kandidatinių teigiamos gamtinės atrankos veikiamų genomo sričių nustatymas Lietuvos populiacijos etnoligvistinėse grupėse, panaudojant plataus masto vieno nukleotido polimorfizmo genotipavimo duomenis

APIBENDRINIMAS

Pagrindinis magistrinio darbo tikslas - išanalizuoti dviejų pagrindinių etnolingvistinių grupių - Žemaičių ir Aukštaičių - teigiamos gamtinės atrankos veikiamas genomo sritis panaudojant plataus masto vieno nukleotido polimorfizmo (VNP) duomenis. Pagrindinis klausimas - ar natūrali atranka vienodai veikia tiriamas etnolingvistines grupes.

Teigiamos natūralios atrankos nustatymas žmonių populiacijoje leidžia išanalizuoti žmonių fenotipus ir kaip jie geba prisitaikyti prie kintančių aplinkos sąlygų. Šiame tyrime nustatytos natūralios atrankos paveiktos genomo sritys užpildo šios srities žinių spragas ir padeda nustatyti jų poveikį fenotipui bei nustatyti adaptacinius alelius.

Plataus masto VNP duomenys buvo surinkti atlikus 425 asmenų, iš šešių Lietuvos geografinių regionų, genotipavimą ir buvo panaudojami natūralios gamtinės teigiamos atrankos signalams surasti ir populiacijos struktūrai patikrinti. Rezultatai rodo, kad Lietuva yra vienalytė populiacija ir visos etnolingvistinės grupės susigrupuoja vieną klasterį. Skirtingi natūralios teigiamos atrankos genomo regionai buvo aptikti Aukštaičių ir Žemaičių etnolingvistinėse grupėse. Tarp stipriausių etnolingvistinėse populiacijose nustatytų teigiamos atrankos paveiktų genomo regionų, nustatyti keli, kuriuose aptinkami su žmonių mityba susiję genai (SMKR1, SEMA6D ir MRA) ir su kūno masės indeksu (KMI) susiję genai (NRF1, LRRC36, ZDHHC1, HSD11B2, PNLIPRP3, PNLIP, DNAH7, HLA- DMA, STAM2, STK17B, DNAJB8).

Rezultatai patvirtina teoriją, kad teigiama natūrali atranka gali atsirasti net mikro-geografiniame lygmenyje.

APPENDICES

Appendix 1. Individuals in the Lithuanian population filtered out after data Quality control

ID	Kinship	F	MDS outlier
781	with ID 785 (0.255)	0.01562	NO
875	with ID 886 (0.2516)	0.01584	NO
972	-	0.02488	NO
1271	-	0.01852	NO
1074	with ID 1105 (0.5012)	-	YES
1105	with ID 1074 (0.5012)	-	YES
333	with ID 813 (0.5001)	-	YES
813	with ID 333 (0.5001)	-	YES
1401	with ID 356 (0.3843)	-	YES
1158	duplicate	-	YES
356	with ID 1401 (0.3843)	-	YES
1158	duplicate	-	YES
294	with ID 793 (0.2485)	-	NO
748	with ID 867 (0.2203)	-	NO
1061	with ID 1056 (0.1416)	-	NO
197	with ID 246 (0.1268)	-	NO
348	with ID 376 (0.2452)	-	NO
665	with ID 616 (0.2478)	-	NO
423	with ID 429 (0.2742)	-	NO

Appendix 2. Genomic regions under positive selection identified with two EHH-based approaches in two ethnolinguistic groups of the Lithuanian population

CHR	START	END	[a]SNP (XPEHH)	[b]SNP (Rsb)	Populations	Genes
1	145564518	145730160	7	7	Aukštaičiai - CEU	*ANKRD35, PIAS3, NUDT17, POLR3C, RNF115, CD160, PDZK1*
1	94550555	94575440	5	10	Aukštaičiai - CEU	*ABCA4*
4	23822328	23901275	5	6	Aukštaičiai - CEU	*PPARGC1A*
4	8557853	8647103	8	6	Aukštaičiai - FIN	*GPR78, CPZ*
5	50012019	50340102	28	27	Aukštaičiai - CEU	*PARP8*
5	150381243	150402490	10	9	Aukštaičiai - YRI	*GPX3*
5	50012019	50254200	6	13	Aukštaičiai - YRI	*PARP8*
5	70975867	70997465	4	4	Aukštaičiai - YRI	[c]*NA*
6	150210681	150247274	4	4	Aukštaičiai - YRI	*RAET1E, RAET1G*
6	27220890	27925367	7	13	Aukštaičiai - YRI	*PRSS16, POM121L2, ZNF391, ZNF184, HIST1H2BL, HIST1H2AI, HIST1H3H, HIST1H2AJ, HIST1H2BM, HIST1H4J, HIST1H4K, HIST1H2AK, HIST1H2BN, HIST1H2AL, HIST1H1B, HIST1H3I, HIST1H4L, HIST1H3J,*

						HIST1H2AM, HIST1H2BO, OR2B2, OR2B6
6	28130194	28869173	40	15	Aukštaičiai - YRI	*ZSCAN9, ZKSCAN4, NKAPL, PGBD1, ZSCAN31, ZKSCAN3, ZSCAN12, ZSCAN23, GPX6, GPX5, SCAND3,*
6	29323838	29528318	4	13	Aukštaičiai - YRI	*OR5V1, OR12D3, OR12D2, OR11A1, OR10C1, OR2H1, MAS1L, UBD, GABBR1*
7	129199328	129350170	4	4	Aukštaičiai - CEU	*NRF1*
7	129187430	129979363	11	5	Aukštaičiai - FIN	*NRF1, UBE2H, ZC3HC1, KLHDC10, TMEM209, SSMEM1, CPA2, CPA4*
8	6817024	6886095	7	6	Aukštaičiai - CEU	*DEFA1, DEFA1B,DEFA3*
9	12352971	129826261	5	9	Aukštaičiai - CEU	*TYRP1, LURAP1L, MPDZ,NFIB*
9	130172485	130311404	4	4	Aukštaičiai - CEU	*ZNF79, RPL12, LRSAM1, FAM129B*
9	71156551	71280103	11	9	Aukštaičiai - CEU	*NA*
9	27124745	27164220	9	5	Aukštaičiai - FIN	*TEK*
9	71156380	71280103	12	18	Aukštaičiai - YRI	*NA*
10	127731197	127755954	5	4	Aukštaičiai - CEU	*ADAM12*

10	33296362	33444837	7	10	Aukštaičiai - CEU	*NA*
10	73740613	73837420	16	18	Aukštaičiai - CEU	*CHST3, SPOCK2*
13	100070907	100080579	4	4	Aukštaičiai - YRI	*NA*
14	66460742	66558515	7	11	Aukštaičiai - YRI	*CTD-2014B16.3*
15	51644775	51673630	5	4	Aukštaičiai - CEU	*GLDN*
16	31394179	31682532	9	7	Aukštaičiai - FIN	*ITGAX, ITGAD, COX6A2, ZNF843, ARMC5, TGFB1I1, SLC5A2, C16orf58, AHSP*
17	4722606	4797305	7	5	Aukštaičiai - FIN	*PLD2, MINK1*
17	4801288	4891956	11	6	Aukštaičiai - FIN	*MINK1, CHRNE, C17orf107, GP1BA, SLC25A11, RNF167, PFN1, ENO3, SPAG7, CAMTA2, INCA1,*
17	59315145	59345321	8	8	Aukštaičiai - YRI	*BCAS3*
18	10424829	10473726	6	7	Aukštaičiai - CEU	*APCDD1*
18	22133526	22491296	4	10	Aukštaičiai - CEU	*NA*
20	25027630	25249734	14	19	Aukštaičiai - CEU	*ACSS1, VSX1, ENTPD6, AL035252.1, PYGB*
20	25043111	25581424	28	26	Aukštaičiai - FIN	*VSX1, ENTPD6, PYGB, ABHD12, GINS1, NINL*

1	1:145564518	1:145730160	6	5	Žemaičiai-CEU	*ANKRD35, PIAS3, NUDT17, POLR3C, RNF115, CD160, PDZK1*
1	1:77863068	1:77924890	13	17	Žemaičiai-CEU	*AK5*
2	2:53846965	2:53854488	4	5	Žemaičiai-CEU	*GPR75-ASB3*
2	2:152919055	2:153391282	9	4	Žemaičiai-FIN	*CACNB4, STAM2, FMNL2*
2	2:196328390	2:197078179	9	5	Žemaičiai-YRI	*SLC39A10, DNAH7, STK17B, HECW2*
3	3:1316884	3:1323078	6	5	Žemaičiai-CEU	*CNTN6*
3	3:153507426	3:153538769	8	5	Žemaičiai-CEU	*NA*
4	4:124337522	4:124387560	9	10	Žemaičiai-CEU	*NA*
4	4:32594107	4:32640624	5	7	Žemaičiai-FIN	*NA*
4	4:8557853	4:8596478	5	5	Žemaičiai-FIN	*GPR78, CPZ*
5	5:50146929	5:50340102	23	23	Žemaičiai-CEU	*LINC02106*
5	5:150381243	5:150402490	8	6	Žemaičiai-YRI	*GPX3*
6	6:32902001	6:32974934	7	4	Žemaičiai-CEU	*HLA-DMB, HLA-DMA, BRD2, HLA-DOA*
6	6:57121684	6:57182068	11	10	Žemaičiai-CEU	*PRIM2*
6	6:11643081	6:11764205	6	7	Žemaičiai-FIN	*ADTRP*
6	6:150210681	6:150247274	4	4	Žemaičiai-YRI	*RAET1E, RAET1G*

6	6:27220890	6:27925827	7	13	Žemaičiai-YRI	*PRSS16, POM121L2, ZNF391, ZNF184, HIST1H2BL, HIST1H2AI, HIST1H3H, HIST1H2AJ, HIST1H2BM, HIST1H4J, HIST1H4K, HIST1H2AK, HIST1H2BN, HIST1H2AL, HIST1H1B, HIST1H3I, HIST1H4L, HIST1H3J, HIST1H2AM, HIST1H2BO, OR2B2, OR2B6*
6	6:28130194	6:28869173	50	20	Žemaičiai-YRI	*ZSCAN9, ZKSCAN4,*
6	6:29323838	6:29528318	4	24	Žemaičiai-YRI	*OR5V1, OR12D3, OR12D2, OR11A1, OR10C1, OR2H1, MAS1L, UBD, GABBR1*
8	8:6801712	8:6829137	4	7	Žemaičiai-CEU	*pseudo DEFA8P, DEFA9P, DEFA10P*
9	9:12352971	9:12510515	12	10	Žemaičiai-CEU	*PTPRD-AS2;TYRP1*
9	9:71156551	9:71280103	7	15	Žemaičiai-YRI	*NA*
10	10:127731197	10:127755954	9	12	Žemaičiai-CEU	*ADAM12*
10	10:73740613	10:73820622	9	20	Žemaičiai-CEU	*CHST3, SPOCK2*
11	11:43201350	11:43550070	6	4	Žemaičiai-CEU	*API5, TTC17,*
12	12:84152210	12:84164303	8	5	Žemaičiai-CEU	*NA*

12	12:104808716	12:104845732	6	4	Žemaičiai-FIN	*NA*
12	12:113124351	12:113209519	6	9	Žemaičiai-FIN	*RPH3A*
13	13:100070907	13:100080579	5	4	Žemaičiai-YRI	*NA*
14	14:65654201	14:66558515	9	7	Žemaičiai-YRI	*FUT8*
15	15:51545454	15:51673630	17	17	Žemaičiai-CEU	*GLDN, CYP19A1*
15	15:48042146	15:48122127	12	9	Žemaičiai-FIN	*SEMA6D, SLC24A5*
16	16:31523113	16:31659915	6	7	Žemaičiai-FIN	*AHSP*
17	17:4712757	17:4917402	28	17	Žemaičiai-FIN	*PLD2, MINK1, CHRNE, C17orf107, GP1BA, SLC25A11, RNF167, PFN1, ENO3, SPAG7, CAMTA2, INCA1, KIF1C*
17	17:59315145	17:59352163	9	8	Žemaičiai-YRI	*BCAS3*
18	18:10405177	18:10473505	7	12	Žemaičiai-CEU	*APCDD1*
20	20:25027331	20:25566485	15	23	Žemaičiai-CEU	*ACSS1, VSX1, ENTPD6, AL035252.1., PYGB, ABHD12, PPIAP2, GINS1, NINL*
20	20:25043111	20:25595868	31	36	Žemaičiai-FIN	*VSX1, ENTPD6, AL035252.1, PYGB, ABHD12, PPIAP2, GINS1, NINL, NANP*