



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Biomedical Science

*Interuniversity Bachelor's degree programme
in Biology of Human and Environmental Health*

Final dissertation

**IMPLEMENTATION OF A COALESCENCE-
BASED METHOD TO INFER METABOLIC
ADAPTATIONS TO COLD CLIMATE IN A
SIBERIAN HUMAN POPULATION**

Supervisor
Professor Marco Sazzini

Candidate *Nora Taskova*
Student ID number: 2105894

Co-Supervisor
Dr. Giulia Ferraretti

TABLE OF CONTENTS

ABSTRACT	3
1.INTRODUCTION	5
2.AIM OF THE STUDY	9
3. METHODOLOGY	11
3.1 Overview of datasets and methodologies employed	11
3.2 Quality Controls	11
3.3 Principal Component Analysis (PCA)	13
3.4 Relate software	14
3.4.1 <i>Ne</i> estimation	15
3.4.2 Selection test and p-value computation	16
3.5 Crosscheck with previous results from LASSI-signet analyses	17
4. RESULTS	19
4.1 Quality Control Procedures	19
4.2 Implementation of the Relate pipeline of analyses	19
4.2.1 Effective Population Size Estimation	19
4.2.2 Implementation of a Selection Test	21
5. DISCUSSION	23
5.1 <i>PLPP3</i>	23
5.2 <i>PRKG1</i>	23
5.3 <i>ERBB4</i>	24
5.4 Cold-induced metabolic adaptations	24
6. CONCLUSION	27
REFERENCES	29

ABSTRACT

The Yakut represent a remarkable example of a population that had successfully adapted to one of the most remote and climatically extreme environments inhabited by modern humans: the harshly cold region of Siberia. Adaptation to such conditions required much more than cultural adaptations, it also necessitated the evolution of biological adaptations with a genetic basis. This thesis just explores the patterns of genomic variation that might have led to relevant metabolic adjustments, which enabled Yakut ancestors to thrive in such an extreme cold environment. By relying on the analysis of whole genome sequence (WGS) data for 25 Yakut individuals from the Human Genome Diversity Panel (HGDP), the effective population size of the population was first inferred through a coalescence-based method, which exploits a Hidden Markov Model (HMM) framework to infer the demographic history of a certain human group. Then the Relate algorithm was used to reconstruct the genealogies of genetic variants across the entire genome, by estimating their ancestral relationships and coalescence times to infer the complete genealogies of each mutation. Such an approach was used to infer the action of natural selection on the Yakut gene pool and the candidate adaptive genes identified, on which the present thesis has been focused, were shortlisted as the loci i) showing the most significant p-values from the performed selection test and ii) having being pinpointed by multiple selection analyses previously performed in other thesis projects. Genes such as *PLPP3*, *ERBB4*, and *PRKG1* emerged from this crosscheck of significant results. The involvement of all of these loci in various signaling pathways regulating insulin signaling, lipid metabolism and cardiovascular processes have been explored as the primary biological framework through which the genetic and phenotypic architecture of Yakut cold adaptation evolved. Although these metabolic adaptations have likely contributed for millennia to determine the evolutionary success of the Yakuts in inhabiting one of the world's most extreme environment, such peculiar lipid and cardiovascular traits may comport an increased susceptibility of this population to complex metabolic disorders under contemporary socio-cultural settings. In fact, the abandonment of traditional diet, the concomitant increasing nutritive impoverishment of the westernized diet and the shifts in lifestyle associated to climate change are becoming the new and more challenging “environmental” pressures.

1. INTRODUCTION

Bioinformatics and computational methods have completely revolutionized the research field of evolutionary biology since the start of their implementation, continuing to open new horizons in this ambit as their implementation evolves. International efforts to create a unified, rich and widely available dataset of human genome-wide data has succeeded with the 1000 Genome Project (1,000GP) initiative, which strived to comprehensively catalogue human genetic variation and to create a public large dataset for human genomic resource (Auton *et al.* 2015) (Figure 1.1). This dataset has found its use in several various contexts, spanning from GWAS and studies aimed at fine mapping specific genomic regions to investigate and filtering likely neutral variants in rare-disease cohorts, or uncovering population-specific genomic variability (Auton *et al.* 2015) (Figure 1.1). Today more than 2,500 individuals from 26 different global populations have their genome sequences as part of the 1,000GP dataset that has been leveraged through a variety of approaches intended to make inferences on the demographic and adaptive history of human populations.

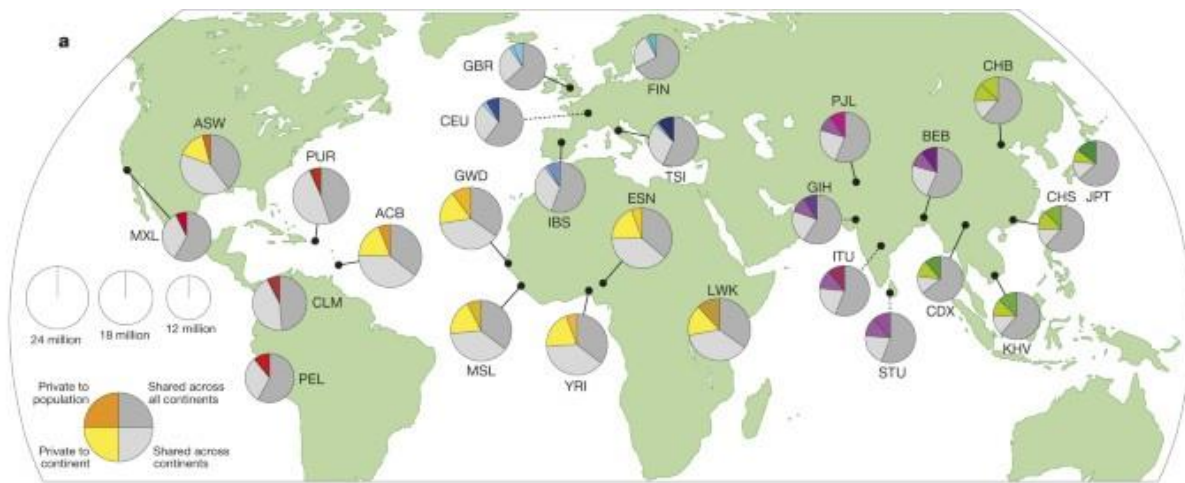


Figure 1.1: Polymorphic variants within sampled populations in the 1,000 Genomes reference panel. Each pie is divided in four slices, representing the ratio between variants private to a population, private to a continental area, shared across continental areas and shared across all continents. The figure was taken from the paper by Auton *et al.* (2015).

A very similar initiative to the 1,000 Genomes Project is the Human Genome Diversity Panel (HGDP) (Bergström *et al.* 2020) (Figure 1.2). The aim of this project was, in fact, to build a reference dataset for those human groups that are still under-represented or not represented at all in public reference datasets such as the 1,000 Genomes Project. For instance, these human groups comprehend indigenous populations from Central and South America, but also populations from remote North Siberia regions,

such as the Yakut people studied in the present thesis. Building on two previous theses conducted at the Centre for Genome Biology (University of Bologna), the present study aims to analyze the Yakut genomes in order to investigate the genetic bases of their biological adaptations evolved in response to extreme cold.

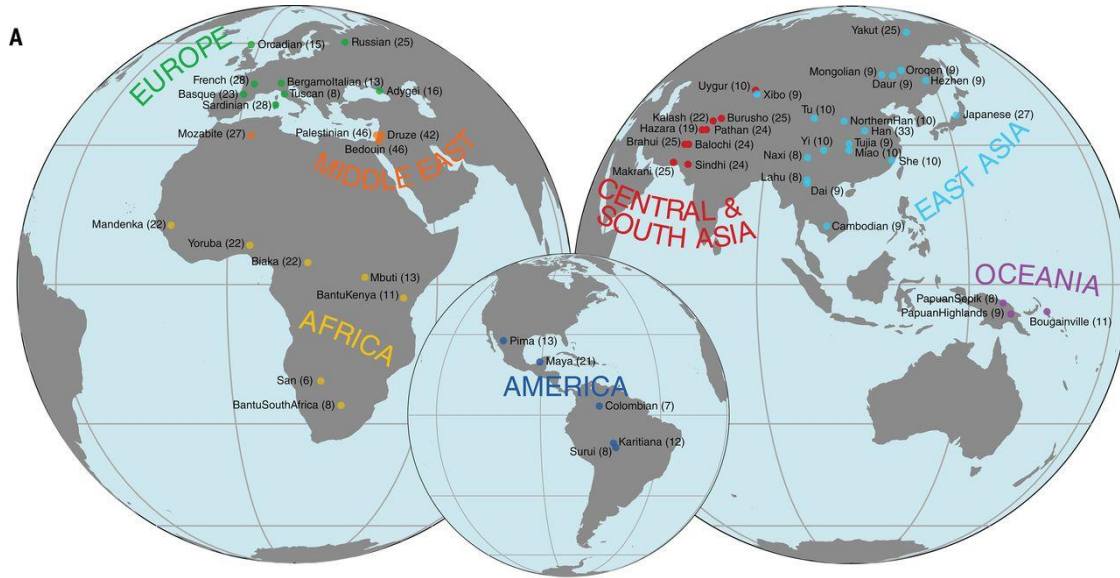


Figure 1.2: Geographical origins of the 54 populations from the HGDP project, along with the number of sequenced individuals for each human group reported in parenthesis. The figure was taken from the paper by Bergström et al. (2020), in which the Yakut whole genomes analyzed in the present thesis were generated.

Of the many statistical approaches useful to analyse genomic data, in this thesis are described the Principal Component Analysis (PCA) (Patterson, N. et al. 2006), the Relate software (Spiedel, L. *et al.* 2019) and the LASSI-signet analyses (Harris & DeGiorgio 2020) (Gouy et al. 2017), with a central place dedicated to the Relate algorithm. In fact, both PCA and LASSI-signet pipelines have been performed in the context of the aforementioned theses and described here for contextualize previous results, while the Relate method was carried out in the present study to validate such results. The new advantage brought by Relate consists of a higher level of accuracy and computational speed compared to other existing methods in terms of performances to process thousands of samples, while using a coalescence-based approach (Spiedel, L. *et al.* 2019). In principle, a sample's historical coalescence, recombination and mutation events are described through the observed patterns of variation, from which Relate cleverly leverages an underlying coalescent model to build a framework of a population's genealogical history and the mutation events it underwent throughout it (Spiedel, L. *et al.* 2019).

A peculiar turn is taken as Relate was used to estimate the genealogical genetic history of the Yakut population, an extremely isolated population with a fairly low effective population size whose demographic history is still a matter of debates among scholars. In this study, a small part of the big

enigma of the ethnically Turkic population (Zlojutro *et al.* 2009) that found its place in the most challenging and uninhabitable lands of Siberia is explored as computational data is interpreted through a variety on genomics, physiology and ethnical studies previously conducted.

2. AIM OF THE STUDY

The primary purpose of this study was understanding the evolutionary mechanisms and the extent to which living for millennia in conditions of extreme cold temperatures has shaped the genetic and metabolic profile of the Yakut population. Since their occupation of the northernmost lands of Siberia has been fairly recent, most probably dating back to the period between the thirteenth to the fifteenth century AD, presuming the pressure of the expanding Mongol Empire of the day (Pakendorf et al. 2006), Some of the metabolic adaptations of the Yakut are supposed to have arisen remarkably quickly relative to the extremity of the conditions they had to adapt to. Conclusions from investigation of the mtDNA sequence diversity and linguistic analysis state that there could have been a certain degree of admixture with the local Evenk population upon their arrival (Pakendorf et al. 2006). Even despite, the Yakut remain an isolated population with a remarkable consistency in their genetic makeup throughout history (Crubézy et al. 2026).

The molecular metabolic adjustments and the time that took to develop such a complex adaptive phenotype are still debated topics, with much more to be understood about them. Especially today, with both demographic and climatic change taking place more rapidly than ever before, the plasticity of this phenotype is put under question. The fact that modern day health problems as diabetes and obesity are even more dangerous for indigenous populations, who have developed alterations in the metabolic pathway has already been argued (Felber & Golay 2002). Nevertheless, the reasoning behind it and especially the strategies to resolve this growing problem are still scarce.

Following a computational analysis of the available Yakut whole genomes to investigate genomic signatures ascribable to the action of natural selection and literature research into prior work on cold-induced adaptations, as well as the correlation between a genetic background to susceptibility patterns of diabetes and obesity, this study aims at reasoning the connection between the harsh Siberian climate, the Yakut genetic makeup and the benefits, or in present-day context the disadvantages resulting from it.

3. METHODOLOGY

3.1 Overview of datasets and methodologies employed

In order to conduct this study, a variety of computational methods and publicly available resources concerning human population genomic data were used. In particular, publicly available WGS (whole genome sequence) datasets from the 1000 Genome Project (1000GP) containing the sequenced genomes of 2,504 individuals from 26 different populations (Auton *et al.* 2015) and from the Human Genome Diversity Panel (HGDP) composed of 929 genomes from 54 geographically, linguistically, and culturally diverse human populations (Bergström *et al.* 2020) were utilized. Specifically, the dataset analyzed in the present study included 25 WGS data of Yakut individuals generated in the context of the HGDP project. Patterns of genomic variation of these samples were preliminary assessed through population structure analyses (e.g. Principal Component Analysis, PCA) carried out within the context of previous theses conducted at the Centre for Genome Biology of the University of Bologna.

Furthermore, the human reference sequence GRCh38 produced in 2013 by the Genome Reference Consortium (UCSC Genome Browser) was employed together with the Genomic mask and Ancestral genome files generated beforehand on the basis of it by the Relate software (Speidel, L., Forest, M., Shi, S. *et al.* 2019). The VCF (Variant Call Format) file was used throughout the process. In order to produce a visual summary of the vast and complex datasets the PCA method was implemented, generating a descriptive bidimensional plot that explores populations' structure relationships (Jolliffe, Cadima 2016). The first step before each analysis was performing Quality Control filtering (QCs) for which the PLINK 1.9 toolset was leveraged by processing the data in the form of .bim and .bed binary files.

The aforementioned Relate software, a newly developed method for genome-wide genealogy estimation, was implemented to estimate the effective population size (N_e) and the selection test. The Relate method was chosen for its improved estimations of coalescence ages and mutation mapping with respect to other approaches. Through modified implementation of the Li and Stephens Hidden Markov model (HMM) and the Markov Chain Monte Carlo (MCMC) algorithm, Relate is able to firstly identify a genealogical framework at each site of the genome and describe ancestral relationships and then build complete genealogies on large-scale datasets in a more time efficient manner than other methods (such as ARGweaver and RENT+), while maintaining and even improving the accuracy (Speidel, L., Forest, M., Shi, S. *et al.* 2019). The R software was implemented for finally visualize Relate results and throughout the whole Relate pipeline the BASH programming language was employed.

Lastly, the KEGG and NCBI databases, along with studies published on the PubMed database held by the National Library of Medicine, were utilized to explore relationships between the outlier genes detected in this and previous theses and the corresponding metabolic pathways.

3.2 Quality Controls

QC procedures included filtering of the data according to four main criteria: the SNVs (Single Nucleotide Variants) with more than 5% missing data were removed, the samples with more than 5% missing data

in the considered genomic sequence were removed, a selection according to the Hardy-Weinberg Equilibrium was performed (i.e. by filtering out variants out of equilibrium) and A/T, T/A, G/C and C/G variants were removed when it was not possible to detect the exact copy of the chromosome to which they belong. The data subjected to the aforementioned filtering criteria were in PLINK binary files comprehending *.bed*, *.bim* and *.fam* extensions.

The first step of the QC procedures was the exclusion of SNVs with more than 5% of missing data, which was done using the PLINK flag `--geno` with 0.05 as filtering threshold. This step was followed by the `--mind` flag, which removed samples with more than 5% of missing data per genomic sequence, so again by setting a threshold of 0.05. The next filtering step was the Hardy-Weinberg Equilibrium test, carried out with the function `--hwe`, which here was used as a method to filter out possible sequencing errors. However, the Hardy-Weinberg Equilibrium is by far a well-established principle in population evolutionary studies. Specifically, the Hardy-Weinberg theorem takes basic Mendelian genetic approaches beyond the two parental species and focuses on the genetic makeup of the population as a dynamic unit. The principle on which the Hardy-Weinberg theorem is based states that the allele frequency should be stable within a population, under the condition that it is not subjected to the external influences of migration, mutation, effects of natural selection or genetic drift, it is infinitely large and panmictic, so that it is possible to define it as an ‘ideal population’. Therefore, under these assumptions and given that there are two alleles at a locus, p and q , the expected genotype frequencies in the following generations are p^2 , $2pq$, and q^2 (Figure 3.1). It can also be described as two alleles, A and B, being present at a locus in a population. The possible outcomes for a genotype in the filial generation would be AA, AB or BB, where $AA = p^2$, $AB = pq$ and $BB = q^2$. A deviation from the Hardy-Weinberg Equilibrium primarily indicates the occurrence of evolutionary processes in a population, but as the extent of the deviation increases it becomes more likely that such SNVs identify sequencing errors. In that matter, the Hardy-Weinberg equilibrium function in this thesis was used to exclude samples possibly containing sequencing errors, with a set threshold of 0.05 divided for the number of variants included in the dataset by applying the Bonferroni correction for multiple tests (i.e., SNVs with a HWE p-values inferior to $9.78e-9$ were excluded). The final step in the QC procedures consists in the exclusion of ambiguous alleles. More specifically these are A/T, T/A, G/C and C/G sites for which it is not possible to determine which chromosome copy they are located on, and therefore to reconstruct the haplotypes within these genomic regions. These alleles were firstly identified through the combination of *cat* and *awk* functions in BASH and then removed by adopting the `--exclude` PLINK flag.

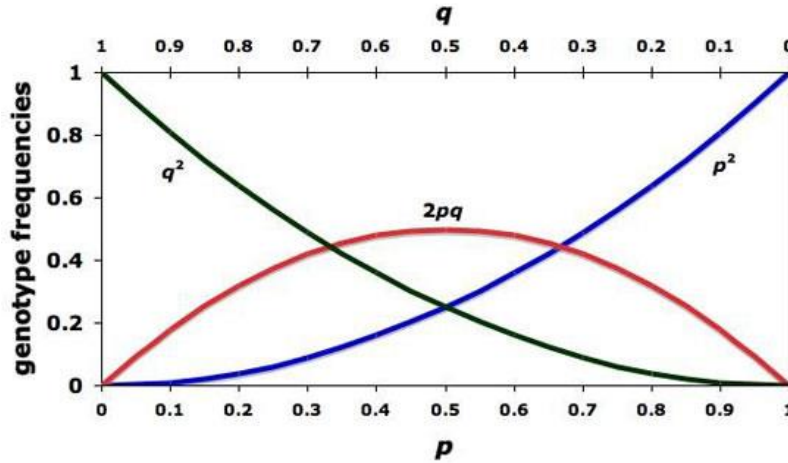


Figure 3.1: A plot of Hardy-Weinberg genotype frequencies (p^2 , $2pq$, q^2) as a function of allele frequencies (p and q). Rare alleles are found primarily in heterozygotes, given that q^2 is much smaller than $2pq$ when q approaches zero, and the same stands for p^2 when it approaches zero. Figure from Andrews et al. (2010).

3.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a powerful tool implemented to visualize complex data of large dimensionality. The principle on which it builds a two-dimensionally structured graph is by ordering the different variables, that are the principal components (PCs), of the dataset according to the amount of information that each PC stores. The first principal component is the variable accounting for most information and variance, followed by the subsequent ones in a decreasing order. Uncorrelated variables from the original dataset are translated into linear functions to produce new and fewer variables that successfully summarize the information with minimal loss of it. Computing these principal components into eigenvectors is what allows to understand their size and to find a way in which to order them according to their eigenvalues. However, the newly generated principal components depend on the dataset rather than being predefined. This allows the PCA method to be easily adaptive and thus descriptive instead of inferential. Although the idea behind it is based on a very demanding approach, PCA is indeed a very computationally feasible method which has found its application in many fields and research areas including GWAS (Genome Wide Association Studies) (Jolliffe, Cadima 2016). In genomic studies, PCA is used as a method to infer ancestry and to explore population genetic structure. There are, however, certain unresolved obstacles that are encountered when processing the data that consist of capturing linkage disequilibrium (LD) structure instead of population structure, shrinkage bias, uneven population sizes and difficulties in detecting sample outliers. That is why additional filtering steps are needed prior of producing the PCA output, which include removing related individuals, pruning variants in linkage disequilibrium and excluding outlier samples that could be indicators of poor genotyping quality or distant relatedness (Privé, Florian et al. 2020).

A PCA analysis was not conducted for this thesis, but instead the results produced in a previous thesis by Pognant Viù (2024), carried out at the Centre for Genome Biology and which had the aim of exploring

the polygenic adaptations evolved by the Yakut population, were shortly reported in this study (Figure 3.2). This PCA analysis provided a valuable insight into the genetic relationship of the Yakut people with other Eurasian populations showing both genomic differentiation of the Yakut samples and genetic affinity with other human groups from North Asian regions (Figure 3.2).

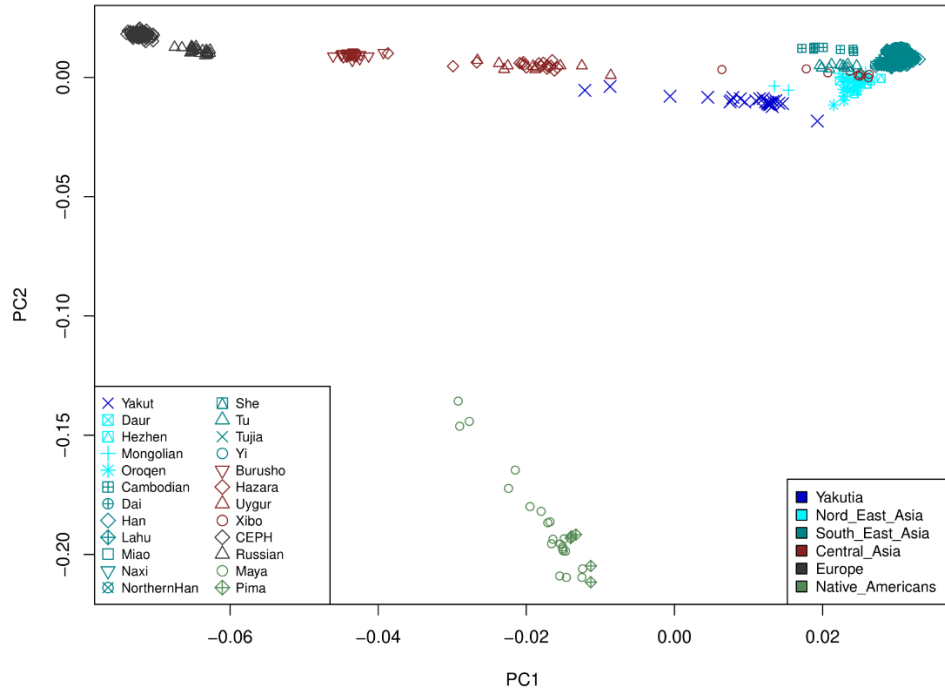


Figure 3.2: A scatter plot from the PCA analysis constructed using PC1 - capturing mainly the variability between North/South East Asian populations and Central Asian/European populations - paired with PC2 - showing the variability observable between Native American and Eurasian populations. Figure adapted from Pognant Viù (2024).

By identifying four principal components (PCs), three scatter plots were generated, one of which is displayed in Figure 3.2. Through these four variables the diversification between the Native American and Asian populations, within Asian populations and the particularly outlying position of the Yakut population were captured. These results suggest that all the populations that were considered in the study were internally homogenous, however the Yakut did notably stand out with their genetic distinctiveness. Geographical isolation appears as the most accountable factor able to explain this differentiation despite the overall shared genetic heritage with other North/East Asian populations.

3.4 Relate software

Relate takes a stepwise analytical approach in recreating the genealogical histories of haplotypes found in modern day populations. The Relate software alongside the respective documentation is available at

<https://myersgroup.github.io/relate/>. The relate pipeline starts by creating a nonsymmetrical distance matrix directly implementing posterior probabilities output by a Hidden Markov Model (HMM) in a similar way as proposed by Li and Stephens (2003), but considering the ancestral and derived state of each genetic variant, taking the accuracy and detail to a new level. This position-specific distance matrix is constructed at each genomic position where the rows estimate the relative order of coalescence between a particular sequence and the observed sequences. If we have the distance matrix d such that every entry d_{ij} stores the rescaled log likelihood of producing haplotype i by copying from haplotype j , which would represent the number of mutations carried by i locally around the focal variant, taking j as a reference sequence. This distance matrix is then employed to coalesce haplotypes and match the observed mutations to corresponding branches among populations. The branch lengths are estimated only after using MCMC algorithm that employs a coalescent prior model. The approach applied here is based on hierarchical clustering algorithm and considers each row of the distance matrix as an indication of the order of coalescence of each genomic position between the two haplotypes under observation, so that in the end the assumption on which the binary tree is built that the pair of clades that are merged first is the pair in which both clades coalesce with each other before coalescing with any other clade. This method allows for changing population sizes and in parallel estimating the varying effective population sizes (N_e) using the genome-wide collection of estimated genealogies. Once the tree topology has been estimated, the subject of focus can be deepened to the branch lengths and assigning particular mutations to their branches of origin. As the possibility of mistaking an introgression event and introduction of a mutation for settling of a mutation within a population, Relate uses a selection test estimating the speed of spread of a mutation within a lineage in comparison to other ‘competing’ lineages over a range of time.

The Relate pipeline in this thesis was implemented for two key steps:

1. To infer the effective population size (N_e) of the Yakut population.
2. To perform the selection test, which yielded results likely informative of selection signatures at specific genes within the studied population.

3.4.1 N_e estimation

For these purposes, the genomic dataset had to be converted to .haps and .sample formats from the VCF file format. The data were phased as required for constructing the Relate input. In its first five columns, the .haps file gives information on the chromosome number, the SNP ID and its position, and the ancestral (represented by 0) and the alternative allele (represented by 1). The successive column pairs correspond to the haplotypes carried by an individual at that SNP. Meanwhile, the sample file is composed of three columns, first individual ID, second individual ID and missing data proportion. Several cycles were run on all the 22 autosomal chromosomes, starting with ones creating input files. The --ancestor and --mask flags were used to call the .ancestor and .mask files, which are both fasta format, the first one containing ancestral genome and the latter of same length as the ancestral genome containing a genomic mask. The genomic mask needs to be specified as a fasta file of same length as the

ancestral genome, such that the x^{th} entry is the x^{th} base of the genome. In the mask file format loci passing the mask are denoted by P, loci not passing the mask are denoted by N. In the end three different files are obtained, all filtered and zipped. The output files prepared in this cycle represent the input files for the next one. If necessary, haplotypes were flipped according to ancestral alleles (using the ancestor fasta file as a reference), meaning they were standardized with the proper 0/1 notation explained above, SNPs were filtered using the genomic mask file and distances between SNPs were computed.

Subsequently, .anc and .mut files were generated to infer the genealogy of each SNP. These file formats are a stepping stone for the effective population size estimation and for the selection test. The coefficients and functions used here were the mutation probability -m, with a value of $1.25e^{-8}$ per generation, the -N flag set at 30,000, which is a standard value for the effective population size (N_e) of the *Homo sapiens* species. The previously outputted and g-zipped .haps, .sample and .dist files were called together with a .map file, which contains a genetic recombination map. The .anc file contains tree information including branch lengths, while the .mut file stores SNP-specific information. Finally, the command --seed for the random number generator used in the MCMC algorithm was implemented for branch length estimation.

The final step for N_e estimation was run remotely using the nohup command and the already provided EstimatePopulationSize.sh instruction script was employed. A .poplabel file was created and exploited. This file format is composed of four columns: individual ID as specified in the .sample file, which must also be listed in the same order as in the .sample file, population label, group label and sex. Another newly introduced flag was --threshold to which a value from the interval [0,1] can be assigned, which is the fraction of trees to be dropped in order to adjust the precision and specificity of the mutation mapping and coalescence inference. In this case the value 0 was assigned, leaving all the information available in use.

3.4.2 Selection test and p-value computation

Once the N_e fluctuations have been estimated, p-values for the selection test implemented in Relate were calculated from the reconstructed genealogy. The previously created popsize file was taken as an input file, the mutation rate -m was once again set to $1.25e^{-8}$ and the flag --years_per_gen was used and set to 29 years. A new .sele file was created. This file format is one of the three types of output files in Relate used to detect signatures ascribable to the action of natural positive selection. The file format .sele in particular detects and stores the $\log_{10}$ p-value for selection evidence at each generation throughout the specified range. Of key importance is the "when_mutation_has_freq2" column that contains the p-value for selection evidence over the lifetime of a mutation of interest. The columns containing significant p-values of interest were selected and moved to a new file to be processed using R Studio, where the top 1% and the top 0.1% of the p-values were filtered and properly aligned. Relate significant results indicative of selection signatures were identified as those SNVs associated with a $\log_{10}$ p-value within the top 1% and/or 0.1% of the distribution of $\log_{10}$ p-values obtained across the Yakut genomes.

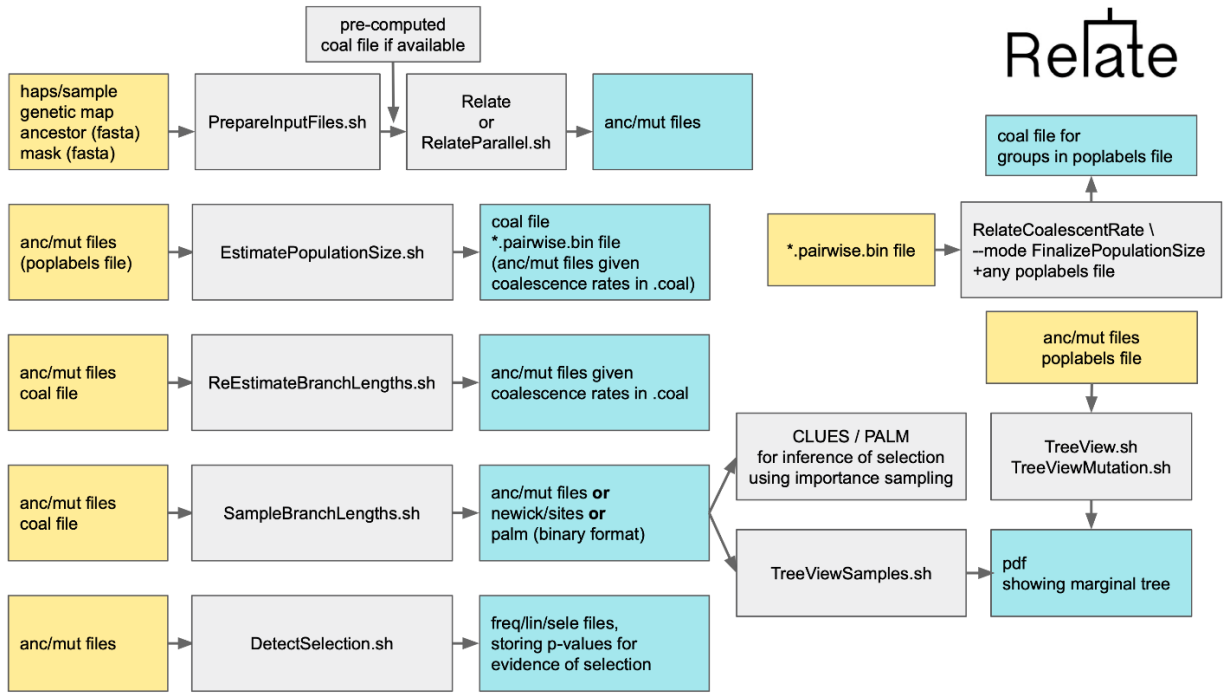


Figure 3.3 A scheme showing the methodological steps of the Relate pipeline of analyses. The left side of the figure captures the process covered in this thesis, starting with the .haps and .sample files that were gradually transformed first into .anc and .mut files and in the end to a .sele file carrying the final information on the inferred evidence of positive selection. The figure has been taken from <https://myersgroup.github.io/relate/>.

3.5 Crosscheck with previous results from LASSI-signet analyses

The primary aim of the present study was that of further validating the results presented in the thesis by Pognant Viù (2024), which were obtained after the application of an integrated approach on Yakut WGS data aimed at detecting genomic signatures ascribable to the polygenic model of adaptation. Such an integrated approach consisted in the combination of two statistical methods LASSI (Harris & DeGiorgio 2020) and signet (Gouy et al. 2017), which together allowed the identification of multiple gene networks characterized by weak signatures of natural selection and functional connections within pathways presumably related to cold-induced biological responses, such as those relative to lipid metabolism and thyroid hormone signaling processes. Collectively, these results are consistent with a polygenic adaptation model and may identify part of the genetic bases underlying the biological adaptations that have evolved in high-latitude Siberian populations to thrive in extremely cold conditions.

The Harris & DeGiorgio (2020) LASSI package based on a likelihood T-statistic methodology was implemented by Pognant Viù (2024), in order to explore the occurrence of potential selective events in the Yakut gene pool, distinguishing them in hard (i.e. very strong) and soft (i.e. weak-to-moderate) selective sweeps. The concept of selective sweep in evolutionary biology explains how a favored mutation establishes itself within a population and, as a consequence of the inevitable linkage disequilibrium (LD), causes a reduced genetic diversity in the surrounding genomic region (Turner-

Hisson et al., 2020). There is a difference between a hard selective sweep, where a single allele becomes fixed in the population, and a soft selective sweep, where a gene bears and fixes multiple beneficial alleles. The likelihood statistic used by Pognant Viù (2024) sets two competing and contrasting explanations for observing a specific event: an alternative hypothesis (i.e., H_1) stating the occurrence of directional positive selection, conflicted by the null hypothesis (i.e., H_0) supporting the idea of neutral evolution.

Phased data from the HGDP version GRCh38 (Bergstrom et al. 2020) were employed and filtered for Yakut individuals using the BCFtools package (Danecek et al., 2021). The average number of SNVs in a haplotype block was then estimated with the PLINK --blocks function followed by the mandatory QC procedures. The LASSI T-statistic was then calculated for all genomic windows detected across the genomic sequences and the entire distribution of the statistical values informative of soft selective sweeps was used to inform the signet algorithm. This latter method is capable of identifying gene networks collectively characterized by soft-sweep like genomic signatures and functional relations within the same biological pathway, thus identifying genomic patterns plausibly ascribable to the polygenic model of adaptation (Ferraretti, G. et al. 2025). Figure 3.4 shows some of the most significant results obtained after the application of these analyses on the Yakut genomes, which were subjected in the present thesis to a crosscheck with Relate significant results in order to provide a further validation of the selection signatures identified for these genes.

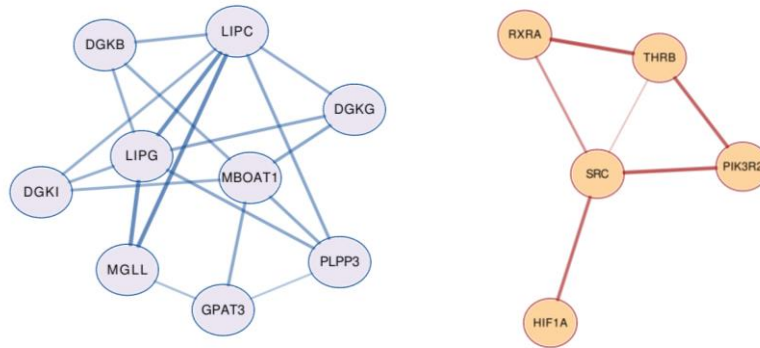


Figure 3.4 Gene networks generated using the Cytoscape software and resulting by LASSI-signet analyses performed by Pognant Viù (2024). The left figure shows the interactions (i.e. edges) between the genes (i.e. nodes) belonging to the Glycerolipid metabolic pathway, with a p-value of 0.021. The right figure represents the gene-network retrieved within the Thyroid hormone signaling pathway, with a p-value of 0.047. These two pathways resulted to be some of the most significant ones in the Yakut genomes and were interpreted as likely involved in the modulation of their metabolic adaptations to cold climate. The two figures were presented in Pognant Viù (2024).

4. RESULTS

4.1 Quality Control Procedures

The Quality Control (QC) filtering steps that were performed on the 25 assembled Yakut WGS from the HGDP panel at the very beginning of the study yielded a dataset ready for further processing. In these QC procedures, by implementing functions of the PLINK 1.9 toolset all the SNVs with more than 5% missing data were excluded, along with those presenting potential sequencing errors (i.e. significantly deviating from the Hardy-Weinberg Equilibrium), and those characterized by ambiguous alleles. These procedures finally resulted in a WGS dataset including 25 Yakut individuals typed for 6,691,957 high-quality SNVs. This filtered dataset was then redirected to the Relate pipeline of analyses to produce the results that are described in the following sections.

4.2 Implementation of the Relate pipeline of analyses

Relate successfully reconstructed the genealogical history of each SNV observed in the Yakut genomes and revealed genomic signatures ascribable to the action of positive selection. By employing the MCMC algorithm, Relate first generated a distance matrix based on which a binary tree was built leading to the estimation of changes in the effective population size (N_e) of the Yakut population across generations. This parameter has been inferred because it is necessary to inform the subsequent steps of analysis aimed at detecting genomic signature ascribable to the action of natural selection. As for proving the occurrence of selective events, a .sele file was indeed generated which, most importantly, contained the "when_mutation_has_freq2" column providing the p-value for the selection test performed for each of the considered mutations.

4.2.1 Effective Population Size Estimation

Results from the effective population size estimation procedure were summarized in the graph shown in Fig.4.1. In particular, the plot shows very accurate estimations considering a large scale perspective informative of the overall demographic history of the ancestral human populations from which the Yakut have originated, also revealing the presence of some interesting more recent demographic events that have been experienced by the immediate Yakuts' ancestors.

According to the obtained plot, four major bottleneck events can be inferred, of which the more recent two appeared to be dated after the *Homo sapiens* lineage diversification. The most prominent of them all is the bottleneck effect occurring between 70,000 and 60,000 YBP (years before present), which is consistent with literature findings informative of a decline in the effective population size of most non-African human groups as consequence of the "Out of Africa" dispersal of early humans (McEvoy et al. 2011). Based on archaeological records, the first modern humans are thought to have settled in South Siberia during the Early Upper Paleolithic period approximately 30,000–40,000 YBP and later having colonized the mammoth-steppes of subarctic Siberia by means of various behavioral adaptations (Zlojutro et al. 2009). These data are in line with a growing effective population size of a well-established

population as observable in the graph in the range corresponding to such a temporal framework. Then, around 10,000 YBP, the Turkic lineage to which the Yakut belong, more particularly the South Siberian populations (Zlojutro et al. 2009) seem to have experienced a new peak in the effective population size as clearly reported in the plot. Most populations throughout Europe and East Asia in that period saw their population sizes reaching a peak like this, and the one of the Siberian Yakut being geographically isolated is even considered to be not as prominent in comparison (McEvoy et al. 2011). Since this timepoint, the graph shows a period of fluctuations of the effective population size with the occurrence of a gradual but significant bottleneck effect. Then, around 1,500 YBP the effective population size started to recover in a very similar gradual manner until it abruptly expanded to around 4,000 estimated individuals remaining stable until present days.

Overall, such results are in accordance with the findings by Zlojutro et al. (2009), which were based on the analysis of Yakut mtDNA variations and aimed at determining their genetic coalescence time and possible historical admixtures with other populations. The authors suggested that although the Yakut migration to the north took place no earlier than the 11th to 13th centuries AD (900-700 YBP), there has been a defined founder event approximately 1,500 to 1,000 YBP, that can be explained by large-scale population movements of the Turkic populations of South Siberia, which might have posed as a migratory pressure for the Yakut within the region. Further on, these findings suggest that the Yakut founder event followed by the migration to North East Siberia involved a significantly small female population of possibly just 150 individuals, who then experienced a sudden gene flow from surrounding indigenous populations. This is supported by the very relevant differences in the diversity of the HVS-I mtDNA sequences and Y-STR haplotypes, the first one having a much higher value, and the latter one being pretty low even in comparison to other Siberian populations. Such a genetic profile is believed to be the result of cultural behavior (Zlojutro et al. 2009).

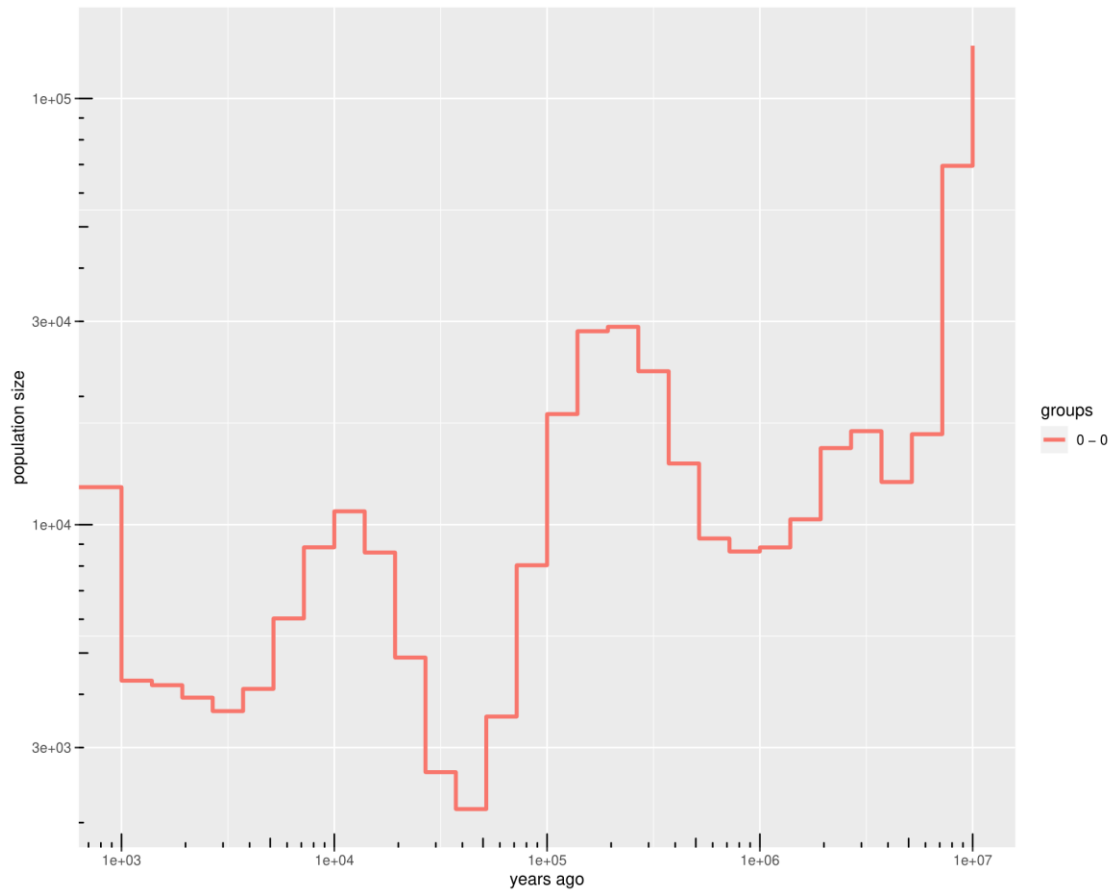


Figure 4.1: Relate-based estimates of the effective population size (N_e) of the Yakuts' ancestors, tracing their earliest genealogical history. Moving from right to left, the sharp fluctuations in N_e can be followed up until present days.

4.2.2 Implementation of a Selection Test

The results obtained after the implementation of the Relate Selection test are summarized in Figure 4.2. Specifically, a Manhattan Plot was built using the R Software, showing 11 different genes and the position of each corresponding SNV with its respective p-value. A Manhattan plot is considered as a fine way to represent genome-wide association studies (GWAS) results through a scatter plot where the x-axis orders the considered SNVs by gene location, while the y-axis shows the negative logarithm of association p-values, and as a result, the highest peaks represent the most significant hits (Liang 2013). In the present thesis, the Manhattan plot was generated by using the p-values returned by the Relate selection test, thus being informative of the possible selective events experienced by the Yakut population. Three significant thresholds were set on the y-axis; a nominal one (i.e. p-value < 0.05; dashed red line), one reporting the -log p-values falling within the top 1% of the overall distribution of p-values obtained (i.e. red line) and the third one showing the top 0.1% threshold (i.e. blue line). The putative adaptive genes selected to be reported in the Manhattan plot are those involved in the thyroid, lipid and

inflammatory signaling pathways, which emerged from previous analyses described in the Methods section and which concern selection signatures likely ascribable to the polygenic model of adaptation. Particularly, the present thesis will focus on the further description of the genes characterized by p-values within the third threshold: *PLPP3*, *ERBB4* and *PRKG1*. These loci are involved in cardiovascular and lipid signaling pathways, whose pivotal role in the polygenic adaptive traits evolved by the ancestors of the Yakut population will be further explored in the *Discussion* section.

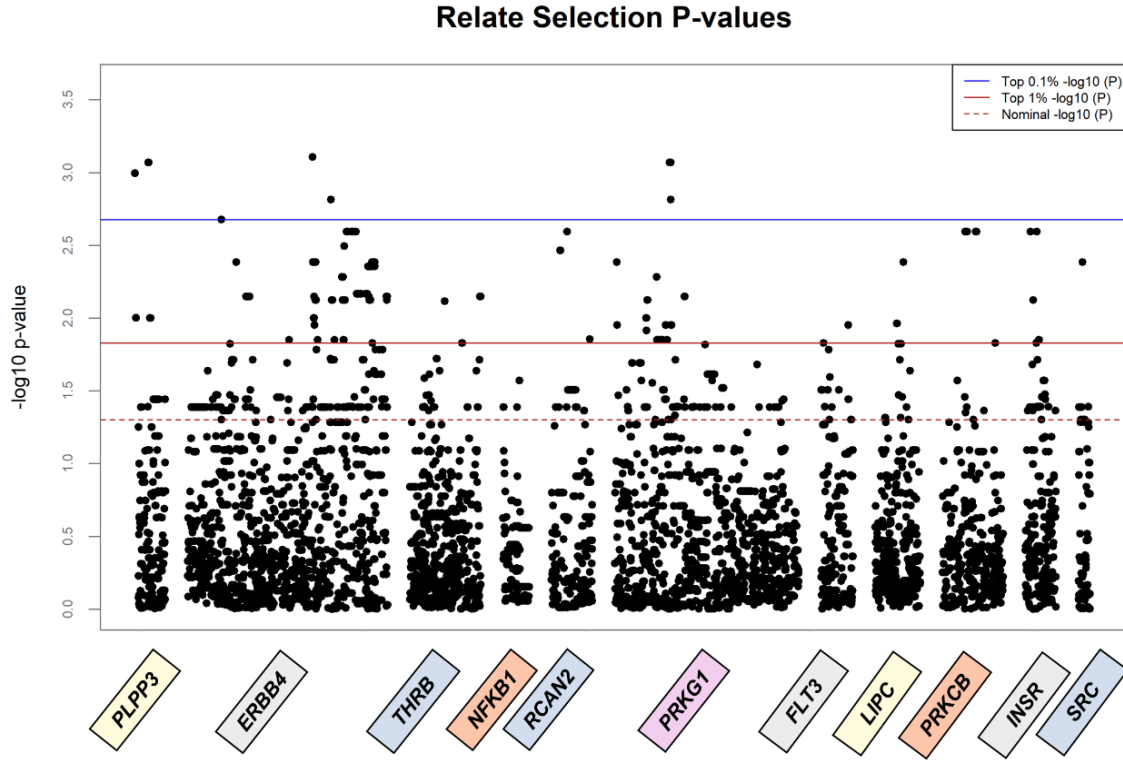


Figure 4.2: Manhattan plot showing the Relate selection test p-values for SNVs within the reported genes. The horizontal lines separate the significant nominal, top 1% and 0.1% thresholds of the outlier SNVs. The figure was produced by exploiting the *plot* function implemented in the R software.

5. DISCUSSION

The three genes that turned out to be the most interesting putative adaptive genes pointed out by this study were selected according to their p-values, all of them falling within the third significance threshold adopted (i.e. top 0.1% of the p-value distribution obtained). These genes, which are *PLPP3*, *ERBB4* and *PRKG1*, belong to different, however tightly connected, metabolic pathways. In particular, the *PI3K-Akt signaling* pathway directly overlaps with the *AGE-RAGE signaling pathway in diabetic complications* and the *Thyroid hormone signaling* pathway, and in all of them the mentioned genes of interest are known to play crucial roles (Kanehisa & Goto 2000). The *AGE-RAGE signaling* pathway, among else, promotes expression of pro-inflammatory cytokines and a variety of atherosclerosis-related genes, and within its network via the receptor for advanced glycation end products (RAGE) induces PI3K-Akt-dependent pathways, which in turn participate in the regulation of cell proliferation and apoptosis. Thus, these intertwined nets of functional interactions might explain how these three genes participated in the evolution of complex polygenic adaptive traits by the ancestors of the Yakut population.

5.1 *PLPP3*

When the pathophysiological role of *PLPP3* was tested in mice, a very clear correlation was revealed between *PLPP3* polymorphisms and CAD (coronary artery disease) susceptibility. In mice subjected to hepatic *PLPP3* gene deletion, the lipid profile of the blood concerning the levels of several plasma lipids involved in the development of atherosclerosis, including lactosylceramides, lysophosphatidic acids, and lysophosphatidylinositols, was significantly altered (Busnelli, M. *et al.* 2016). Even though *PLPP3* does not exhibit a direct involvement in the development of obesity and type diabetes 2 (TD2) pathways, its regulatory role in the glycolipoid metabolic pathway and direct connection to vascular pathologies like CAD and atherosclerosis would certainly impact any already existing condition of the aforementioned. Further plasma lipidomic analysis indicated low-abundant lipid species as potential drivers of the development of atherosclerosis associated with *PLPP3* hepatic deficiency. An additional study (Jose & Kienesberger 2021) on the impact of *PLPP3*, or rather cardiac *PLPP3* deficiency in mice, explored the key role of *PLPP3* for mitochondrial functioning and energy homeostasis in the heart. The results of the study therefore suggested that the adult-onset disruption in *PLPP3* function could accelerate a cardiac-metabolic and functional decline in the setting of diet-induced obesity.

5.2 *PRKG1*

The *PRKG1* gene plays an essential role in regulating multiple and diverse functions including cardiovascular and neuronal ones, such as smooth muscle tone, also preventing platelet aggregation and modulating cell growth, all of it through the nitric oxide/cGMP signaling pathway. Although without direct experimental data, multiple studies (Cardona *et al.* 2014) have proposed variants at *PRKG1* as

having played a central role in the evolution of biological adaptation in Siberian populations, where efficient cardiovascular regulation is required to tackle the increase of cardiac pressure due to cold exposure. Furthermore, in the study by Ferraretti et al. (2026), the importance of *PRKG1* in cell differentiation in brown adipose tissue (BAT) was emphasized together with experimental results inferring a strong correlation between the absence of the gene and lower mitochondrial content and reduced amounts of adipogenic factors in mouse BAT and even lower body temperature in wild-type mice (Haas, B. et al. 2009).

5.3 *ERBB4*

The evolution of the *ERBB4* gene and its involvement in the development of cold-induced adaptive responses in Yakuts is believed to have been to some extent shaped by an archaic introgression event, which was due to gene flow from the Neanderthal species into the ancestral populations from which Yakuts have originated (Ferraretti et al. 2026). This same study also highlighted the well-known significance and active participation of the *ERBB4* gene in adaptive thermoregulation, in accordance to which is the inhibition of BAT thermogenesis and metabolic functioning upon suppression of the expression of the aforementioned gene. Interesting observations were also reported by Santiago-Marrero et al. (2023), who focused on the role of *ERBB4* in the brain to control energy expenditure and body weight. The deletion of *ERBB4* in the paraventricular nucleus of the hypothalamus (PHV) in mice indeed resulted in increased body weight and fat mass, lower energy expenditure and decreased BAT thermogenesis. Another intriguing correlation was inferred, showing that *ERBB4* is also expressed in oxytocin neurons of the PVH and its deletion leads to reduced serum oxytocin levels. The final conclusion of the study was indicative of *ERBB4* exerting a pivotal regulatory role in the PVH likely through regulation of oxytocin expressing neurons, and revealing a novel function of *ERBB4* in the pathophysiological mechanisms of depression-associated obesity.

5.4 Cold-induced metabolic adaptations

As it can be inferred from the previous paragraphs, BAT plays a pivotal role in thermoregulatory metabolic pathways. In relation to this, Yakut young adults exhibited greater evidence of BAT thermogenesis in response to mild cooling in comparison to young adults from the area of Chicago. In line with this, Yakut subjects displayed a stronger relationship between BAT thermogenesis and change in energy expenditure, exhibiting greater metabolic response to cold stress and maintaining warmer sternum temperatures (Stephanie et al. 2022).

Another study exploring the effects of partial and whole-body cryotherapy (PBC and WBC) emphasized the beneficial effects of WBC, which in our case could mimic the extremely cold Siberian environment. One of the very first WBC effects to be described was the increased release of endorphins in the body

and relief of chronic pain problems, making the beneficial WBC effects comparable to the ones from physical exercise (Fontana et al. 2024).

Among the other advantages of cold exposure in the form of WBC, it is found the shift in the hormones fibroblast growth factor 21 (*FGF21*) and irisin. In particular, *FGF21*, acts extracellularly activating the PI3K-Akt signaling pathway and has a strong regulatory effect on lipid, glucose and energy metabolism. *FGF21* promotes glucose uptake in fat, whereas in the liver, it stimulates gluconeogenesis and decreases lipolysis. Irisin, on the other hand, is a hormone secreted from skeletal muscles in response to exercise, mediating its beneficial effects such as weight loss and thermoregulation, boosts thermogenesis in humans and plays a role in exercise-induced weight loss. Increased levels of irisin have been shown to lead to browning of white adipose tissue (WAT), further improving adiposity and glucose homeostasis as it has a stimulatory effect on p38 and ERK signaling, enabling UCP-1 expression in adipocytes (Fontana et al. 2024). These effects match those described by Chen & Guo (2025) in their paper focused on the protective effects of exercise on diabetic cardiomyopathy.

However, with increasingly sedentary lifestyles, globalization of the nutritive impoverished diet and altered mean temperatures, all these adaptive traits in such tight connection to the Glycerolipid metabolic pathway and adipose tissue disposal, can potentially increase the Yakuts' susceptibility of developing diabetes and obesity. As major factors for obesity remain environmental, ethnological studies have already presented clear proof that there is a genetic origin for susceptibility to many metabolic diseases. Impaired insulin secretion is in the central spot of the subject as it is the main and an irreversible driver to T2DM, which was shown to have further correlation to obesity (Felber & Golay 2002).

6. CONCLUSION

This study was based on the implementation of bioinformatic tools to investigate patterns of adaptive genomic variation of the Yakut population. In particular, the Relate software was used to process 25 Yakut WGS and the results that have been obtained by means of such an approach were then cross-checked with previous findings obtained by means of different methods aimed at uncovering genomic signatures putatively ascribable to polygenic adaptations.

The results pinpointed by such a crosscheck were further explored through literature research. Overall, these methodological strategies enabled to gain a deeper understanding into the evolutionary history of the Yakut people and to identify the genes which were significantly impacted by the action of natural selection in targeted metabolic pathways putatively involved in the development of polygenic adaptations and which have likely enabled this population to thrive in the extreme cold.

The results obtained in the present study were in line with findings previously described for this population. The graph showing the effective population size fluctuations experienced by the Yakut ancestors, which was yielded by Relate with an input from the human reference sequence GRCh38 together with Genomic mask and Ancestral genome files generated beforehand by the software, described the historical fluctuations of the effective population size of the Yakut within a large time span (depicting a range of evolutionary history from 10^7 YBP and on). Significant bottle-neck effects were easily identifiable, the major of which could be traced back to well-studied demographic events like the “Out of Africa” migration of ancestral *Homo sapiens* populations and later in relation to the migration of the South Siberian populations to the north of the region.

The second set of findings obtained with the Relate approach concerned signatures of natural positive selection identified for multiple genes in the Yakut gene pool. Particularly, highly significant results for the Relate selection test were identified for the *PLPP3*, *PKRG1* and *ERBB4* genes. The functional contextualization of these results was performed by further exploring the roles played by these genes in the context of complex biological adaptations evolved by the Yakuts to cope with cold climates. In details, it was conducted through database and literature research.

Although having distinct functions, these three genes intertwine in several metabolic pathways, such as PI3K-Akt signaling pathway, Thyroid hormone signaling pathway, Glycerolipid metabolism and the AGE-RAGE signaling pathway in diabetic complications. The pervasive selection processes occurred at these genes may have led to the evolution of a distinct physiology in the Yakut population, including a greater BAT specialization and activity.

Nevertheless, this powerful set of molecular-level physiological adjustments is deemed to be a subject of an evolutionary mismatch in present-day conditions. The stimuli which originally led to increased BAT levels in the body are now lost and/or substituted. Furthermore, the prolonged extreme cold is undergoing a major climate change with constantly increasing temperatures. The traditional North-East Siberian diet has been recently lost in favor to a post-industrial fast food diet and so the adipocytes are pushed into the process of ‘whitening’, converting the tissue into a white adipose tissue (WAT). The Yakut evolutionary (i.e. genomic) background puts the Yakut in the center of today’s diabetes and obesity ‘epidemic’ and as such is a subject of an issue that requires even greater attention than before.

REFERENCES

- 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Andrews, C. (2010) The Hardy-Weinberg Principle. *Nature Education Knowledge* 3(10):65
- Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., Durbin, R., ... Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science (New York, N.Y.)*, 367(6484), eaay5012. <https://doi.org/10.1126/science.aay5012>
- Busnelli, M., Manzini, S., Hilvo, M., Parolini, C., Ganzetti, G. S., Dellera, F., Ekroos, K., Jänis, M., Escalante-Alcalde, D., Sirtori, C. R., Laaksonen, R., & Chiesa, G. (2017). Liver-specific deletion of the Plpp3 gene alters plasma lipid composition and worsens atherosclerosis in apoE^{-/-} mice. *Scientific reports*, 7, 44503. <https://doi.org/10.1038/srep44503>
- Cardona A, Pagani L, Antao T, Lawson DJ, Eichstaedt CA, et al. (2014) Genome-Wide Analysis of Cold Adaptation in Indigenous Siberian Populations. *PLOS ONE* 9(5): e98076. <https://doi.org/10.1371/journal.pone.0098076>
- Chen, H., & Guo, L. (2025). Exercise in Diabetic Cardiomyopathy: Its Protective Effects and Molecular Mechanism. *International journal of molecular sciences*, 26(4), 1465. <https://doi.org/10.3390/ijms26041465>
- Crubézy, É., Guarino-Vignon, P., Seguin-Orlando, A., Der Sarkissian, C., Hanghøj, K., Duchesne, S., Gérard, P., Thèves, C., Alcouffe, A., Romanova, L., Nikolaeva, D., Alekseeva, L., Hochstrasser-Petit, C., Zvenigorosky, V., Keyser, C., Ludes, B., Petit, M., Dabernat, H., Géraut, A., Jyrkov, E., ... Orlando, L. (2026). An ancient DNA perspective on the Russian conquest of Yakutia. *Nature*, 650(8101), 389–398. <https://doi.org/10.1038/s41586-025-09856-5>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), giab008. <https://doi.org/10.1093/gigascience/giab008>
- Felber, J. P., & Golay, A. (2002). Pathways from obesity to diabetes. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, 26 Suppl 2, S39–S45. <https://doi.org/10.1038/sj.ijo.0802126>
- Ferraretti, G., Rill, A., Abondio, P. et al. Convergent evolution of complex adaptive traits modulates angiogenesis in high-altitude Andean and Himalayan human populations. *Commun Biol* 8, 377 (2025). <https://doi.org/10.1038/s42003-025-07813-6>

- Ferraretti, G., Alberti, M., Cicolini, R. *et al.* Adaptive evolution at thyroid hormone, insulin and glycerolipid pathways improved energy metabolism in high-latitude Eurasian populations. *Commun Biol* (2026). <https://doi.org/10.1038/s42003-026-10540-1>
- Felber, J. P., & Golay, A. (2002). Pathways from obesity to diabetes. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, 26 Suppl 2, S39–S45. <https://doi.org/10.1038/sj.ijo.0802126>
- Fontana, J.M., Dugué, B., Capodaglio, P. (2024). Prolonged or Repeated Cold Exposure: From Basic Physiological Adjustment to Therapeutic Effects. In: Capodaglio, P. (eds) Whole-Body Cryostimulation. Springer, Cham. https://doi.org/10.1007/978-3-031-18545-8_1
- Gouy, A., Daub, J. T., & Excoffier, L. (2017). Detecting gene subnetworks under selection in biological pathways. *Nucleic acids research*, 45(16), e149. <https://doi.org/10.1093/nar/gkx626>
- Haas, B., Mayer, P., Jennissen, K., Scholz, D., Berriel Diaz, M., Bloch, W., Herzig, S., Fässler, R., & Pfeifer, A. (2009). Protein kinase G controls brown fat cell differentiation and mitochondrial biogenesis. *Science signaling*, 2(99), ra78. <https://doi.org/10.1126/scisignal.2000511>
- Harris, A. M., & DeGiorgio, M. (2020). A Likelihood Approach for Uncovering Selective Sweep Signatures from Haplotype Data. *Molecular biology and evolution*, 37(10), 3023–3046. <https://doi.org/10.1093/molbev/msaa115>
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Jose, A., & Kienesberger, P. C. (2021). Autotaxin-LPA-LPP3 Axis in Energy Metabolism and Metabolic Disease. *International journal of molecular sciences*, 22(17), 9575. <https://doi.org/10.3390/ijms22179575>
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), 27–30. <https://doi.org/10.1093/nar/28.1.27>
- Levy, S. B., Klimova, T. M., Zakharova, R. N., Fedorov, A. I., Fedorova, V. I., Baltakhinova, M. E., Bondy, M., Atallah, D., Thompson-Vasquez, J., Dong, K., Debertine, A., & Leonard, W. R. (2022). Brown adipose tissue thermogenesis among young adults in northeastern Siberia and Midwest United States and its relationship with other biological adaptations to cold climates. *American journal of human biology : the official journal of the Human Biology Council*, 34(6), e23723. <https://doi.org/10.1002/ajhb.23723>
- McEvoy, B. P., Powell, J. E., Goddard, M. E., & Visscher, P. M. (2011). Human population dispersal "Out of Africa" estimated from linkage disequilibrium and allele frequencies of SNPs. *Genome research*, 21(6), 821–829. <https://doi.org/10.1101/gr.119636.110>

- Pakendorf, B., Novgorodov, I. N., Osakovskij, V. L., Danilova, A. P., Protod'jakonov, A. P., & Stoneking, M. (2006). Investigating the effects of prehistoric migrations in Siberia: genetic variation and the origins of Yakuts. *Human genetics*, 120(3), 334–353. <https://doi.org/10.1007/s00439-006-0213-2>
- Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics*, 2(12), e190. <https://doi.org/10.1371/journal.pgen.0020190>
- Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J., & Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* (Oxford, England), 36(16), 4449–4457. <https://doi.org/10.1093/bioinformatics/btaa520>
- Pognant Viù, Sabrina Thriving in the Cold: Exploring Biological Adaptations of the Yakut Population to a Continental Subarctic Climate, Department of Biological, Geological and Environmental Sciences - BiGeA Master's degree in Global Change Ecology and Sustainable Development Goals
- Santiago-Marrero, I., Liu, F., Wang, H., Arzola, E. P., Xiong, W. C., & Mei, L. (2023). Energy Expenditure Homeostasis Requires ErbB4, an Obesity Risk Gene, in the Paraventricular Nucleus. *eNeuro*, 10(9), ENEURO.0139-23.2023. <https://doi.org/10.1523/ENEURO.0139-23.2023>
- Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, 51(9), 1321–1329. <https://doi.org/10.1038/s41588-019-0484-x>
- Turner-Hissong, S. D., Mabry, M. E., Beissinger, T. M., Ross-Ibarra, J., & Pires, J. C. (2020). Evolutionary insights into plant breeding. *Current opinion in plant biology*, 54, 93–100. <https://doi.org/10.1016/j.pbi.2020.03.003>
- Zlojutro, M., Tarskaia, L. A., Sorensen, M., Snodgrass, J. J., Leonard, W. R., & Crawford, M. H. (2009). Coalescent simulations of Yakut mtDNA variation suggest small founding population. *American journal of physical anthropology*, 139(4), 474–482. <https://doi.org/10.1002/ajpa.21003>