VILNIUS UNIVERSITY

FACULTY OF MEDICINE

DEPARTMENT OF HUMAN AND MEDICAL GENETICS

INSTITUTE OF BIOMEDICAL SCIENCES

2nd Year Master's Student of Molecular Biotechnology Study Program

Kristina GRIGAITYTĖ

Inferring Archaic Alleles: Assessing Their Impact on Balancing Selection

Master Thesis

Supervisor:

Faculty of Medicine Institute of Biomedical Sciences

Department of Human and Medical Genetics

Dr. Alina URNIKYTĖ

Signature:

Student:

Kristina GRIGAITYTĖ

Signature:

Vilnius, 2024

CONTENTS

ABBREVIATIONS

ANN - Artificial Neural Networks

API - Application Programming Interface

ArchIE - Archaic Introgression Explore

CNN - Convolutional Neural Networks

HKA - Hudson-Kreitman-Aguadé

HLA - Human Leukocyte Antigen

HIV - Human Immunodeficiency Virus

IFS - Individual Frequency Spectrum

JSON - JavaScript Object Notation

LD - Linkage Disequilibrium

MHC - Major Histocompatibility Complex

MK - McDonald-Kreitman

ML - Machine Learning

SARS-CoV - Severe Acute Respiratory Syndrome Coronavirus

SNP - Single Nucleotide Polymorphism

SFS - Site Frequency Spectrum

URL - Uniform Resource Locator

INTRODUCTION

In population genetics, the importance of balancing selection as a main factor of evolution has been more widely recognized in recent years. However, not as many studies have shown exactly how balancing selection impacts recent or ancient human evolution, even though it is predicted that short-term balancing selection can be quite common in nature (Olivia L. Johnson, 2023).

Balancing selection frequently preserves beneficial genetic variations. Studying these variations in ancient fragments gives an interesting view of the possible adaptive ways that shaped modern human evolution (Sankararaman et al., 2014). This becomes incredibly important when examining immune response-related genes and trying to decipher how ancestors adapted to environmental challenges and various external factors like pathogens (Abi-Rached et al., 2011).

Moreover, population-specific research expands the understanding of the genetic makeup in specific human groups/populations. This additional knowledge not only contributes to the field of genetics, particularly evolutionary genetics, but also holds potential regarding personalized medicine and disease resistance related to specific historical regions or even populations (Enard et al., 2016).

Contemporary Lithuanians are one of such target populations. Lithuanians are an outcome of a blend of ancient Baltic tribes, contributing to the deep and ancient genetic roots. Due to this historical intermingling, it is likely that current Lithuanians retain elements of this ancient genetic makeup within their genome (Urnikyte et al., 2019).

Balancing selection within the recent human evolution of archaic sequences in a population-specific scope is overall a compelling journey into the complex forces that shape the human species. The idea was to search for balancing selection footprints in the Lithuanian population, variants that could potentially be researched further on to provide more insights into human adaptability.

The objective of the research work:

The objective of the research is to identify ancient balancing selection signatures across generations in modern human genomes, employing genetic population statistics and deep learning techniques. Answering the question of which variants or proteins potentially indicate that balancing selection occurred in ancient times but could continue to influence populations today?

The tasks of the research work:

1. To parse and transform the genetic data files to appropriate formats for further analysis.
2. Utilize AI techniques to identify archaic signatures of balancing selection within modern human DNA.
3. Enhance the study by incorporating *in silico* functional analysis of the found archaic balancing selection variants.

1. LITERATURE REVIEW

1.1. Mechanisms of evolution

In populations, genetic diversity is usually maintained over generations, with the frequency of different genes remaining fairly stable in the absence of significant external factors. This genetic balance can be influenced by several key mechanisms, each with different outcomes: mutations, gene flow (or in other words, migration), genetic drift, and natural selection. Like gene flow and genetic drift, mutations are random processes. They do not directly increase an organism's adaptation to the environment. These mechanisms introduce some level of randomness into the genetic makeup of a population and can change gene frequencies in ways that may not favor the reproductive success of an organism (K.A.Stewart, 2019).

In contrast, natural selection is a non-random process that significantly shapes genetic frequencies based on how well organisms are adapted to their environments (Oscar Lao, 2021). It is possible to predict the impact of natural selection on specific genetic variants, especially when looking at alleles that show marked differences from the population's general genetic pattern. These 'outlier loci' are striking because they either confer strong benefits or disadvantages upon the organism, hence becoming the target of strong selective pressure. The fact that they deviate from the norm in the population's genome already proves the intervention of natural selection in retaining or eliminating the given genetic variant. In contrast to predictions of regularity in natural selection, it acts through a diverse array of mechanisms that can change, reduce, or enhance genetic diversity and the distinctiveness of populations and species. Two primary types of natural selection are implicated in changing gene allele frequencies - directional selection and balancing selection (Angela M. Hancock, 2008)

It is widely held that directional selection represents the quintessential model of natural selection. It favors one allele over its alternatives, significantly increasing its frequency. On the other hand, directional selection also involves the suppression of deleterious alleles, hence decreasing their frequency to give way to the more favorable alternative. Whether increasing or decreasing an allele's frequency, directional selection changes the frequency of alleles by favoring one over the other (T. Ryan Gregory, 2009).

Although genetic variation can be reduced by random genetic drift and directional selection, some genetic differences persist for longer periods due to balancing selection, helping to understand how human species evolve and adapt to the environment over the years.

1.2. Balancing selection

Balancing selection (sometimes also referred to as stabilizing selection), was first proposed by Dobzhansky in 1951. This phenomenon arises when two different genetic variants maintain equal frequencies within a population, a state referred to as balanced polymorphism. This implies that both genetic forms could be advantageous (William E. Gundling Jr., 2023) (Figure 1.2.1.). Maintaining a genetic variation is achieved through the compensation of stochastic elimination or fixation of one allele due to genetic drift, thanks to balancing selection (Alber Aqil, 2023).

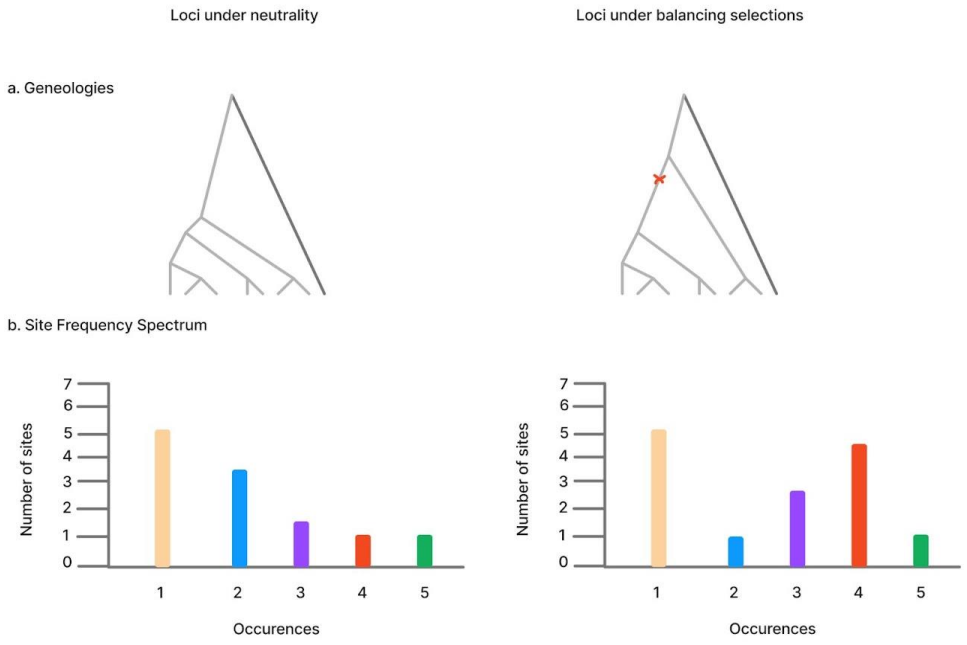

Figure 1.2.1. A graphical outline of balancing selection indicators includes a. The ancestral lineages of genetic locations in a state of neutrality (left) and under the influence of balancing selection (right). The red star marks the emergence of a favored mutation within the lineage. b. The distribution of allele frequencies at each site when neutral (left) and affected by balancing selection (right), with the latter condition often leading to a higher proportion of alleles at moderate frequencies (third and fourth bar) (Deborah Charlesworth, 2006)

1.2.1. Balancing selection mechanisms

There are two main mechanisms for balancing selection that ensure the maintenance of genetic diversity within populations or a group. One of the mechanisms is referred to as heterozygote advantage (or overdominant selection), where individuals with two different alleles of a given gene, heterozygotes, show greater fitness as opposed to both homozygote types (Charlesworth, 2006). This superiority can be due to their resistance to diseases or even their adaptability to environmental changes (Hedrick, 2011). In such a situation, the relative frequency of an allele in a population is an indicator of fitness. The rare alleles are favored, and the frequency thus increases to reach a stable point (Hartl and Clark, 2007). At this point, the relative fitness of each genotype, including the reproductive success and survival rate of all the individuals carrying the different combinations of alleles, determines the equilibrium frequency of that allele. For example, if the heterozygotes have more fitness than either type of homozygote, then the frequency of each of the alleles would get to an equilibrium, which is usually about 50% (Fisher, 1930). However, for the maintenance of genetic diversity, it must lie within a range of, say, 20% to 80%. Beyond that range, it may result in genetic drift (Takahata and Nei, 1990; Ewens and Thomson, 1970).

The second mechanism of negative frequency-dependent selection is the process by which a rare allele is favored, hence increasing in frequency until it reaches equilibrium or experiences negative selection. This mechanism plays a crucial role in maintaining diversity within a population (Mark R. Christie, 2023).

Apart from the above two factors, another major factor is spatiotemporally varying selection. This phenomenon takes place when the benefit of a particular gene variant is dependent on the environment and period. A gene variant could be helpful in one scenario or under specific environmental conditions but harmful in other scenarios. This constant change retains the significance and usefulness of different gene variants for a wide range of setups and times and serves as a prominent factor in maintaining genetic diversity (Bell, 2010).

Besides, balancing selection acts on a variety of timescales, running from very long-term selection, influencing distinct species to short-term, acting only within a population. The duration of the selection has a direct effect on the genomic signature produced (Fijarczyk and Babik, 2015). In this way, we can realize the fine patterns and processes that support the maintenance of genetic diversity.

1.2.2. Balancing selection cases

In contrast to positive and negative selection, well-established instances of balancing selection are less common (Charlesworth and Charlesworth, 2017). In humans, balancing selection plays a crucial role in diversifying genes related to metabolism (Matteo Fumagalli, 2019), among other biological functions (Barbra D Bitarello, 2018). Significantly, it has been discovered that variants influenced by pathogen-driven balancing selection are linked to increased risk for various autoimmune diseases (Matteo Fumagalli, 2011). Consequently, by deciphering the genomic indicators of balancing selection, we can pinpoint prevalent alleles that have crucial functional impacts.

The heterozygote advantage linked with sickle cell anemia in African populations serves as a powerful illustration of genetic traits impacting disease resistance. Sickle cell anemia arises from recessive alleles of the hemoglobin gene, necessitating two copies of the 'diseased' allele for severe symptoms. While typically deleterious, in malaria-prevalent African regions, carriers of one sickle cell allele (HbS allele, rs334) demonstrate resistance to malaria, as the parasites find it difficult to infect the altered shape of red blood cells. This heterozygote advantage means individuals with one normal hemoglobin allele (HbA) and one sickle cell allele (HbS) have increased survival rates in areas with high malaria incidence. The hemoglobin-β locus is crucial, where homozygous individuals for the sickle cell allele (rs334) experience sickle cell anemia, those homozygous for the normal allele are more vulnerable to malaria, and heterozygotes exhibit resistance to malaria with a milder sickle-cell condition (Malaria Genomic Epidemiology Network, 2015, Luzzatto, 2012; Aidoo et al., 2002). This phenomenon provides a clear example of how genetic diversity can confer survival advantages in specific environmental contexts.

One more fitting example is the major histocompatibility complex. The Major Histocompatibility Complex (MHC), known for its dense concentration of transspecies polymorphisms (trans-SNPs, ancient genetic variations maintained across distinct species over extended periods), exemplifies the concept of balancing selection in genetic studies. In-depth research into these polymorphisms, like the well-documented HLA-B*57:01 allele (rs2395029) associated with human immunodeficiency virus (HIV) control but also with hypersensitivity to certain drugs, reveals that individuals heterozygous for certain MHC genes tend to have a more robust immune response without triggering excessive inflammation. This balance between effective pathogen defense and controlled inflammatory response highlights the critical role of genetic diversity within the MHC

in shaping adaptive immunity. The enduring presence of these alleles, such as those found within the *HLA* region on chromosome 6, underscores their evolutionary significance in maintaining health and combating a wide range of pathogens, reflecting a delicate equilibrium between immune efficiency and overactivity (Azevedo et al., 2015, Leffler et al., 2013). Chromosome 6 of Lithuanian whole-genome sequencing data contains Neanderthal introgressed selection fragments with *HLA* genes, which are important for acquired immunity, including *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB*1, and *HLA-DQA1*. Many genes were identified that are connected to innate immunity, including *REG3G*, *MAPK10*, *PHACTR2*, *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1*, *IL17A*, *IL17F*, *DLC1*, *MMP20*, *CLEC7A*, *CDH1*, and *TOM1*. Nonsynonymous variants were detected in the particularly important immunity genes *HLA-A*, *HLA-B*, and *HLA-C*. Using gene ontology analysis, further investigation demonstrated that these genes are involved in a series of pathways due to infections, including HIV, *Listeria monocytogenes* entry, and SARS-CoV-2 interactions. The pathogen–host interaction database PHILM2Web revealed associations between specific human immune-related genes such as *HLA-G*, *HLA-DQB1*, *HLA-B*, and *HLA-DRB1*, and pathogens like human papillomavirus, HIV, and unidentified influenza viruses (Alina Urnikytė, 2023).

As research continues in this field, it is becoming increasingly clear that the history of the human species is deeply intertwined with the genetic complexities of the human immune system, a testament to the power of natural selection and genetic diversity. Selection events, particularly pronounced during infectious disease outbreaks, tend to imprint distinct markers within the human genome.

1.3. Methods to detect balancing selection

In the field of genetics, genetic signatures of recent balancing selection are like those of recent positive selection (Ulas Isildak, 2020). This becomes clear when noting that beneficial alleles typically exist at intermediate frequencies in contemporary genomes (Deborah Charlesworth, 2006). Such frequencies are neither exceedingly rare nor overly common. In balancing selection, this equilibrium is essential for maintaining genetic diversity, whereas in positive selection, it signifies a transitional phase where a beneficial allele is gradually becoming more prevalent. Understanding the precise evolutionary mechanisms behind balancing selection, like

overdominance or negative frequency-dependent selection, continues to be a daunting task (Violaine Llaurens, 2017).

While identifying balancing selection theoretically benefits from a more localized genomic impact compared to positive selection, thus potentially easing the narrowing of causal variants, the situation is actually more complex. Balanced loci in humans that have a clearly defined phenotype or identified selective pressure are unexpectedly rare, highlighting a substantial gap in our comprehension of human evolutionary genetics. To accurately identify genome regions under balancing selection, researchers use a range of statistical tools from population genetics, which are helpful in detecting signs of balancing selection (Table 1.3.1.).

Table 1.3.1. Tools/tests for detecting balancing selection (Anna Fijarcczyk, 2015)

Timescale	Test	Pattern
Intermediate, ancient	HK, BALLET, Tajima's D, Allele frequency spectrum, McDonald Kreitman test	Increased diversity around selected locus, excess of common polymorphisms
Recent, intermediate	Fst outliers	Genetic differentiation between populations
Recent	Long range haplotype test, integrated extended haplotype homozygosity	LD

1.3.1. Tajima's D Test

One of the main tools in the study of balancing selection is Tajima's D, a statistical test introduced by Tajima in 1989 (Tajima, 1989). This test is normally applied to identify anomalies in the frequency of polymorphisms by comparison of two critical measures: π and θw. Tajima's D measures deviations in the site frequency spectrum from the expectations under neutral

evolutionary conditions. A positive Tajima's D, therefore, indicates an excess of alleles at intermediate frequencies - something not expected to be seen by the neutral model of evolution.

On the contrary, a neutral model would normally cause a Tajima's D value to be close to zero. The deviation is a significant indicator of balancing selection, characterized by higher genetic diversity than expected at the selected locus and a prominent presence of polymorphisms occurring at intermediate frequencies (Joanne R. Chapman, 2019)

In this regard, Tajima's, and the related statistics, π and θw, are indispensable instruments in the detection and characterization of balancing selection within genomes.

1.3.2. HKA

The Hudson-Kreitman-Aguade (HKA) test is one of the basic methods in evolutionary genetics, introduced by Hudson and colleagues in 1987 (Stephen I. Wright, 2004).

It contrasts the level of genetic variation within species with the genetic divergence of interspecies. Neutral variants linked to a site under balancing selection are maintained within a population, leading to increased diversity around a target of interest. The Hudson–Kreitman–Aguade test can detect this pattern in specific genes or flag selection signatures in a gene group. In genomic surveys, in the absence of prior information, a π/d ratio, within-species diversity to between-species divergence, can be calculated for each locus. Loci that are significant outliers from the expected range or an appropriate null model may be inferred as under-selection (Thomas, 2012).

The neutral theory predicts a constant relation of polymorphism to divergence in genomic regions. However, balancing selection disrupts this balance by causing increased polymorphism in a species compared to divergence between species, especially in contrast to other parts of the genome. This high polymorphism would indicate that genetic variation in populations is being actively maintained (Cutter and Payseur, 2013).

Applying the HKA test, researchers look for significant polymorphism-to-divergence ratio variations across different loci. These variations are not merely random occurrences but rather can be indicative of evolutionary processes such as balancing selection affecting certain genome regions.

1.3.3. MK

In addition, nonsynonymous polymorphisms can effectively bring out balancing selection. One such important method in this regard is the test developed by McDonald and Kreitman in 1991, often referred to as the MK test (John H. McDonald, 1991). This test is insightful in that it analyzes the ratio of nonsynonymous to synonymous variations present in both polymorphism within species and divergence among species.

In the ideal neutrality scenario posited by the neutral theory of evolution, the ratios of nonsynonymous to synonymous changes should be similar for both polymorphisms and divergence. However, a more frequent occurrence of nonsynonymous polymorphisms within a species compared to synonymous ones may strongly suggest balancing selection. The reason for this is simple: balancing selection tends to maintain a wide variety of alleles in a population, including those that change amino acids.

Further refinement and validation of the use of the MK test in detecting balancing selection by recent studies (Vitti et al., 2013; Fijarczyk and Babik 2015) have deepened the understanding of the way the MK test can be applied to discrimination of subtle but significant signals of balancing selection in the genetic data.

1.3.4. LD

As noted, the detection of recent balancing selection is problematic because it generates patterns indistinguishable from those expected under positive selection. This results in heightened linkage disequilibrium near the selected locus and reduced population differentiation, which resembles what would be observed in cases of incomplete sweeps. In this respect, linkage disequilibrium (LD)-based methods can be utilized for detection. In such polymorphisms, LD will be found around the selective site, meaning that alleles in haplotypes are non-randomly associated, and haplotypes tend to cluster by allele type rather than by population or species. However, this signature is like that of an incomplete sweep and is, therefore, difficult to separate between the signal of balancing selection. Allele beneficial in fitness is increasing in frequency due to positive selection but has not yet become universal in the population and is considered under incomplete

sweep. The allele is therefore favored and under selection, but the process has not been fully completed (Hermisson and Pennings, 2005).

These phenomena often overlap in occurrence. For instance, a genetic variant may first experience a partial selective sweep because of an increase in the selective benefit it confers. Then, through changes in environmental or population factors, the same genetic variant may be subject to balancing selection thereafter, which maintains it at a particular frequency instead of letting it sweep. These dynamics will represent the complicated and fluctuating nature of evolution in the wild. It is convenient to set up two-time points to detect the hallmarks of very recent balancing selection, typically starting from 0.02–0.4 Ne (effective population size) generations ago (Anna Fijarczyk, 2015), which unfortunately resemble the hallmarks of recent positive selection, such as partial or soft selective sweeps (Hermisson & Pennings, 2005).

1.4. Balancing selection challenges

The challenges of finding balancing selection using these methods are not perfectly accurate when used alone. One way to find potential genes is by using tests that look at different genetic patterns expected under balancing selection and combining them (Ochola et al. 2010).

While Tajima's D and the HKA test are useful, they cannot distinguish selection effects from demographic and population structure influences that also produce similar polymorphism patterns. Thus, it is vital to construct and use a demographic model as a baseline for comparison (Quach et al. 2013). Andrés et al. (2009) exemplified this by analyzing 13,400 human genes with these tests, using inferred demographic history as a reference model. They identified genes deviating from this model as balancing selection candidates, a method noted for its low false-positive rate. Moreover, the effectiveness of these tests is further greatly impacted by high recombination rates, which localize selection signatures and complicate detection.

1.5. Machine Learning (ML) Approaches

The challenges may be addressed by a promising area from the side of supervised machine learning, which recently has been introduced to population genetics (Daniel R. Schreder, 2018). ML

algorithms self-optimize their parameters to improve predictive performance. In contrast to classical ML approaches that rely on summary statistics, deep-learning methods can automatically learn important data features relevant for prediction during the training process (Yann LeCun, 2015).

Recent developments also feature two likelihood-based approaches to identifying ancient balancing selection signatures employing simulations (DeGiorgio et al. 2014). The first focuses on the spatial distribution of polymorphisms and substitutions around a selective site, while the second analyzes allele frequencies flanking a polymorphic site. The effect of balancing selection on the genealogy surrounding a selected locus permits even finer-grained analysis. This precision is increased using a model of the genealogical process under balancing selection (Kaplan NL, 1988). Composite likelihood approaches are specifically useful in the analysis of genetic variation data with complex models of population genetics, providing estimates for complex models without the need to explicitly calculate full likelihoods.

1.6. Archaic Sequences

Detection of archaic genetic fragments within modern human DNA could be quite a challenging task due to the rarity of intact archaic DNA and the difficulty in distinguishing haplotypes that are distinctly different from already known archaic genomes. New statistical methods, including ArchIE, enable the detection of archaic DNA segments without comparison to known archaic DNA (Durvasula and Sankararaman, 2019). These methods allow for the extension of human evolutionary history insights (Sankararaman, 2020).

ArchIE, which is short for ARCHaic Introgression Explorer, uses genetic population statistics within a logistic regression framework to detect archaic DNA segments without directly comparing them to known archaic sequences. It represents the allele frequency using the individual frequency spectrum (IFS) then calculates the Euclidean distance between haplotypes while implementing statistics, such as mean, variance, and skewness/kurtosis. The main attribute used in this study is the minimum Euclidean distance between the target haplotype and any other haplotype in the reference group (measuring how likely it is to have experienced introgression). The model also considers the number of unique single nucleotide polymorphisms (SNP) in the haplotype, ignoring those shared with the reference group. Furthermore, it employs the S statistic,

obtained from the imbalance in LD, to identify introgressed haplotypes. This logistic regression approach has been quite effectively trained with these parameters, to discern between archaic and contemporary genetic fragments (Plagnol and Wall, 2006).

One telling example of the implementation of such a tool is the identification of Neanderthal gene variants of *LZTFL1* that have been associated with a higher risk of developing severe COVID-19. These variants have been found located at chr3:45,859,651–45,909,024, hg19 (Zeberg and Pääbo, 2020). These are examples of the influence that ancient genetic variation has on human health today and how selection is still having its effect on our response to modern diseases.

Briefly, ArchIE is one of the major breakthroughs in genomics, which delves into the mystery of human evolutionary history. Primarily, the tool has proved to be remarkable in its application of sophisticated statistical techniques to identify archaic DNA segments in modern human genomes. It underlines the innovation behind ArchIE and establishes it as a milestone in deepening our understanding of the intriguing play of genetic elements across various periods in human life.

2. MATERIALS AND METHODS

2.1. Materials

2.1.1. Subjects and Genotyping Data

Genotyping data is a part of the project ‘Genetic Diversity and Structural Changes in the Lithuanian Population Related to Evolution and Common Diseases’ (acronym LITGEN, project code VP1-3.1-ŠMM-07-K-01-013) carried out by the Faculty of Medicine at Vilnius University (A. Urnikyte, 2019). The research sample includes 424 individuals from the Lithuanian population, all with parents and grandparents who lived in Lithuania.

Genomic DNA extraction was performed from blood leukocytes using the phenol-chloroform method and the TECAN Freedom EVO automated DNA extraction system. The concentration of the extracted DNA was measured using the NanoDropR ND-1000 spectrophotometer. Large-scale genotyping of DNA samples for SNP markers was done using Illumina Infinium® HD and Illumina Infinium HTS technology chips. During the LITGEN project, the DNA of 295 individuals was genotyped with Illumina 770K HumanOmniExpress-12 v1.1 (719,666 SNPs) chips, and the DNA of 144 individuals was genotyped with Infinium OmniExpress-24v1-2 (713,599 SNPs) chips.

Genotyping was based on the Illumina Infinium® HD SNP genotyping protocol, as outlined in the Illumina Infinium® HD Assay Ultra user guide (Illumina, 2009). This research, part of the LITGEN project, was approved by the Vilnius Regional Research Ethics Committee (Approval No. 158200-05-329-79, dated May 3, 2011). Written informed consent was obtained from each participant according to the Declaration of Helsinki.

2.1.2. Data Preparation

Data preparation for analysis was conducted via SSH (secure shell) terminal (macOS, Darwin Kernel Version 23.3.0) on the ZMGKSERV server. Genetic data quality control was performed using the PLINK v1.90b6.21 64-bit software according to quality control metrics (Table 2.1.1.). Overall, these steps involve converting data formats, conducting quality control, removing duplicates, organizing data by chromosome, and phasing the data for further genetic analysis (SHAPEIT program v2.904.2.6, with architecture compatibility - x86-64).

Table 2.1.1. Data processing steps and related code lines. All sequential steps and cleaned data can be found in the following repository: https://github.com/Tsukinome/Files/tree/main

Preprocessing Step	Code Line	Description
Converting .ped and .map to PLINK Binary Format	plink --file data --make-bed --out data_binary	This step converts genotype data from the text-based .ped and .map formats to PLINK's binary formats (.bed, .bim, .fam). The binary format is more space-efficient and allows for faster processing.

Quality Control Steps with PLINK	a. Excludes SNPs and Individuals Based on Missingness Rates plink --bfile data_binary --geno 0.05 --make-bed --out data_snp_callrate_filtered plink --bfile data_snp_callrate_filtered --mind 0.1 --make-bed --out data_indiv_callrate_filtered b. Filters by Minor Allele Frequency (MAF) plink --bfile data_indiv_callrate_filtered --maf 0.01 --make-bed --out data_maf_filtered c. Excludes SNPs Deviating from Hardy-Weinberg Equilibrium (HWE) plink --bfile data_maf_filtered --hwe 1e-6 --make-bed --out data_hwe_filtered	a. Filters out SNPs with more than 5% missing data and individuals with more than 10% missing data. Elevated levels of missing data can indicate inferior quality and could bias the analysis. b. Excludes SNPs with a minor allele frequency below 1%. Rare alleles might not have enough statistical power for analysis and could lead to false positives. c. Removes SNPs that significantly deviate from Hardy-Weinberg equilibrium, as this may indicate genotyping errors or population stratification (Bowang Chen, 2017)

Remove Related Individuals and Duplicated Ids	plink --bfile data_hwe_filtered --remove --out plink_autosomes	Removed genotype data IDs: WA LTG-356, EA LTG-1401, WA LTG-333, EA LTG-813, WA LTG-1158, NZ LTG-1271, SZ LTG-875, SA LTG-781
Dividing by chromosomes	for chr_num in range(1, 23): run_command([plink, "--bfile", “plink_autosomes”, "--chr", str(chr_num), "--make-bed", "--out", “{file_prefix}_filtered_chr{chr_num}"])	-
Phasing with SHAPEIT	shapeit -P data_for_phasing.ped data_for_phasing.map -M genetic_map.txt -O data_phased shapeit -B .bed .bim .fam -O phased_chr1	Executes the phasing of genotype data using SHAPEIT, which determines the most likely combination of alleles inherited together on each chromosome. Phased data are essential for accurate haplotype analysis and imputation (Vivek Appadurai, 2023)

2.2. Methods

2.2.1. Data Augmentation

The SLiM 3.2 simulation program was used to generate training data (Benjamin C. Haller, 2019). SLiM is a powerful forward-in-time genetic simulation software that differs from backward-in-time simulations by focusing on phylogenetic trees. Instead, forward-in-time simulations model the genetic evolution of populations over time and provide a dynamic view of genetic changes.

Four scenarios were simulated, each representing one aspect of genetic evolution:

- Neutrality: genetic drift without any selection pressure (offers a guideline for the natural fluctuation of allele frequencies).
- Incomplete Sweep: a beneficial mutation that has not yet been fixed in the population, helping understand the initial stages of selection.
- Overdominance (locus under balancing selection): heterozygous individuals have a fitness advantage, a case in which genetic diversity is maintained at the locus.
- Negative Frequency-Dependent Selection (locus under balancing selection): fitness of a phenotype decreases when it gets common (hence increasing diversity in the population).

All simulations were performed with the following parameters: a mutation rate of 1.44e-8 per base pair per generation (Julien Jouganous, 2017), reflecting more or less realistic mutation rates observed in human populations; a generation time of 29 years (human generation interval) and a recombination rate from a normal distribution with a mean of 1e-8 and a standard deviation of 1e-9 (reflecting the natural variability of the recombination rate).

To simulate scenarios of natural selection, sequences with 50,000 base pairs (bp) were used, in the middle of which a specific mutation is located. This mutation was introduced into a simulated European population at 21 different time points ranging from 40,000 to 20,000 years ago. These time points were divided into three categories by period: recent selection (20,000 to 26,000 years ago), intermediate selection (27,000 to 33,000 years ago), and ancient selection (34,000 to 40,000 years ago). The simulation was restarted in the generation of the introduction of the variant when

and if the final frequency of the selected allele was not between 40% and 60%. This provided a method for recreating the impact of a variant under selection but included only those variants that had major evolutionary consequences.

No selection pressure was thus exerted in this case of neutrality. Data were instead generated with a neutral variant in the middle of the sequence, at a frequency lying between 40% and 60%. This kind of simulation framework is thorough, generating a strong dataset for training that captures a broad spectrum of evolutionary dynamics.

2.2.2. Artificial Neural Networks Model Training

To train an artificial neural network (ANN), the scikit-allele package is used to compute a set of potentially informative summary statistics for each simulation to represent the output sequence. The main statistics and their derivatives (median/mean/maximum) are shown in Table 2.2.1.

Table 2.2.1. Main range of statistics used for simulations.

Statistic	Description
Pairwise Distance	Genetic distance between pairs of sequences
Tajima's D	A neutrality test statistic comparing the number of segregating sites to the average number of nucleotide differences (Tajima, 1989)
Watterson's Theta	An estimate of genetic diversity based on the number of segregating sites (Sankar Subramanian, 2016)

Observed Heterozygosity	Heterozygosity observed in the population as well as ratio of observed and expected heterozygosity
r2	Linkage disequilibrium measure (Montgomery Slatkin, 2008)
Haplotype Diversity	A measure of the uniqueness of a particular haplotype in the population. As well as diversity statistics H1, H12, H123, H2/H1. Total count of different haplotypes in the sample.
EHH	Extended haplotype homozygosity (Alexander Klassmann, 2022)
iHS	Integrated haplotype score (Alexander Klassmann, 2022)
nSL	nSL statistic value (Anna Ferrer-Admetlla, 2014)
Kelly's Zns	A measure of linkage disequilibrium (Pleuni S. Pennings, 2006)
Fay and Wu's H	A test statistic to detect population growth or selection (KEi Zeng, 2006)
Number of Singletons	Count of unique variants appearing only once

Summary statistics were calculated using the scikit-allel library, which can be found at https://github.com/cggh/scikit-allel. These statistics were then scaled using the StandardScaler function from the sklearn library (https://scikit-learn.org/stable/).

A Python package called BaSe (Balancing Selection), available at https://github.com/ulasisik/balancing-selection, was used for selection detection. Apart from convolutional neural networks (CNN), BaSe uses artificial neural networks (ANN) to distinguish between incomplete sweep and balancing selection. All further analyses were performed in Python 3.10, using a Miniconda environment for package management and reproducibility and Jupyter Notebook for interactive data analysis and visualization.

ANN was created in Python using the Keras library (https://keras.io) (a high-level API (application programming interface) for building and training neural networks). Keras is based on TensorFlow (tensorflow.org) (ML platform developed by Google that serves as a backend for the actual calculations). All tools used are open-sourced.

The ANN model consists of several layers (Wesam Salah Alaloul, 2018) (Figure 2.2.1.):

- Input layer: the first layer in a neural network where the data is fed into the model. The number of neurons in this layer corresponds to the number of input features (variables) in a data set.
- Hidden layer: layers termed hidden because they have no direct contact with inputs or outputs. The many neurons in each of the hidden layers make up the processing of the input data to help the model learn complex patterns.
- Output layer: the last layer in the network provides the output of the model. If the task is binary classification, the output layer consists of one neuron. It consists of a neuron with a sigmoidal, logarithmic, or logistic activation function that outputs a continuous value between 0 and 1, which may be interpreted as the probability that a given input belongs to a certain class.

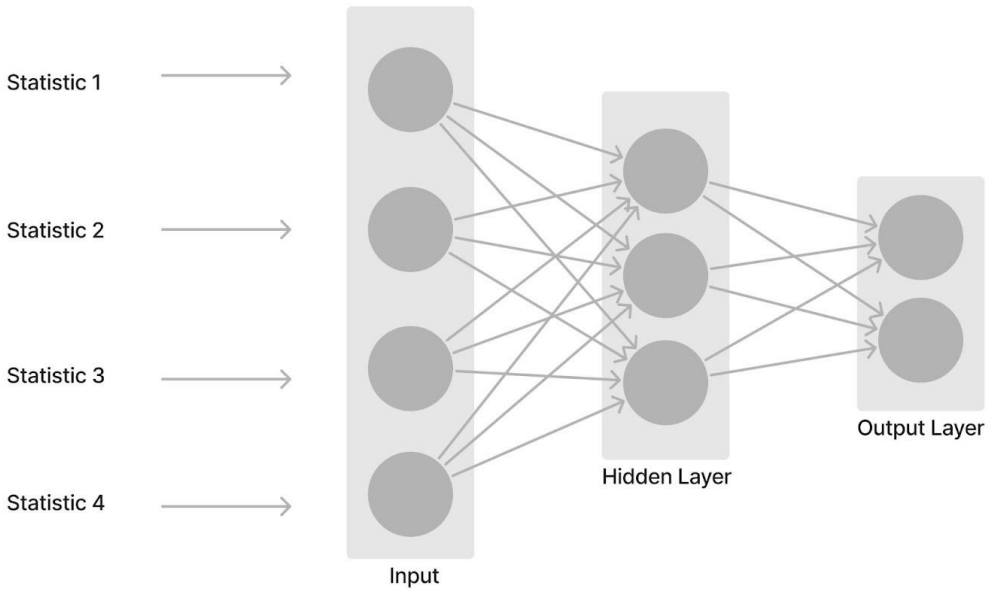

Figure 2.2.1. Artificial neural network model layers

The network adjusts its internal parameters (weights and biases) during training to minimize the error in its predictions and thus improve its ability to correctly classify new, unseen data. This process is usually performed using optimization algorithms such as stochastic gradient descent (SGD), which iteratively updates the parameters to reduce the loss function, a measure of the discrepancy between the predicted results and the actual targets. By constantly fine-tuning these parameters, the network improves its generalization capabilities so that it can do well with various samples of data different from the ones in the training set. Effective training lies at the heart of the creation of robust neural networks that are able to adapt to different input conditions and yield accurate results in real-world applications (Goodfellow et al., 2016).

2.2.3. Model Predictions

The models are categorized by selection category to indicate when the selection started - new, intermediate, or ancient. In addition, each model is given a test number from 1 to 3, depending on

what it was trained for: discrimination between neutrality and selection (including positive and balancing selection), balancing selection and incomplete sweep, or negative frequency-dependent selection and overdominance. All models are stored in Keras h5 format.

Each test performs a binary classification. The first test checks whether the target allele is under selection. If it is, the second test checks whether it is a balancing selection or an incomplete sweep. If it is a balancing selection, the third test distinguishes between the several types of balancing selection: overdominance and negative frequency-dependent selection.

2.2.4. Archaic Fragments Mapping

The references to the archaic regions come from the latest research results on Neanderthal alleles in a set of 50 Lithuanian genomes (Alina Urnikyte, 2023). The archaic region data in this study was found using the Archaic Introgression Explorer (ArchIE) program. Genotype data from 108 individuals of African ancestry from the Yoruba population were used for this study obtained from the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015). The final output included the chromosome number, segment, region of interest, and a probability score for Neanderthal introgression between 0 and 1.

For further analysis, only those genomic fragments were considered to be of Vindija Neanderthal origin. Introgressive sequences with a match of 85% or more with the Vindija Neandertal genome were used. Matches with the Denisovan genome were less significant and were not analyzed because the main source of introgression in Lithuanians is the Neanderthal genome.

To understand the association between SNPs and genes, each SNP was mapped via Ensembl. The function get_gene_and_consiquence_for_snp (available at https://github.com/Tsukinome/Files/blob/main/BSModel/CSVs/Prediction.ipynb), requires two inputs: snp_id - the identifier of a SNP, and chr_column - chromosome number.

The process starts with the creation of a uniform resource locator (URL) to query the Ensembl REST API for information about the variations of the SNP. If the request is successful, it gets the JavaScript object notation (JSON) string response from the function. After that, if the response contains a mapping key, that means it includes the genomic locations for that SNP. It constructs

another URL for each mapping - one can also think about it in terms of a genomic location - to query overlapping genes using the Ensembl REST API. If this is a success, it collects all the gene IDs. The output of the function would be a list of unique gene IDs for the given SNP. Example response for available features is shown as below:


```
{
"evidence":

["Frequency", "1000Genomes", "Cited", "TOPMed", "gnomAD"],
"most_severe_consequence": "intron_variant",
"name": "rs123",
"MAF": 0.292133,
"mappings": [
{
        "end": 24966446,
         "coord_system": "chromosome",
         "ancestral_allele": "C",
         "assembly_name": "GRCh37",
         "allele_string": "C/A/G/T",
         "seq_region_name": "7",
         "location": "7:24966446-24966446",
        "strand": 1,
         "start": 24966446
}
         ],
"ambiguity": "N",
"source": "Variants (including SNPs and indels) imported from dbSNP",
"synonyms": [
         "NM_001177519.3:c.*235C>A",
         "rs17614680",
         "rs57332242"
         ],
"minor_allele": "C",
```

"var_class": "SNP"
}

2.2.5. Protein Functionality Analysis

The data were further processed using the GRCh37 (hg19) genome build and RefSeqGene. GRCh37 is a well-established reference build of the human genome, and RefSeqGene provides comprehensive and curated gene information; both are necessary for the precision of genomic studies. In support, a function get_protein_info_for_gene was designed to fetch protein information from a gene. It retrieves information about proteins related to the meaning of the gene. It works in the following way: the input to this function is gene_id. The process is initiated by constructing a URL in order to query Ensembl REST API for information about the gene. If there is no error in the request, the function retrieves the JSON response. From there, it picks up the display name, which is the protein name, in addition to the description of the gene. In case the protein name or description is not found, it defaults to the 'Unknown' and 'No description available' categories. So, this function output is a list that includes the protein name and its description. The following is an example of a response:

{
"species": "homo_sapiens",
"start": 32889611,
"db_type": "core",
"end": 32973805,
"seq_region_name": "13",
"object_type": "Gene",
"logic_name": "ensembl_havana_gene_homo_sapiens_37",
"display_name": "BRCA2",
"id": "ENSG00000139618",
"source": "ensembl_havana",
"description": "breast cancer 2, early onset [Source:HGNC Symbol;Acc:1101]",
"canonical_transcript": "ENST00000544455.1",
"version": 10,

"strand": 1,
"biotype": "protein_coding",
"assembly_name": "GRCh37"
}

This function provides a straightforward way of obtaining relevant protein information, which is essential for gene analysis, functional annotation, and understanding the potential implications of genetic variations.

RESULTS

3.1 Genotype data analysis

Initially, genotype data was imported from text-based formats, and 733,133 (Table 3.1.1.) variants and 424 individuals were loaded onto the platform, with 212 males and 212 females. Related persons, as well as duplicated IDs, were removed from the dataset to avoid bias due to genetic relatedness; the result was a final count of 415 individuals, 208 males and 207 females.

After filtering, 600,627 variants remained, and 415 individuals passed the quality control steps. The final genotyping rate was 99.87%, which is significantly higher than the original 96.81%. This high genotyping rate suggests that about 99.87% of the genotyping attempts were successful and hence guaranteed quality data for further analyses.

These preprocessing steps ensure that the dataset being used for further genetic analysis is of excellent quality, hence minimizing errors and biases.

Table 3.1.1. Initial and final population statistics.

Preprocessing Step	**Removed Data**

Initial Population	733,133 variants loaded from .bim file 424 people (212 males, 212 females) loaded from .fam file
Variants Removed during Quality Control Steps	38,249 variants removed due to missing genotype data 80,032 variants removed due to minor allele threshold(s) 8,196 variants removed due to Hardy-Weinberg exact test
Remove Related Individuals and Duplicated Ids	Final individual count: 415 people (208 males, 207 females) 2 duplicates and 7 related individuals removed
Final Population	600,627 variants and 415 people pass filters and Quality Control Total genotyping rate is 0.998718 (before data cleaning: 0.968083)

3.2. Selection Type Distribution and Probability

Strongest selection signals (mixture of both balancing and positive selection scenarios) were observed on chromosomes 6, 13, 15, and 22, indicating their possible role in genetic adaptation (Figure 3.2.1/2/3.). These chromosomes account for a total of 5,311 SNPs, with an average selection probability of 0.998 (Figure 3.2.4/5/6., zero values indicate lack of selection type pattern found).

Significant patterns of recent balancing, frequency-dependent selection and overdominance are notable on chromosomes 3, 6, 13, 15, 16, 18, 19, 21, and 22. For instance, recent balancing

selection on these chromosomes comprises 1,544 SNPs with an average probability of 0.610. Similarly, recent negative frequency-dependent selection is prominent on mentioned chromosomes 6, 13, 15, and 22, encompassing 5,233 SNPs with an average probability of 0.557. Recent overdominance is prevalent across 1-20 chromosomes (>1,000 SNPs per chromosome), totaling 93,292 SNPs (Supplementary Table 1).

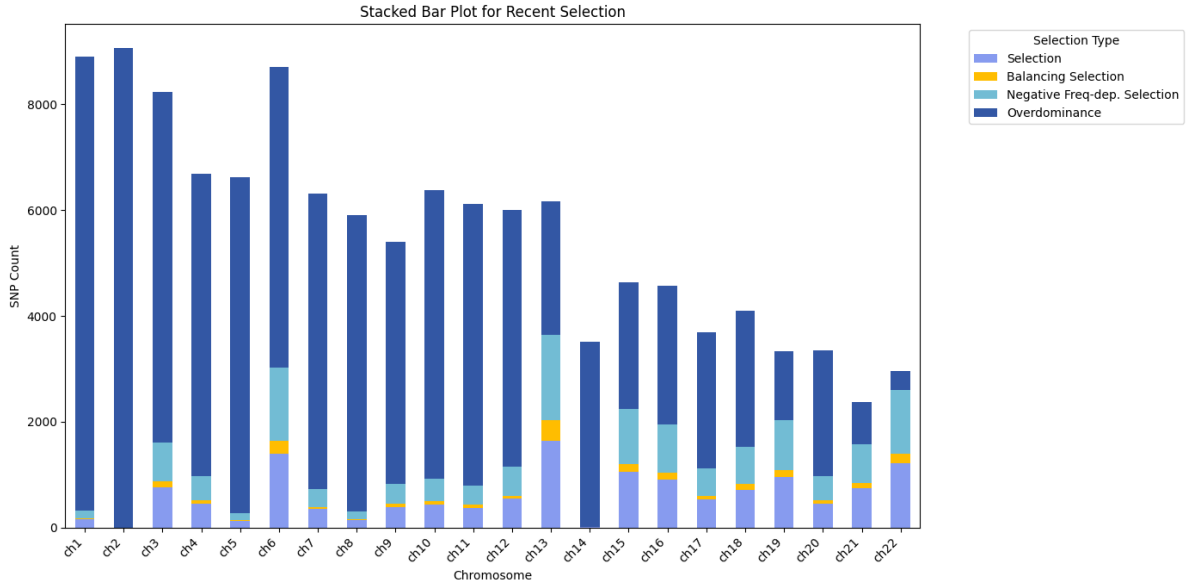

Figure 3.2.1. Recent selection. SNP count distribution by chromosome and selection type. Selection class includes a mixture of both balancing and positive selection scenarios.

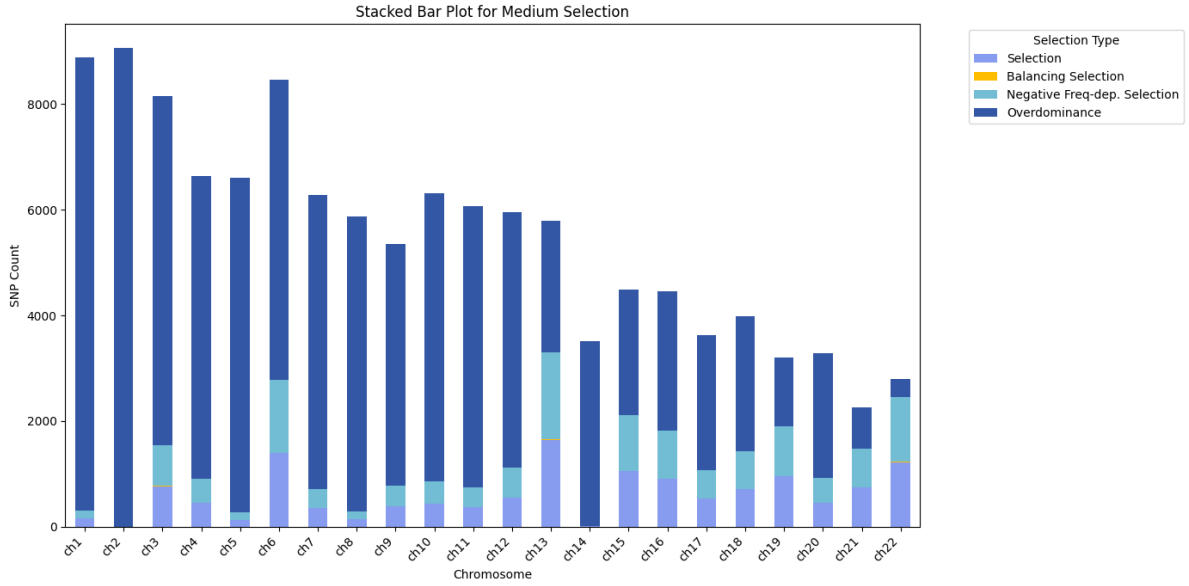

Figure 3.2.2. Intermediate selection. SNP count distribution by chromosome and selection type. Selection class includes a mixture of both balancing and positive selection scenarios.

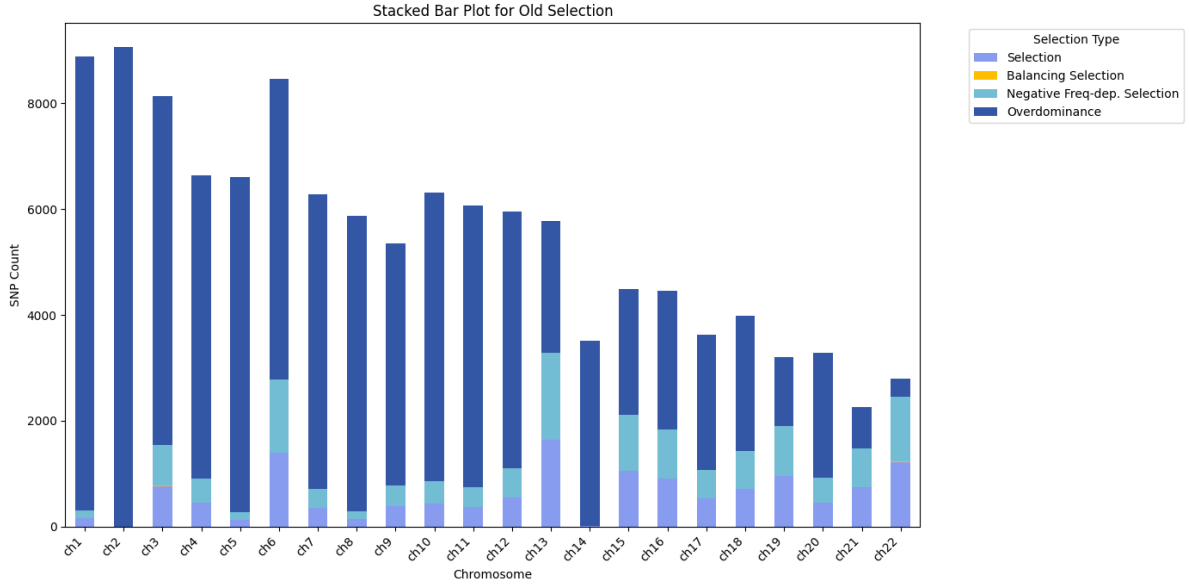

Figure 3.2.3. Ancient selection. SNP count distribution by chromosome and selection type. Selection class includes a mixture of both balancing and positive selection scenarios.

The distinction between balancing selection and incomplete sweep was the most prominent and the most accurate during recent selection, hence only this pattern is explored further.

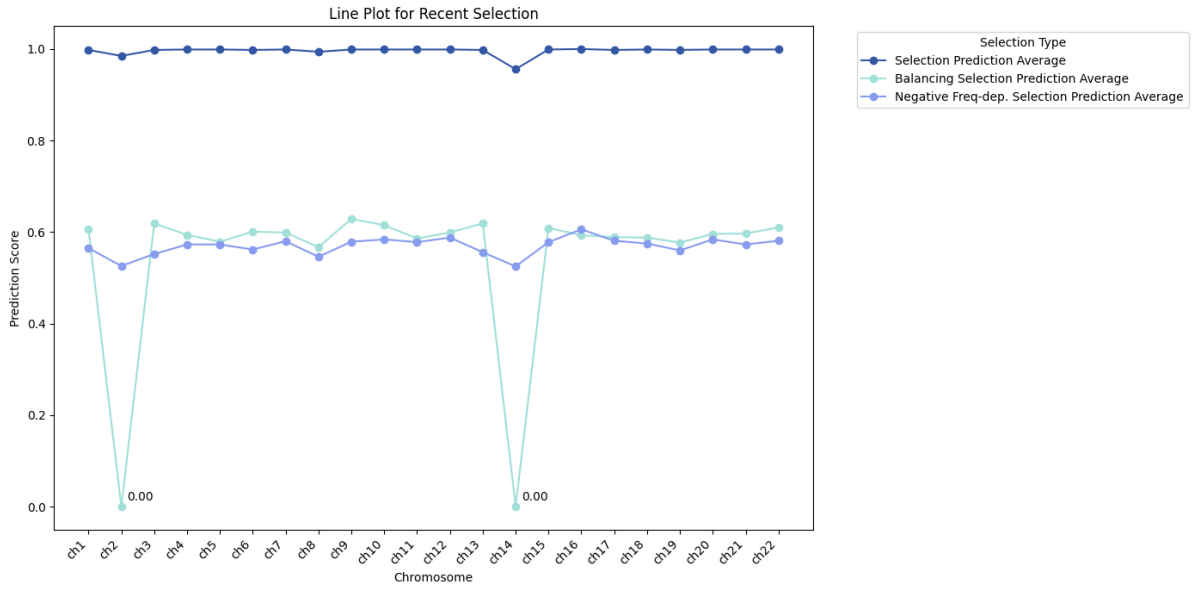

Figure 3.2.4. Recent selection. Selection probability average by selection type and chromosome.

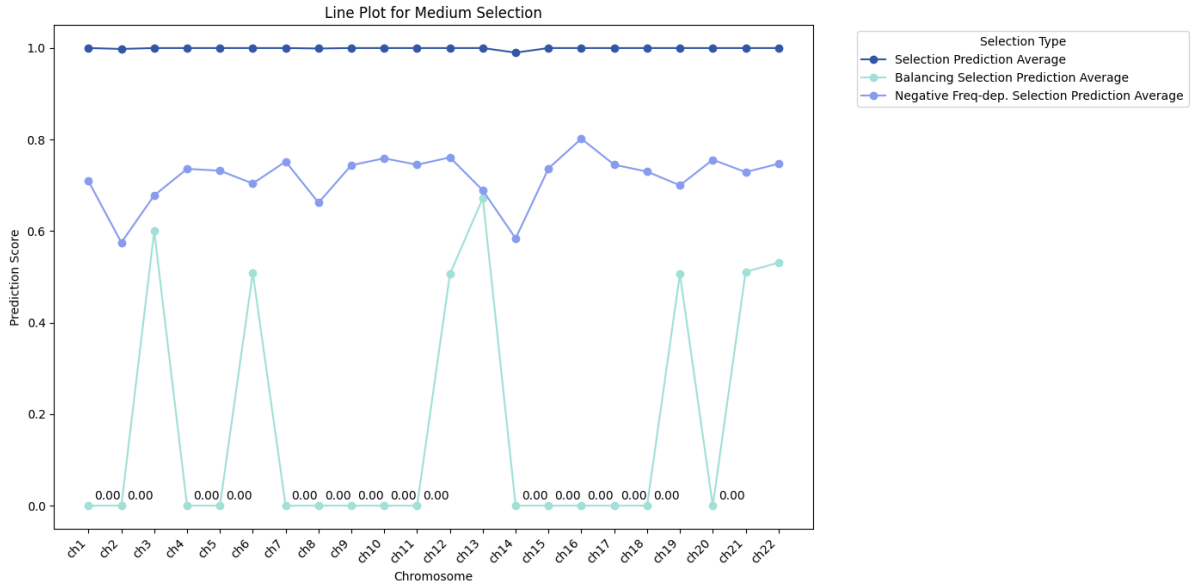

Figure 3.2.5. Intermediate selection. Selection probability average by selection type and chromosome.

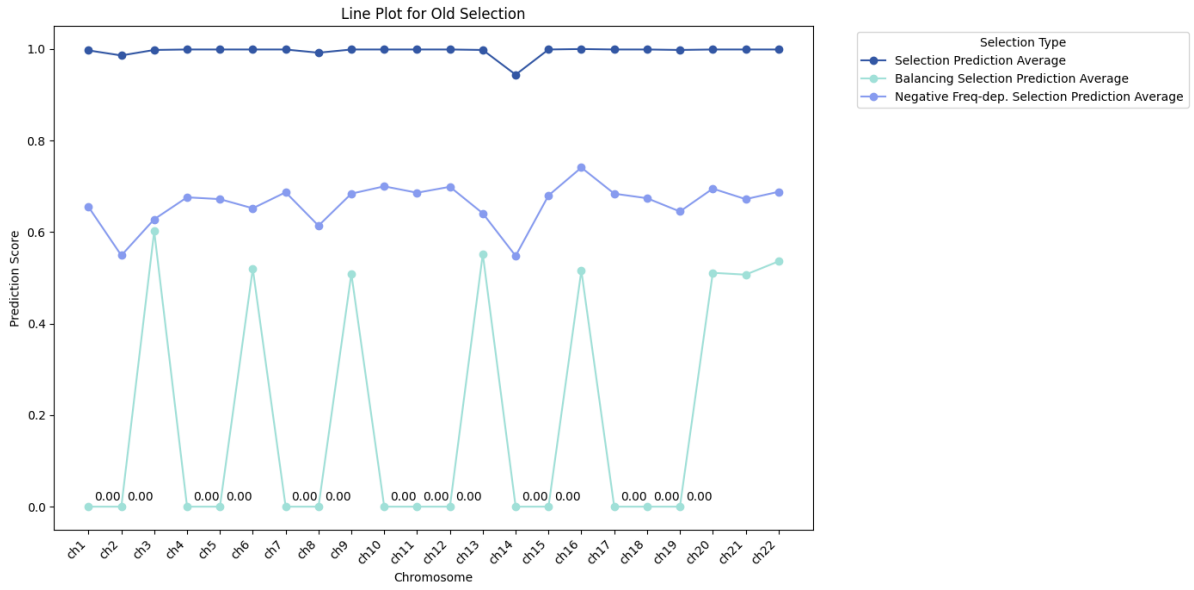

Figure 3.2.6. Ancient selection. Selection probability average by selection type and chromosome.

Additionally, balancing selection, negative frequency-dependent, and overdominance variants are chosen only if they were categorized as a selection class in the first prediction test (distinguishing between neutrality and balancing/positive selection scenarios).

3.3. Related Protein Functionally

Out of 1598 archaic fragments, 175 were determined to contain balancing selection-related variants. One hundred forty-five fragments were filtered out having <85% probability of having Neanderthal introgression. In total, 53 variants were selected to be processed further.

Examining further these genetic variants (SNPs) across various chromosomes (both predicted to be under selection and balancing selection), it was found that chromosome 3 features 10 SNPs, Chromosome 13 includes 9 SNPs, and Chromosome 22 has 11 SNPs, making them the most SNP-rich chromosomes under balancing selection in this dataset (Table 3.3.1. and Supplementary Table 2). Most of these SNPs lead to changes in intron variants that most often do not affect protein function. Several regulatory and missense variants were observed on Chromosomes 3, 6, 13, and

22. Such specific variants could play important roles in the regulation of genes and proteins and, therefore, may have strong implications for gene expression and protein activity (Table 3.3.2/3.).

The only missense variant was found to be rs5764698. In the reference genome, the nucleotide at this position is G (guanine). In some individuals, however, this position may also have other nucleotides, namely C (cytosine) or T (thymine). The SNP at position chr22:45354103 in the *SMC1B* gene results in a missense mutation that causes a change in the amino acid at position 1050 in the protein, resulting in leucine (L) becoming either valine (V) for the C allele or methionine (M) for the T allele.

Overall, proteins predicted to be under balancing selection are categorized into distinct functional groups. Among the signaling and regulatory proteins are interleukin 1 receptor accessory protein, PARK2 co-regulated, extracellular leucine-rich repeat and fibronectin type III domain containing 2, and family with sequence similarity 118, member A. The metabolic and transport proteins include adenylate kinase 1, tripeptidyl peptidase II, and solute carrier family 39, member 11. Structural maintenance of chromosomes 1B is identified as a structural and maintenance protein. Several hypothetical and novel proteins with unknown functions, such as chromosome 6 open reading frame 99, must be mentioned as well.

Table 3.3.1. SNPs intron variants and related protein functional aspects

SNP	Position	Gene	Biological Function
rs2193880	chr3:190262799	*IL1RAP* (ENSG00000196083)	IL-1 Receptor Accessory Protein
rs3773989	chr3:190547640		
rs9457507	chr6:159316430		

rs9457511	chr6:159325462	*LINC02901* (ENSG00000203711)	Chromosome 6 Open Reading Frame 99
rs12200537	chr6:159326830		
rs9356058	chr6:163151399	*PACRG* (ENSG00000112530)	Parkin Co-Regulated Gene Protein
rs3780663	chr9:130631519	Novel Gene (ENSG00000257524), *AK1* (ENSG00000106992)	ST6-N-Acetylgalactosaminide Alpha-26-Sialyltransferase 6 Isoform 2, Epididymis Secretory Sperm Binding Protein
rs12251249	chr10:127720923	*ADAM12* (ENSG00000148848)	ADAM Metallopeptidase Domain 12
rs3736972	chr13:103275386	*TPP2* (ENSG00000134900)	Tripeptidyl Peptidase II

rs6496435	chr15:88214711	Novel Gene (ENSG00000259560)	Novel
rs12905479	chr15:91711209	*SV2B* (ENSG00000185518)	Synaptic Vesicle Glycoprotein 2B
rs17594552	chr15:91712789		
rs1110614	17:70750674	*SLC39A11* (ENSG00000133195)	Zinc Transporter ZIP11
rs4793484	17:70752633		
rs9906450	17:70763732		
rs2334295	19:48263635	*NOP53-AS1* (ENSG00000269656)	NOP53 Antisense RNA 1
rs2223743	21:29503271	*LINC01695* (ENSG00000236532)	Long Intergenic Non-Protein Coding RNA
rs6516819	21:29510893		
rs9976944	21:29512586		

rs2831534	21:29514609		
rs4820286	22:37781591	*ELFN2* (ENSG00000166897) , *ELFN2* (ENSG00000243902)	Extracellular Leucine Rich Repeat and Fibronectin Type III Domain Containing 2
rs5750428	22:37794269		
rs11704481	22:45732328	*FAM118A* (ENSG00000100376)	Family With Sequence Similarity 118 Member A
rs5771906	22:49054594	*TAFA5* (ENSG00000219438)	TAFA Chemokine Like Family Member 5
rs6010568	22:49056481		

Table 3.3.2. SNPs 3 prime UTR or missense variants and related protein functional aspects

SNP	Position	Gene	Biological Function
rs2064068	22:45736485	*FAM118A* (ENSG00000100376)	Family With Sequence Similarity 118 Member A
rs1044742	22:45737290		
rs5764698	22:45749983	*SMC1B* (ENSG00000077935)	Mitosis-Specific Chromosome Segregation Protein Like Protein Bet

Because these proteins were subject to balancing selection, particularly negative frequency-dependent selection, they exhibit genetic diversity that could provide significant advantages.

One notable case concerns the SNP rs9356058 in the *PACRG* gene. Previous research has identified rs9356058 as one of two important regulatory polymorphisms in the *PARK2* and *PACRG* genes associated with susceptibility to leprosy (A.Bakija-Konsou, 2011). In this study, Caucasian individuals from Mljet, an island formerly used as a leprosy quarantine, were compared with two control groups. The population of Mljet showed a significant increase in the frequency of the rs9356058 C allele compared to the control groups. However, the allele frequencies of the analyzed polymorphisms did not differ between the control groups. These results suggest that leprosy exposure and mortality on Mljet led to the selection of the protective C allele rs9356058 in the *PARK2* gene, a pattern that is absent in the control groups, which have no documented leprosy history. This balancing selection on rs9356058 suggests that the retention of both the C and T alleles might be beneficial for a better resistance to leprosy or other pathogens. Therefore, the protective effect against leprosy has led to the prevalence of the C allele on the island of Mljet. In contrast, in Lithuania, various influences have maintained a balance between the two alleles. The

allele frequency variations indicate that Lithuanians have an intermediate C allele frequency (0.2674), which is lower than the European average (0.417) but higher than that in Africans (0.161) and South Asians (0.172), positioning it as intermediate globally. This indicates the existence of balancing selection, whereby both T and C alleles give evolutionary advantages under different conditions, thus retaining genetic variability. Such variability can be advantageous against a variety of diseases and environmental factors and can highlight the Lithuanian unique evolutionary pressures.

Additionally, the identified *PARK2* gene, also known as parkin, is associated with autosomal recessive juvenile parkinsonism (Marcelo T. Mira, 2004). Mutations in this gene cause Parkinson's disease and autosomal recessive juvenile Parkinson's disease. In addition, both *PARK2* and *PACRG* are expressed in Schwann cells and macrophages, the primary host cells for *Mycobacterium leprae*, the bacterium that causes leprosy (Louis de Leseleuc, 2013). The same genetic changes that conferred protection against leprosy may also have survival effects relevant to vulnerability to Parkinson's disease. The fact that these genes, by their association with expression in the same cell types, namely Schwann cells and macrophages, suggest that there must be a shared pathway operating under balancing selection. It will be important to learn more about these genetic interplays, which may offer new inroads into mechanisms operating both in infectious and neurodegenerative diseases.

Other promising SNPs were found to be related to IL-1 Receptor Accessory Protein: rs2193880 (chr3:190262799) and rs3773989 (chr3:190547640). This protein plays a key role in the signaling pathways of various proinflammatory cytokines (Jame Frenay, 2022). Polymorphic forms of this gene have been linked to a wide range of inflammatory diseases. Various polymorphisms of the IL-1RAcP contribute to the pathophysiology of various diseases, including obesity, endometriosis, pre-eclampsia, and neurodegenerative diseases (Ali Zarezadeh, 2022). Significantly, the rs12053868-G polymorphism of the IL-1RAcP gene has been linked to Alzheimer's disease. Investigations show that this polymorphism leads to increased accumulation of amyloid peptides, which in turn decreases the activity of cortical microglial cells essential for the clearance of amyloid fibrils from the brain. Such activity can trigger cerebral atrophy, most especially in the temporal cortex, which contributes to accelerated disease progression and memory loss in Alzheimer's patients. Due to the elevated levels of expression of IL-1RAcP in inflammatory diseases compared to its low expression in healthy tissues, these polymorphisms may modify

susceptibility or resistance to a diverse array of inflammatory diseases. Therefore, IL-1RAcP and the associated polymorphisms found, rs2193880 and rs3773989, may have adaptive utility by moderating immune responses to balance effective defense against pathogens and prevention of excessive inflammation leading to chronic disease (Khaled Khazim, 2018). This is true of neuroinflammatory diseases, such as Alzheimer's disease, where neurodegenerative and inflammatory mechanisms play a critical role (Jose Miguel Rubio-Perze, 2012).

In total 7 genes were found to be related to the immune system or inflammatory responses:

- *IL1RAP*, there is evidence for the significant role in airway inflammation driven by IL-1 and IL-33, encouraging further research to explore the benefits of blocking the IL1RAP coreceptor in chronic respiratory diseases (G. Kasetty, 2022).
- *ADAM12* is a naive T cell costimulatory molecule that mimics CD28 signaling to activate Th1 cells and induce IFNγ production. Genetic or knockout ablation of *ADAM12* decreased the activity of Th1 cells, reduced production of IFNγ, and attenuated Th1-mediated neuroinflammation in models of multiple sclerosis. These findings position *ADAM12* as a potential therapeutic target for the treatment of Th1 cell-mediated inflammatory diseases (Yawei Liu, 2020).
- *SLC39A11* is associated with chronic gastritis in the Korean population and encodes a zinc transporter, suggesting involvement in zinc homeostasis. Although not directly associated with inflammatory responses, chronic gastritis involves inflammation of the lining in the stomach, indicating that *SLC39A11* may be involved in influencing inflammation through gastritis progression (Eunyoung Ha, 2018).
- *PACRG*, regarding this variant, additional related studies were done with Chinese (Jinghui Li, 2012) and Amazon ethnic admixed population (Andre Luiz Leturiondo, 2020).
- *TPP2*, its deficiency causes immunodeficiency and immune dysregulation due to disrupted protein catabolism, favoring lysosomal pathways. This results in lysosome accumulation, reduced glycolysis, and impaired cytokine production (Clare Stockdale, 2021).
- *ELFN2*, genes co-expressed with *ELFN2*, are involved in the PI3K-Akt signaling pathway. This observation has indicated that *ELFN2* and its associated genes play key

roles in regulating such a pathway, which is crucial for cellular processes including growth, cell survival, metabolism, and inflammation (Ying Dong, 2022)

- *TAFA5* is upregulated in gastric cancer (GC) and linked to poor differentiation, advanced stages, and worse patient prognosis. Its downregulation inhibits GC cell proliferation and migration. *TAFA5* is associated with genes involved in epithelial-mesenchymal transition, making it a potential therapeutic target for GC (Zhiqing Hu, 2019)

Apart from SNPs related to neurodegenerative and inflammatory processes, one SNP rs5764698, which is the only missense variant (in the *SMC1B* gene), was found to be related to azoospermia and severe/moderate oligozoospermia (Kenneth I. Aston, 2010). The *SMC1B* gene is crucial for spermatogenesis and plays a significant role in the correct segregation of chromosomes during meiosis. This genotyping study included patient and control samples from individuals of European ancestry and Mediterranean origin.

The finding that rs5764698 is subject to balancing selection in Lithuanians suggests that this variant, despite its association with azoospermia and oligozoospermia, provides some adaptive advantages and may optimize reproductive success under certain environmental or physiological conditions. The allele frequency of rs5764698 varies significantly across populations. The G allele is more common across the globe than the T allele. Lithuanians have a higher G allele frequency and a lower T allele frequency than average and the other Europeans. Conversely, Africans have a much higher G allele, while Asians and Latin Americans have intermediate values.

Such balancing selection could be accounted for by a trade-off in which the variant provides some advantages - fertility or other reproductive-related - that are of sufficient magnitude to maintain the allele in the population despite negative effects on sperm production in some individuals. Knowing what specific advantages are conferred by the variant will shed deeper light on evolutionary forces acting upon genes responsible for spermatogenesis and the eventual consequences for reproductive health.

Additionally, variation in the regulatory proteins with *ELFN2* and *FAM118A* could affect cell signaling and growth, where different alleles convey differential advantages under varying physiological or environmental contexts. Lastly, variations in metabolic and transport proteins -

for example, *AK1*, *TPP2*, and *SLC39A11* - could be allowed to vary in accord with the different nutritional and metabolic conditions across populations and, further, give importance to these proteins in adapting to different challenges.

DISCUSSION

Selection events, particularly pronounced during the outbreak of infectious diseases, leave clear markers in the human genome. Deep learning techniques have proven to be powerful tools for detecting genomic signals for recently acted balancing selection (Ulas Isildak, 2020). By training forward-in-time simulation-based deep neural networks for population genetics using data augmentation, such methods have shown promise.

This work demonstrates how deep learning can predict things that are currently impossible to predict using traditional methods for small populations based on summary statistics. The accuracy of these predictions can be drastically improved by expanding the size of the training dataset, by doing a more comprehensive search of the hyperparameters while avoiding overfitting of the model, and by treating overdominance and negative frequency-dependent selection as different classes for the prediction.

The same applies to the relevance of individual statistics to be evaluated. To measure the false positive rate, it is also important to test ANN in neutral scenarios. This step ensures that the networks within these neutral control regions do not predict false selection signals. With this validation in place, it will be possible to enhance the reliability of the deep learning models and further refine their accuracy in identifying true signals for balancing selection.

Presently, the recent compensatory selection is best represented in the population genomic data available now for Lithuanians, distinguished with greater accuracy from the incomplete sweep. In addition, the variants selected for analysis show negative frequency-dependent selection as the most generic form. This means that rarer phenotypes have a selection advantage simply because they are less common. As a result, negative frequency-dependent selection promotes and maintains genetic diversity within the population.

Variants found to be under balancing selection display a few distinct categories relating to genes involved in neurodegenerative, inflammatory, or reproductive processes. Inflammatory response and immune function are common themes regarding Neanderthal introgressed fragments identified by previous studies (A. Urnikyte, 2023). Particularly with genes like *REG3G*, *IL17A* and *IL17F* compared to the IL-1 Receptor Accessory Protein of this study, as well as with CDH1, involved in cell adhesion and immune response regulation, compared to TAFA Chemokine Like Family Member 5.

Despite the inflammatory response, balancing selection could also affect reproductive health, as shown by the SNP rs5764698 in the *SMC1B* gene, which is associated with azoospermia and oligozoospermia (Kenneth I. Aston, 2010). This variant may be widespread in the population despite an association with negative reproductive outcomes.

In all, the persistence of those alleles in such genes may indicate that each of them gives some advantages under different environmental conditions or selection pressures. For instance, in the case of the C allele rs9356058, this may confer protection against leprosy, with the T allele conferring resistance to other diseases or having adaptive benefits that are presently unknown.

More precisely, the *PARK2* gene associated with autosomal recessive juvenile parkinsonism and *PACRG* are expressed in the main host cells for *Mycobacterium leprae*: the Schwann cells and macrophages (Louis de Leseleuc, 2013). This suggests a common pathway under balancing selection. Genetic variants that confer protection against leprosy may, in turn, increase vulnerability to Parkinson's disease; thus, the trade-off would be in that direction, such that alleles that are beneficial in one context may have a negative effect on another. This could be an example of how balancing selection maintains alleles that bring an advantage for resistance to infectious diseases, but that may make individuals more prone to neurodegenerative diseases.

The fact that these genes and their expression in immune cells are linked allow to hypothesize that balancing selection causes adaptations that maximize the immune response. In other words, populations exposed to the broadest spectrum of pathogens in time may evolve a genetic architecture that balances the traits of multiple immune-related characteristics, ensuring a versatile and effective immune defense. The investigation of these interactions may lead to new interpretations of the mechanisms of infectious and neurodegenerative diseases and may shed light

on the complex interrelations of genetic diversity with susceptibility to diseases in human populations.

This work, analyzing genetic adaptation within local populations, thus sheds more light on the evolutionary mechanisms underlying human health and disease. Looking at these different selection pressures that each population faces, the genetic basis of adaptation will be apparent and deepened in our understanding of human biology. Balanced selection maintains not only the genetic diversity of a population but also contributes to molding our physiology and shaping our responses to a variety of health challenges. Integrative deep learning applied to genetic investigation can better identify and understand these selection markers, which hold the clue to the advancement of medical science using evolutionary history to improve health.

CONCLUSIONS

1.1. Successful identification of balancing selection variants depends on the preparation of the genomic data according to appropriate quality control metrics, removing related individuals and non-biallelic variants.

1.2. One hundred seventy-five archaic fragments were determined to contain balancing selection-related SNP variants.

2.1. The majority of found SNPs result in intron variants, which generally are not known to impact protein function but could impact important regulatory aspects. Often, variants under selection are found in non-coding fragments. Several regulatory and missense variants on chromosomes 3, 6, 13, and 22 were found. The only missense variant was found to be rs5764698.

2.3. Strongest selection signals are observed on chromosomes 3, 6, 13, 15, 16, 18, and 22, indicating their potential role in genetic adaptation.

3.1. Proteins predicted to be under balancing selection, categorized into a few distinct functional groups such as signaling and regulatory, metabolic and transport, structural and maintenance protein, as well as several hypothetical and novel proteins.

3.2. SNP rs9356058 in the *PACRG* gene is one of the two important regulatory polymorphisms associated with susceptibility to leprosy.

3.3. In total, 7 genes were related to the immune system or inflammatory responses. Two of which (*PACRG* and *ADAM12*) are related to neurodegenerative disorders and two (*TAFA5* and *SLC39A11*) with gastro-related disorders.

VILNIUS UNIVERSITY

DEPARTMENT OF BIOMEDICAL SCIENCES

Kristina Grigaitytė

Inferring Archaic Alleles: Assessing Their Impact on Balancing Selection

Master thesis
Vilnius University, Institute of Biotechnology

SUMMARY

Balancing selection continues to be a captivating target because it sheds more light on how genetic diversity is maintained within populations despite external evolutionary pressures. Aspects of

balancing selection are crucial for understanding how populations adapt over time to different environmental conditions, disease pressure, and other selective forces.

In the context of the Lithuanian population, the study of balancing selection helps to clarify why certain alleles remain predominant in other populations despite the possible advantages of alternative alleles. This may reveal underlying mechanisms of disease resistance, metabolic adaptations, or other fitness-related traits that are of historical and contemporary importance. In addition, the identification of archaic SNPs that are subject to balancing selection is important as they highlight the contributions of ancient alleles to modern genetic diversity and adaptation.

Objective of the research was to identify ancient balancing selection signatures across generations in modern human genomes, employing genetic population statistics and deep learning techniques, while identifying which variants or proteins potentially indicate that balancing selection occurred in ancient times but could continue to influence human species today.

Deep learning models displayed great capability of effectively predicting loci under selection for a large pool of samples and by different selection types, namely between neutrality and selection (including positive and balancing selection), balancing selection and incomplete sweep, or negative frequency-dependent selection and overdominance. Overall, out of analyzed SNPs in 175 archaic fragments, the strongest selection signals are pronounced on chromosomes 3, 6, 13, 15, 16, 18, and 22, indicating their possible role in genetic adaptation. SNP-related genes are found to be involved in such processes as signaling, metabolism, transport, structural and maintenance functions.

Although the most prominent pathways of influence were found to be related to the immune system or inflammatory responses, few of them are also related to neurodegenerative disorders (*PACRG* and *ADAM12*) and (*TAFA5* and *SLC39A11*) with gastro-related disorders. Additionally, one of the most previously researched and explored in different populations, SNP rs9356058 in the *PACRG* gene, is one of the two important regulatory polymorphisms associated with susceptibility to leprosy. Understanding the role of rs9356058 and similar polymorphisms helps clarify the genetic factors that influence disease susceptibility and resistance, providing insight into how historical pathogen exposure has shaped the immune systems of modern humans.

LITERATURE

1. Abi-Rached, L., Jobin, M.J., Kulkarni, S., McWhinnie, A., Dalva, K., Gragert, L., Babrzadeh, F., Gharizadeh, B., Luo, M., Plummer, F.A., Kimani, J., Carrington, M., Middleton, D., Rajalingam, R., Beksac, M., Marsh, S.G.E., Maiers, M., Guethlein, L.A., Tavoularis, S., Little, A-M., Green, R.E., Norman, P.J., and Parham, P., 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science, 334(6052), pp.89-94.
2. Alaloul, W.S., Liew, M.S., Zawawi, N.A.W., Mohammed, B.S., Adamu, M., and Abdul Aziz, H., 2018. An Artificial Neural Networks (ANN) model for evaluating construction project performance based on coordination factors. [Journal Name], Article 1507657. DOI: [Insert DOI if available]. Published online: 28 August 2018.
3. Andrés, A.M., Hubisz, M.J., Indap, A., Torgerson, D.G., Degenhardt, J.D., Boyko, A.R., Gutenkunst, R.N., White, T.J., Green, E.D., Bustamante, C.D., Clark, A.G., Nielsen, R., 2009. Targets of balancing selection in the human genome. Molecular Biology and Evolution, 26(12), pp.2755-2764.
4. Appadurai, V., Bybjerg-Grauholm, J., Krebs, M.D., Rosengren, A., Buil, A., Ingason, A., Mors, O., Børglum, A.D., Hougaard, D.M., Nordentoft, M., Mortensen, P.B., Delaneau, O., Werge, T., and Schork, A.J., 2023. Accuracy of haplotype estimation and whole genome imputation affects complex trait analyses in complex biobanks. Communications Biology, 6, p.101. DOI: 10.1038/s42003-023-04477-y. PMCID: PMC9876938. PMID: 36697501.
5. Aston, K.I., Krausz, C., Laface, I., Ruiz-Castané, E., and Carrell, D.T., 2010. Evaluation of 172 candidate polymorphisms for association with oligozoospermia or azoospermia in a large cohort of men of European descent. Human Reproduction, 25(6), pp.1383-1397. DOI: 10.1093/humrep/deq081. Published: 08 April 2010.

6. Azevedo, L., Serrano, C., Amorim, A., and Cooper, D.N., 2015. Trans-species polymorphism in humans and the great apes is generally maintained by balancing selection that modulates the host immune response. Human Genomics, 9, Article number: 21.
7. Bakija-Konsuo, A., Mulić, R., Boraska, V., Pehlic, M., Huffman, J.E., Hayward, C., Marlais, M., Zemunik, T., and Rudan, I., 2011. Leprosy epidemics during history increased protective allele frequency of PARK2/PACRG genes in the population of the Mljet Island, Croatia. European Journal of Medical Genetics, [e-pub ahead of print]. DOI: 10.1016/j.ejmg.2011.06.010. PMID: 21816242.
8. Bell, G., 2010. Fluctuating selection: the perpetual renewal of adaptation in variable environments.
9. Bitarello, B.D., de Filippo, C., Teixeira, J.C., Schmidt, J.M., Kleinert, P., Meyer, D., and Andrés, A.M., 2018. Signatures of Long-Term Balancing Selection in Human Genomes. Genome Biology and Evolution, 10(3), pp.939-955.
10. Charlesworth, B., and Charlesworth, D., 2017. Population genetics from 1966 to 2016. Heredity, 118, pp.2-9.
11. Charlesworth, D., 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. PLoS Genetics, 2(4), e64.
12. Chen, B., Cole, J.W., and Grond-Ginsbach, C., 2017. Departure from Hardy-Weinberg Equilibrium and Genotyping Error. Frontiers in Genetics, 8, p.167. DOI: 10.3389/fgene.2017.00167. PMCID: PMC5671567. PMID: 29163635.
13. Christie, M.R. and McNickle, G.G., 2023. Negative frequency dependent selection unites ecology and evolution.
14. Cutter, A.D., and Payseur, B.A., 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics, 14(4), pp.262-274.
15. de Léséleuc, L., Orlova, M., Cobat, A., Girard, M., Huong, N.T., Ba, N.N., Thuc, N.V., Truman, R., Spencer, J.S., Adams, L., Thai, V.H., Alcais, A., and Schurr, E., 2013. PARK2 Mediates Interleukin 6 and Monocyte Chemoattractant Protein 1 Production by Human Macrophages. PLoS Neglected Tropical Diseases, 7(1), p.e2015. DOI: 10.1371/journal.pntd.0002015. PMCID: PMC3547867. PMID: 23350010.
16. DeGiorgio, M., Lohmueller, K.E., Nielsen, R., 2014. A Model-Based Approach for Identifying Signatures of Ancient Balancing Selection in Genetic Data. PLOS Genetics.

17. Dong, Y., Zhang, T., Li, X., Yu, F., Yu, H., and Shao, S., 2022. Identification of Key Prognostic-Related miRNA-mRNA Pairs in the Progression of Endometrial Carcinoma. [Journal Name], [Volume and Issue Number]. DOI: 10.1159/000520339. PMID: 35081534.
18. Durvasula, A., Sankararaman, S., 2019. A statistical model for reference-free inference of archaic local ancestry. (Version 2).
19. Durvasula, A., Sankararaman, S., 2020. Recovering signals of ghost archaic introgression in African populations. Science Advances, 6(7).
20. Enard, D., Cai, L., Gwennap, C., and Petrov, D.A., 2016. Viruses are a dominant driver of protein adaptation in mammals. eLife.
21. Ewens, W.J. and Thomson, G., 1970. Heterozygote selective advantage.
22. Ferrer-Admetlla, A., Liang, M., Korneliussen, T., and Nielsen, R., 2014. On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure. Molecular Biology and Evolution, 31(5), pp.1275-1291. DOI: 10.1093/molbev/msu077.
23. Fisher, R.A., 1930. The genetical theory of natural selection. Clarendon Press.
24. Fijarczyk, A., and Babik, W., 2015. Detecting balancing selection in genomes: limits and prospects. Molecular Ecology.
25. Fumagalli, M., Camus, S.M., Diekmann, Y., Burke, A., Camus, M.D., Norman, P.J., Joseph, A., Abi-Rached, L., Benazzo, A., Rasteiro, R., Mathieson, I., Topf, M., Parham, P., Thomas, M.G., and Brodsky, F.M., 2019. Genetic diversity of CHC22 clathrin impacts its function in glucose metabolism. eLife, 8, e41517.
26. Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., and Nielsen, R., 2011. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution.
27. Frenay, J., Bellaye, P.S., Oudot, A., Helbling, A., Petitot, C., Ferrand, C., Collin, B., and Dias, A.M.M., 2022. IL-1RAP, a Key Therapeutic Target in Cancer. International Journal of Molecular Sciences, 23(23), p.14918. DOI: 10.3390/ijms232314918. PMCID: PMC9735758. PMID: 36499246.
28. Gregory, T.R., 2009. Understanding Natural Selection: Essential Concepts and Common Misconceptions. Evolution: Education and Outreach, 2, pp.156-175.
29. Goodfellow, I., et al., 2016. Deep Learning. MIT Press, Cambridge, MA. Available at: http://www.deeplearningbook.org.

30. Gundling Jr, W.E., Post, S., Illsley, N.P., Echalar, L., Zamudio, S., and Wildman, D.E., 2023. Ancestry dependent balancing selection of placental dysferlin at high-altitude. Frontiers in Cell and Developmental Biology, 11.
31. Haller, B.C., and Messer, P.W., 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Molecular Biology and Evolution, 36(3), pp.632-637. DOI: 10.1093/molbev/msy228. PMCID: PMC6389312. PMID: 30517680.
32. Hancock, A.M. and Di Rienzo, A., 2008. Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. Annual Review of Anthropology, 37, pp.197-217.
33. Ha, E., and Bae, J.H., 2018. Zinc transporter SLC39A11 polymorphisms are associated with chronic gastritis in the Korean population: the possible effect on spicy food intake. [Journal Name]. Available online 22 April 2018, Version of Record 13 July 2018.
34. Hartl, D.L. and Clark, A.G., 2007. Principles of Population Genetics. 4th Edition, Sinauer Associates. Review in Ecoscience, 14(4), pp.544-545. ISBN: 978-0-878-93308-2.
35. Hedrick, P.W., 2011. Population genetics of malaria resistance in humans. Heredity, 107, pp.283-304.
36. Hermisson, J., and Pennings, P.S., 2005. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics, 169(4), pp.2335-2352.
37. Hu, Z., Niu, G., Ren, J., Wang, X., Chen, L., Hong, R., and Ke, C., 2019. TAFA5 promotes proliferation and migration in gastric cancer. Molecular Medicine Reports, 20(5), pp.4477-4488. DOI: 10.3892/mmr.2019.10724. PMCID: PMC6797941. PMID: 31702029.
38. Jouganous, J., Long, W., Ragsdale, A.P., and Gravel, S., 2017. Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. Genetics, 206(3), pp.1549-1567. DOI: 10.1534/genetics.117.200493.
39. Kaplan, N.L., Darden, T., Hudson, R.R., 1988. The Coalescent Process in Models with Selection. Genetics, 120(3), pp.819-829. PMID: 3066685. PMCID: PMC1203559.
40. Kasetty, G., Jönsson, M., Persson, J., Grundevik, P., Nys, J., Cohen, S., Borde, A., and Birrell, M.A., 2022. IL1RAP is an emerging target in chronic airway inflammatory diseases. European Respiratory Journal, 60, p.1948. DOI: 10.1183/13993003.congress-2022.1948.

41. Khazim, K., Azulay, E.E., Kristal, B., and Cohen, I., 2018. Interleukin 1 gene polymorphism and susceptibility to disease. Immunological Reviews, 281(1), pp.40-56. DOI: 10.1111/imr.12620. PMID: 29247999.
42. Klassmann, A., and Gautier, M., 2022. Detecting selection using extended haplotype homozygosity (EHH)-based statistics in unphased or unpolarized data. PLoS One, 17(1), p.e0262024. DOI: 10.1371/journal.pone.0262024. PMCID: PMC8765611. PMID: 35041674.
43. Lao, O., 2021. Selection still shapes our genome. Nature Human Behaviour, 5, pp.1600-1601.
44. LeCun, Y., Bengio, Y., and Hinton, G., 2015. Deep Learning. Nature, 521(7553), pp.436-444. DOI:10.1038/nature14539.
45. Leturiondo, A.L., Noronha, A.B., Mendonça, C.Y.R., Ferreira, C.O., Alvarado-Arnez, L.E., Manta, F.S.N., Bezerra, O.C.L., de Carvalho, E.F., Moraes, M.O., Rodrigues, F.C., and Talhari, C., 2020. Association of NOD2 and IFNG single nucleotide polymorphisms with leprosy in the Amazon ethnic admixed population. PLoS Neglected Tropical Diseases, 14(5), p.e0008247. DOI: 10.1371/journal.pntd.0008247. PMCID: PMC7239438. PMID: 32433683.
46. Li, J., Liu, H., Liu, J., Fu, X., Yu, Y., Yu, G., Chen, S., Chu, T., Lu, N., Bao, F., Yuan, C., and Zhang, F., 2012. Association study of the single nucleotide polymorphisms of PARK2 and PACRG with leprosy susceptibility in Chinese population. European Journal of Human Genetics, 20(5), pp.488-492. DOI: 10.1038/ejhg.2011.190. PMCID: PMC3330206. PMID: 22009144.
47. Liu, Y., Bockermann, R., Hadi, M., Safari, I., Carrion, B., Kveiborg, M., and Issazadeh-Navikas, S., 2021. ADAM12 is a costimulatory molecule that determines Th1 cell fate and mediates tissue inflammation. Cell & Molecular Immunology, 18(8), pp.1904-1919. DOI: 10.1038/s41423-020-0486-8. PMCID: PMC8322154. PMID: 32572163.
48. Malaria Genomic Epidemiology Network; Band, G., Rockett, K.A., Spencer, C.C.A., Kwiatkowski, D.P., 2015. A novel locus of resistance to severe malaria in a region of ancient balancing selection.
49. McDonald, J.H., and Kreitman, M., 1991. Neutral mutation hypothesis test. Nature, 354, p. 116.

50. Mehrabadi, A.Z., Aghamohamadi, N., Khoshmirsafa, M., Aghamajidi, A., Pilehforoshha, M., Massoumi, R., and Falak, R., 2022. The roles of interleukin-1 receptor accessory protein in certain inflammatory conditions. Immunology. First published: 01 March 2022. DOI: 10.1111/imm.13462.
51. Mira, M.T., Alcaïs, A., Nguyen, V.T., Moraes, M.O., Di Flumeri, C., Vu, H.T., Mai, C.P., Nguyen, T.H., Nguyen, N.B., Pham, X.K., Sarno, E.N., Alter, A., Montpetit, A., Moraes, M.E., Moraes, J.R., Doré, C., Gallant, C.J., Lepage, P., Verner, A., Van De Vosse, E., Hudson, T.J., Abel, L., and Schurr, E., 2004. Susceptibility to leprosy is associated with PARK2 and PACRG. Nature, 427(6975), pp.636-640. DOI: 10.1038/nature02326. PMCID: Free article. PMID: 14737177.
52. Plagnol, V., Wall, J.D., 2006. Possible Ancestral Structure in Human Populations. PLOS Genetics.
53. Quach, H., Wilson, D., Laval, G., Patin, E., 2013. Different Selective Pressures Shape the Evolution of Toll-like Receptors in Human and African Great Ape Populations. Human Molecular Genetics, 22(23),
54. Rubio-Perez, J.M., and Morillas-Ruiz, J.M., 2012. A Review: Inflammatory Process in Alzheimer's Disease, Role of Cytokines. The Scientific World Journal, 2012, p.756357. DOI: 10.1100/2012/756357. PMCID: PMC3330269. PMID: 22566778.
55. Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., & Reich, D., 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature, 507, p.354-357.
56. Schrider, D.R., and Kern, A.D., 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics. Available online 10 January 2018, Version of Record 26 March 2018.
57. Schmidt, J.M., and Huber, C.D., 2023. Fluctuating selection and the determinants of genetic variation. Trends in Genetics, 39(6), pp.491-504.
58. Stewart, K.A., Draaijer, R., Kolasa, M.R., & Smallegange, I.M. The role of genetic diversity in the evolution and maintenance of environmentally cued, male alternative reproductive tactics. BMC Evolutionary Biology, 19, 2019.
59. Stockdale, C., Rice, L., Carter, C., Berry, I., Poulter, J., O'Riordan, S., Pollard, S., Anwar, R., Tooze, R., and Savic, S., 2021. Novel Case of Tripeptidyl Peptidase 2 Deficiency Associated with Mild Clinical Phenotype. Journal of Clinical Immunology,

41(5), pp.1123-1127. DOI: 10.1007/s10875-021-01006-6. PMCID: PMC7937547. PMID: 33682069.

60. Subramanian, S., 2016. The effects of sample size on population genomic analyses – implications for the tests of neutrality. BMC Genomics, 17, p.123. DOI: 10.1186/s12864-016-2441-8. PMCID: PMC4761153. PMID: 26897757.
61. Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123(3), pp.585-595.
62. Takahata, N. and Nei, M., 1990. Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics, 124(4), pp.967-978.
63. The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature, 526, pp.68-74.
64. Urnikyte, 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.
65. Urnikyte, A., Masiulyte, A., Pranckeniene, L., and Kučinskas, V., 2023. Disentangling archaic introgression and genomic signatures of selection at human immunity genes. Infection, Genetics and Evolution, 116, Article number: 105528.
66. Vitti, J.J., Grossman, S.R., and Sabeti, P.C., 2013. Detecting natural selection in genomic data. Annual Review of Genetics, 47, pp.97-120.
67. Wright, S.I. and Charlesworth, B., 2004. The HKA Test Revisited: A Maximum-Likelihood-Ratio Test of the Standard Neutral Model. Genetics, 168(2), pp.1071-1076. doi: 10.1534/genetics.104.026500. PMCID: PMC1448833. PMID: 15514076.
68. Zeberg, H., and Pääbo, S., 2020. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature, 587, pp.610-612.
69. Zeng, K., Fu, Y.X., Shi, S., and Wu, C.I., 2006. Statistical Tests for Detecting Positive Selection by Utilizing High-Frequency Variants. Genetics, 174(3), pp.1431-1439. DOI: 10.1534/genetics.106.061432. PMCID: PMC1667063. PMID: 16951063.

ACKNOWLEDGMENTS

Thank you very much to Dr. Alina Urnikytė for the opportunity to work on a remarkably interesting topic and for all the support despite the busy schedule from both sides. Hope to continue collaboration further on in the future and all good luck with upcoming projects and students!

SUPPLEMENTARY MATERIAL

Supplementary Table 1. SNP variants and predicted selection probability by chromosome and selection type.

chromosome	ch1	ch2	ch3	ch4	ch5	ch6	ch7	ch8	ch9	ch10	ch11	ch12	ch13	ch14	ch15	ch16	ch17	ch18	ch19	ch20	ch21	ch22
recent_selection	150	1	763	454	133	1393	353	145	390	429	375	555	1635	7	1059	913	531	714	951	458	738	1224
recent_selection_avg	0.998	0.985	0.998	0.999	0.999	0.998	0.999	0.994	0.999	0.999	0.999	0.999	0.998	0.956	0.999	1.000	0.998	0.999	0.998	0.999	0.999	0.999
recent_balancingSelection	23	0	106	58	10	252	28	20	54	74	52	48	397	0	136	116	62	111	140	58	109	177
recent_balancingSelection_avg	0.6006	0.000	0.6119	0.594	0.579	0.6001	0.599	0.5667	0.629	0.6115	0.586	0.599	0.6119	0.000	0.6009	0.593	0.589	0.588	0.577	0.596	0.597	0.610
recent_negativeFreq-dep.Selection	147	1	738	453	133	13	350	144	385	425	373	550	16	7	10	913	529	695	933	452	729	12

						75							04		48						06
recent_negativeFreq-dep.Selection_avg	0.5665	0.5526	0.5552	0.5573	0.5573	0.5562	0.5580	0.5546	0.5579	0.5584	0.5578	0.5588	0.5556	0.5225	0.5578	0.6006	0.5581	0.5575	0.5560	0.5584	0.5573	0.5581
recent_overdominance	85583	90600	66333	57330	63339	56693	55583	55590	45579	54545	53223	48449	25225	35000	23889	26225	25573	25572	13115	23376	7889	3557
old_selection	1510	1	7633	4544	1333	13393	3533	1445	3900	4299	3755	5555	16355	7	10559	9133	5331	7144	9511	4558	7388	12244
old_selection_avg	0.9997	0.9986	0.9998	0.9999	0.9999	0.9999	0.9999	0.9992	0.9999	0.9999	0.9999	0.9999	0.9998	0.9944	0.9999	1.0000	0.9999	0.9999	0.9998	0.9999	0.9999	0.9999
old_balancingSelection	0	0	10	0	0	1	0	0	1	0	0	0	11	0	0	1	0	0	0	1	1	3
old_balancingSelection_avg	0.0000	0.0000	0.6002	0.0000	0.0000	0.5220	0.0000	0.0000	0.5009	0.0000	0.0000	0.0000	0.5552	0.0000	0.0000	0.5116	0.0000	0.0000	0.0000	0.5111	0.5037	0.5336
old_negativeFreq-dep.Selection	1510	1	7633	4544	1333	13393	3533	1445	3900	4299	3755	5555	16355	7	10559	9133	5331	7144	9511	4558	7388	12244

old_negativeFreq-dep.Selection_avg	0.655	0.549	0.628	0.676	0.672	0.652	0.687	0.614	0.684	0.700	0.686	0.699	0.641	0.548	0.680	0.741	0.684	0.674	0.645	0.695	0.672	0.688
old_overdominance	85800	90600	66008	57229	63399	56675	55580	55589	45574	54511	53211	48444	24944	35000	23678	26225	25571	25553	12970	23700	7800	3339
medium_selection	150	1	763	454	133	1393	353	145	390	429	375	555	1635	7	1059	913	531	714	9511	458	738	1224
medium_selection_avg	1.000	0.998	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000	0.990	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
medium_balancingSelection	0	0	17	0	0	1	0	0	0	0	0	1	24	0	0	0	0	0	5	0	1	6
medium_balancingSelection_avg	0.000	0.000	0.600	0.000	0.000	0.509	0.000	0.000	0.000	0.000	0.000	0.506	0.672	0.000	0.000	0.000	0.000	0.000	0.506	0.000	0.511	0.531
medium_negativeFreq-dep.Selection	150	1	763	454	133	1393	353	145	390	429	375	555	1635	7	1059	913	531	714	9511	458	738	1224
medium_negativeFreq-dep.Selection_avg	0.7	0.5	0.6	0.7	0.7	0.7	0.7	0.6	0.7	0.7	0.7	0.7	0.6	0.5	0.7	0.8	0.7	0.7	0.7	0.7	0.7	0.7

	09	75	78	36	32	04	52	62	44	59	45	61	89	84	37	02	45	30	00	56	29	47
medium_overdominance	8580	9060	6608	5729	6339	5675	5580	5589	4574	5451	5321	4844	2494	3500	2378	2625	2571	2553	1297	2370	780	339

Supplementary Table 2. Genes and proteins under balancing selection by chromosome and SNP variant.

chr	from	to	AltaiNean	SNPs	PredClass	Genes	Consequences	Proteins
3.0	189261797.0	189309846.0	0.9394	['rs2037184', 'rs6444269']	Negative Freq-Dep. Selection	[]	['intergenic_variant', 'intergenic_variant']	[]
3.0	191112052.0	191161035.0	1.0	['rs12696598', 'rs724438', 'rs4687112']	Negative Freq-Dep. Selection	[]	['intergenic_variant', 'regulatory_region_variant', 'regulatory_region_variant']	[]
3.0	191661171.0	191710630.0	0.9479	['rs3943979', 'rs9857115', 'rs2042865']	Negative Freq-Dep. Selection	[]	['intergenic_variant', 'intergenic_variant', 'regulatory_region_variant']	[]

3.0	191711064.0	191755302.0	0.9818	['rs2193880', 'rs3773989']	Negative Freq-Dep. Selection	['ENSG00000196083']	['intron_variant', 'intron_variant']	[['IL1RAP', 'interleukin 1 receptor accessory protein [Source:HGNC Symbol;Acc:5995]']]
4.0	186117497.0	186164767.0	0.9268	['rs4355430']	Negative Freq-Dep. Selection	[]	['intergenic_variant']	[]
6.0	1438673366.0	143897977.0	0.9661	['rs2128977']	Negative Freq-Dep. Selection	[]	['regulatory_region_variant']	[]
6.0	156898474.0	156946230.0	0.8947	['rs288976']	Balancing Selection	[]	['intergenic_variant']	[]
6.0	156898474.0	156946230.0	0.8947	['rs288976']	Negative Freq-Dep. Selection	[]	['intergenic_variant']	[]
6.0	159198260.0	159247967.0	0.9672	['rs9457507', 'rs9457511',	Negative Freq-Dep. Selection	['ENSG00000203711']	['intron_variant', 'intron_variant', 'intron_variant']	[['C6orf99', 'chromosome 6 open reading frame 99

				'rs12200537']				[Source:HGNC Symbol;Acc:21179]']]
6.0	161248552.0	161291617.0	0.9286	['rs1962358']	Negative Freq-Dep. Selection	[]	['intergenic_variant']	[]
6.0	163048514.0	163097724.0	0.9091	['rs9356058']	Negative Freq-Dep. Selection	['ENSG00000112530']	['intron_variant']	[['PACRG', 'PARK2 co-regulated [Source:HGNC Symbol;Acc:19152]']]
9.0	129660977.0	129709828.0	0.8684	['rs3780663']	Balancing Selection	['ENSG00000257524', 'ENSG00000106992']	['intron_variant']	[['RP11-203J24.9', 'No description available'], ['AK1', 'adenylate kinase 1 [Source:HGNC Symbol;Acc:361]']]
9.0	129660977.0	129709	0.8684	['rs3780663']	Negative Freq-Dep. Selection	['ENSG00000257524',	['intron_variant']	[['RP11-203J24.9', 'No

		828.0				'ENSG0000010 6992']		description available'], ['AK1', 'adenylate kinase 1 [Source:HGNC Symbol;Acc:361]']]
10.0	12767117 5.0	127711 500.0	0.9787	['rs12251 249']	Negative Freq-Dep. Selection	['ENSG000001 48848']	['intron_variant']	[['ADAM12', 'ADAM metallopeptidase domain 12 [Source:HGNC Symbol;Acc:190]']]
13.0	10207181 2.0	102116 884.0	0.8611	['rs37369 72']	Balancing Selection	['ENSG000001 34900']	['synonymous_variant']	[['TPP2', 'tripeptidyl peptidase II [Source:HGNC Symbol;Acc:12016]']]
13.0	10207181 2.0	102116 884.0	0.8611	['rs37369 72']	Negative Freq-Dep. Selection	['ENSG000001 34900']	['synonymous_variant']	[['TPP2', 'tripeptidyl peptidase II [Source:HGNC

								Symbol;Acc:12016]']]
13.0	107421329.0	107469601.0	0.9365	['rs9520690', 'rs9520691', 'rs9520696', 'rs9514782', 'rs9514792', 'rs12871532', 'rs1022909']	Negative Freq-Dep. Selection	[]	['regulatory_region_variant', 'intergenic_variant', 'intergenic_variant', 'intergenic_variant', 'intergenic_variant', 'intergenic_variant']	[]
15.0	86002672.0	86049988.0	1.0	['rs6496435']	Negative Freq-Dep. Selection	['ENSG00000259560']	['intron_variant']	[['RP11-648K4.2', 'No description available']]
15.0	89501124.0	89544688.0	1.0	['rs12905479', 'rs17594552']	Negative Freq-Dep. Selection	['ENSG00000185518']	['intron_variant', 'intron_variant']	[['SV2B', 'synaptic vesicle glycoprotein 2B [Source:HGNC Symbol;Acc:16874]']]

17.0	65702659.0	65749953.0	0.8519	['rs236586', 'rs236523', 'rs236531']	Negative Freq-Dep. Selection	[]	['intergenic_variant', 'intergenic_variant', 'intergenic_variant']	[]
17.0	68256221.0	68299201.0	0.9333	['rs1110614', 'rs4793484', 'rs9906450']	Negative Freq-Dep. Selection	['ENSG00000133195']	['intron_variant', 'intron_variant', 'intron_variant']	[['SLC39A11', 'solute carrier family 39, member 11 [Source:HGNC Symbol;Acc:14463]']]
19.0	52928986.0	52976774.0	0.8989	['rs2334295']	Negative Freq-Dep. Selection	['ENSG00000269656']	['intron_variant']	[['CTD-2571L23.6', 'No description available']]
21.0	28411654.0	28456611.0	0.8571	['rs2223743', 'rs6516819', 'rs9976944', 'rs2831534']	Negative Freq-Dep. Selection	['ENSG00000236532', 'ENSG00000232079']	['intron_variant', 'intron_variant', 'intron_variant', 'intron_variant']	[['AL035610.2', 'No description available'], ['AL035610.1', 'No description available']]

22.0	34153779.0	34203496.0	0.9706	['rs5755730']	Balancing Selection	[]	['intergenic_variant']	[]
22.0	34153779.0	34203496.0	0.9706	['rs5755730']	Negative Freq-Dep. Selection	[]	['intergenic_variant']	[]
22.0	35654133.0	35703551.0	0.8596	['rs11089816', 'rs5750348']	Balancing Selection	[]	['intergenic_variant', 'intergenic_variant']	[]
22.0	35654133.0	35703551.0	0.8596	['rs11089816', 'rs5750348']	Negative Freq-Dep. Selection	[]	['intergenic_variant', 'intergenic_variant']	[]
22.0	36103709.0	36133397.0	1.0	['rs4820286', 'rs5750428']	Negative Freq-Dep. Selection	['ENSG00000243902', 'ENSG00000166897']	['intron_variant', 'intron_variant']	[['RP1-63G5.5', 'No description available'], ['ELFN2', 'extracellular leucine-rich repeat and fibronectin type III domain containing 2 [Source:HGNC

								Symbol;Acc:29396]']]
22.0	44107968.0	44149935.0	0.9167	['rs11704481', 'rs2064068', 'rs1044742', 'rs5764698']	Balancing Selection	['ENSG00000100376', 'ENSG00000077935']	['intron_variant', '3_prime_UTR_variant', '3_prime_UTR_variant', 'missense_variant']	[['FAM118A', 'family with sequence similarity 118, member A [Source:HGNC Symbol;Acc:1313]'], ['SMC1B', 'structural maintenance of chromosomes 1B [Source:HGNC Symbol;Acc:11112]']]
22.0	44107968.0	44149935.0	0.9167	['rs11704481', 'rs2064068', 'rs1044742',	Negative Freq-Dep. Selection	['ENSG00000100376', 'ENSG00000077935']	['intron_variant', '3_prime_UTR_variant', '3_prime_UTR_variant',	[['FAM118A', 'family with sequence similarity 118,

				'rs5764698']			'missense_variant']	member A [Source:HGNC Symbol;Acc:1313]'], ['SMC1B', 'structural maintenance of chromosomes 1B [Source:HGNC Symbol;Acc:11112]']]
22.0	47432419.0	47445957.0	0.9107	['rs5771906', 'rs6010568']	Negative Freq-Dep. Selection	['ENSG00000219438']	['intron_variant', 'intron_variant']	[['FAM19A5', 'family with sequence similarity 19 (chemokine (C-C motif)-like), member A5 [Source:HGNC Symbol;Acc:21592]']]