UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea Magistrale in Molecular Biology

TESI DI LAUREA

The genetic landscape of Kyrgyzstan: admixture and genetic history of a population from Central Asia

Relatore: Dr. Massimo Mezzavilla
Dipartimento di Biologia

Laureanda: Anna Carolina E. L. Tanada

ANNO ACCADEMICO 2022/2023

INDEX

1. ABSTRACT

Kyrgyzstan was crossed by the ancient Silk Road, being a melting point of cultures, languages, and genes, and later it was invaded by different tribes and groups. This complex history may be reflected in their genome and further understanding is essential for understanding the genetic landscape of this region. To achieve it, 184 Kyrgyz SNP genotype data were analysed through Principal Component Analysis, Runs of homozygosity, and linkage disequilibrium pattern to describe the level of population structure and level of inbreeding. To estimate the level of present and past gene flow from neighbouring regions, allele frequency-based methods were performed. Then we estimated the effective population size through time using IBDNe. Considering the level of admixture and the possible different ancestries present in Kyrgyz's genomes it was also used AS-IBDNe to estimate the ancestry-specific effective population size. Finally, signatures of local adaptation were searched using haplotype-based scans to detect positive selection in the Kyrgyzstan population. The results showed that different patterns of admixture exist within the population, and most of the individuals are an admixture between Russian and Han Chinese ancestral populations. Furthermore, evidence of recent ongoing selection in genes related to the Behçet syndrome, the "Silk Road disease", was found.

2. INTRODUCTION

2.1. Silk Road

It is widely acknowledged that humans are wanderers beings, conquering lands and tracing roads all around the world. More than wanderer beings, humans are also social beings, therefore interactions between populations are extremely common and occurred abundantly throughout history. With the emergence of civilizations and the advancement of technology, these interactions have become more complex, and eventually, the creation of trade routes was needed.

One of the most important trade routes of history was the ancient Silk Road, which existed from the second century BCE until the 15th century CE. It was an extensive trading route that connected all of Eurasia, and its main purpose was transporting goods, especially silk produced in China, hence the name. The trading of goods also led to the trading of cultures, food, people, and hypothetically also, genes. Furthermore, the extensive period of the Silk Road comprehends major historical events such as the Black Plague and the Mongol Conquest, events that doubtless impacted the demography and genome of the populations that inhabited its surroundings. Moreover, it was shown that the ancient Silk Road crossed what was a principal geographical area of human expansion after the Out of Africa (Maca-Meyer et al. 2001), which emphasizes the importance of this region.

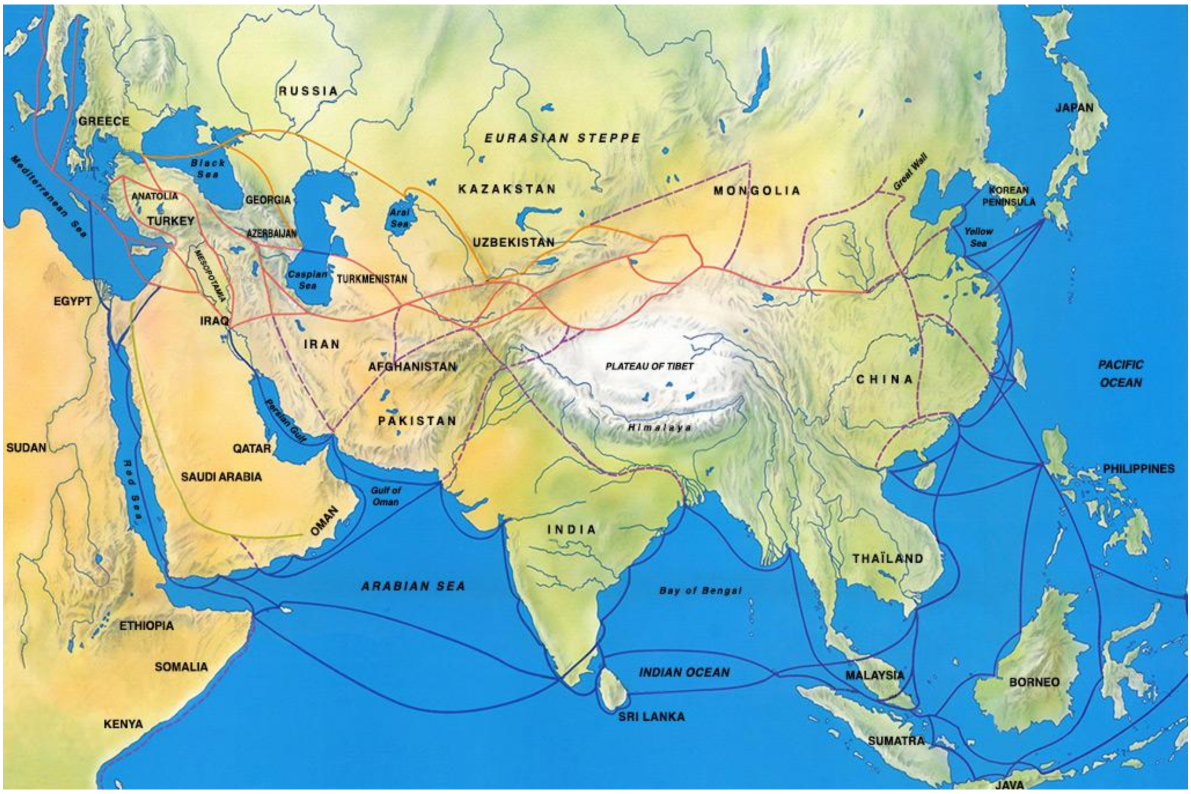


Figure 1. Ancient Silk Road Map. Map illustrating the main routes of the Ancient Silk Road across the whole Eurasia and surrounding areas. Figure obtained from the official UNESCO website, available at: https://en.unesco.org/silkroad/silkroad-interactive-map

Previous studies have shown through mitochondrial DNA (mtDNA) that populations of the Silk Road have intermediate features to West and East Eurasia, which is strong evidence of admixture between Europe and East Asia populations (Comas et al. 1998; Yao et al. 2000). More specifically, a study demonstrated that most of the populations received a contribution of 50% from Europe and Asia approximately, and that admixture with Asian populations probably occurred around 25 generations ago (Mezzavilla et al. 2014).

The ancient Silk Road induced a great transit of people within Central Asia, bringing people from Europe and East Asia together. This raises the question of whether genes were also “traded” during this period, and whether part of the genetic diversity seen in Central Asia can be accredited to it. Previous studies helped elucidate this question, and yet a deeper analysis of each population is still needed to fully comprehend Central Asia's genome and history.

2.2. Kyrgyzstan

Central Asia is composed of five recognized countries: Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, and Uzbekistan, and includes Indo-Iranian and Turkic speaker populations. It is an ample region in Asia that extends from Russia in the North to Afghanistan in the South, and from Mongolia in the East to the Caspian Sea in the West. As mentioned before, this region was once crossed by the ancient Silk Road, and one of its main trading routes was located exactly in Kyrgyzstan.

Kyrgyzstan, one of the countries in Central Asia, is bordered by Kazakhstan, Uzbekistan, and China. The country is inhabited mainly by Kyrgyz people, a Turkic ethnic group. It has a population of 7.1 million people according to the official 2022 census (available at http://www.stat.kg/en/), and its official languages are Kyrgyz and Russian.

Kyrgyzstan has a very intricate history, of which previous studies demonstrate evidence of pastoralists that occupied the region by the early Bronze Age (Taylor et al. 2018; Motuzaite Matuzeviciute et al. 2020). It is also known that Kyrgyzstan faced the Turko-Mongol expansion during the medieval period, and was later invaded by the Russian Empire after the 17th century (Adle, Habib, and Baipakov 2003).

All of these historical events that occurred within the region probably are reflected in their genetic and demographic history. Previous studies about Central Asia showed that the ancestors of Kyrgyz people seem to have merged with other local tribes such as Mongol, Uygur, Khitan, and Han (Peng et al. 2018). Therefore, it is expected that modern-day Kyrgyz individuals present some sort of admixture within their genome. The complex anthropological history of Kyrgyzstan stimulated further research to better describe the genetics of Kyrgyz people, and they showed a close genetic relationship with Kazakh and Uyghur populations (Guo et al. 2018; Chen et al. 2019). Nonetheless, further studies are still needed to fully understand Kyrgyz's genetic and demographic history.

2.3. Aim of the Study

“Why we are here” and “Where did we come from” are questions that perpetuate humanity for millennia. We may never have the answer to the first question, however, population genetics can help unveil the mystery of the past at the same time it can open the doors to the future. Answering this question is not a trivial task, after all, it is widely acknowledged that humans are social beings. Therefore interactions between populations are extremely common and occurred abundantly throughout history, usually with little or no physical archaeological evidence of it.

By analyzing the genome of several individuals from different populations, it is possible to have insights into past migration flows and their courses, the interaction between populations, and investigate natural selection signatures in the genome. Therefore, population genetics is crucial to not only studying human origins and evolution but also to increasing life quality through medical breakthroughs (Jorde, Watkins, and Bamshad 2001).

The goal of this study is to determine the population structure of Kyrgyzstan, investigate the presence of admixture events, and also analyze if there is any evidence of positive ongoing selection using Single-Nucleotide Polymorphism (SNP) data. SNPs are a one-nucleotide variation in a given location that naturally occurs within a population with a frequency higher than 1% in a population, being usually bi-allelic in humans (Brookes 1999). This is a normal and the most abundant variation of the genome, therefore insertions and deletions cannot be classified as SNPs, and the analysis of such variations is widely used for population genetics and evolutionary studies.

The main methods of this study include Principal Components Analysis, admixture analyses, F statistics, Runs of Homozygosity, inbreeding coefficient, estimation of effective population size, and statistical approaches to detect signals of ongoing recent selection. The application of all of these methods should give deeper insights into the population structure and admixture history of Kyrgyz people, and then expand our knowledge of this population's genome. A detailed description of the Kyrgyzstan population structure has consequences and implications for the design of future analyses in the field of molecular anthropology, including genotype-phenotype association studies, facilitating the identification of alleles associated with diseases and traits and providing insights into the effect of evolutionary forces shaping their variation.

3. MATERIALS AND METHODS

3.1. Data

The dataset used as reference was Allen Ancient DNA Resource (AADR) release version 54.1, which contains both ancient and present DNA (Mallick and Reich 2023; Mallick et al. 2023). The “1240K + HO” dataset was used, which comprehends a present-day and ancient individuals' dataset merged with present-day individuals from Human Origins. The former has 1,233,013 sites, and the latter has 597,573. These reference datasets were then merged with our samples, the Kyrgyzstan genome dataset previously collected during the Marco Polo Project (official website available at https://medialab.sissa.it/mp/il-progetto/marcopolo2010/english/). The Kyrgyz data set contains 184 individuals in total, with a total of 645,559 Single-Nucleotide Polymorphisms (SNPs). The Kyrgyz and the reference genotype data were merged using PLINK 1.9 (Chang et al. 2015).

3.2. Marco Polo Project

The Marco Polo Project started in 2010 as a scientific expedition along the Silk Road to assess the genetic diversity among different populations. It focused on three main topics: population genetics, genetics of taste, and genetics of food preference. Genetics of color perception, smell, and hearing also followed these main topics. The project lasted two years and had three expeditions (2010, 2011, and 2012). They collected samples from 8 countries (Crimea, Georgia, Armenia, Azerbaijan, Turkmenistan, Uzbekistan, Tajikistan, and Kazakhstan).

Kyrgyzstan was visited in 2012. The procedures were the same as the other expeditions, where the scientists led by Professor Gasparini met the corresponding representatives of each isolated community. They aimed to sample 50-100 individuals and the DNA was collected by either mouth rinse or saliva samples. The analysis of the samples collected allows having a better understanding of the relationship between genetics, food preferences, and eno-gastronomy traditions.

Furthermore, this project's relevant and positive feature is that a group of scientific communication experts accompanied the scientist during the expeditions. Therefore, the project, whenever possible, was displayed publicly.

3.3. Quality Control of the Kyrgyz Dataset

3.3.1. Linkage Disequilibrium Pruning

Pruning the data based on linkage disequilibrium is a useful step for several analyses. Linkage disequilibrium (LD) describes when variant alleles close to each other are associated in a non-random manner. Alleles at different loci are expected to be correlated randomly, therefore, unliked. However, if the association frequencies are higher than expected, they are in linkage disequilibrium. Therefore, when a dataset presents high LD, several alleles are linked, which is equivalent to redundant information. This redundancy can interfere with forward analyses, producing noise to the results. To avoid that redundancy, the data can be pruned by LD. This will result in keeping only the SNPs with approximate LD between them.

The merged file was pruned using PLINK 1.9. The parameter used was "--indep-pairwise", with a window size of 50 Kilobases (KB), step size of 10, and r^2 threshold of 0.2. This parameter in PLINK prunes the data based on a pairwise genotypic correlation, in other words, an SNP-SNP metric. Essentially, the command will calculate the LD between all possible pairs within the designated window size (in this case, 50 KB) and then remove one SNP of a pair based on the r^2 threshold input (0.2). The step size input will make the command shift the window on the number of SNPs input in order to proceed.

3.3.2. Runs of Homozygosity and Inbreeding

Runs of Homozygosity are long contiguous portions of the homozygous genotype due to the inheritance of identical alleles from each parent with a shared common ancestor, first discovered by Broman and Weber (1999). It serves as a powerful tool in population genetics and evolutionary biology, being an indicator of genomic autozygosity. This phenomenon can be caused by several events, such as population bottleneck, natural selection, inbreeding, genetic drift, and consanguineous mating, and it is seen in both inbred and non-inbred populations (Curik, Ferenčaković, and Sölkner 2014). Assessing ROH can be useful to give insights into a certain population's demographic history and structure.

In this study, the ROH was assessed using PLINK 1.9. The criteria used were: minimum density of 50 SNPs; gap length threshold between two SNPs as 1000 Kb; sliding window of 2000 Kb; minimum number of SNPs per window of 50 SNPs; and finally only one heterozygous SNPs per window was allowed. The results were plotted using matplotlib pyplot in Python 3.

Inbreeding, as the name suggests, is the mating within closely related individuals, and it has important implications in evolutionary genetic studies (Charlesworth and Charlesworth 1987). To assess this, PLINK 1.9 was used again. The command used was --ibc in the merged dataset. The algorithm gives three different inbreeding coefficients (Fhat). The one used in this study was the "Fhat2" which corresponds to the "Excess homozygosity-based inbreeding estimate". The results were plotted using matplotlib pyplot in Python 3.

3.3.3. Relatedness

Related individuals can sometimes bias some type of analysis. Therefore, removing one of a pair of related individuals is necessary. This step was performed on the merged dataset using KING (Manichaikul et al. 2010), using the specific command "unrelated". The program first calculates the relatedness of the data, and then extracts a list of unrelated individuals from it. If the option "degree 2" is specified, as in this case, then it will remove individuals with a second-degree relationship.

3.4. Population Structure

3.4.1. Principal Component Analysis

Principal Component Analysis (PCA) is a technique to assess large datasets with multiple dimensions. This approach decreases the dimensionality of the data, facilitating its interpretation and visualization. In this study, a PCA was performed to investigate better and visualize the population structure of Kyrgyzstan. Two PCAs analyses were done in this study: one with all modern samples, and one as projected PCA. The principal components (PCs) were calculated using PLINK 1.9 and then plotted with R. PLINK extracts the first 20 principal components (PCs). The first and second PC was plotted using R. Additionally, the data was also filtered by Minor Allele Frequency (MAF), with a threshold of 0.05.

The difference between the PCA and the projected PCA is that the former uses all the individuals of the dataset, while the latter uses only a subset of the data to calculate the principal components. Then, all the other populations are then projected into the already existing PCs.

In other words, only a few selected populations are used for the computation. This not only decreases the computational cost but also evades the artifacts that could potentially interfere with the analysis. Therefore, the unsupervised PCA is generally used as an initial visualization of how the data is distributed, and the projected PCA gives a more accurate representation allowing a better understanding of the data distribution.

3.4.2. Admixture

ADMIXTURE is a program capable of quickly estimating the ancestry of large datasets with multilocus SNPs genotypes (Alexander, Novembre, and Lange 2009). This program estimates individual ancestries from multilocus SNPs datasets by the maximum likelihood approach. However, ADMIXTURE is significantly faster than other global ancestry methods due to a numerical optimization algorithm. Moreover, it updates allele frequency and ancestry fraction parameters using a block relaxation approach and solves several optimization problems using a sequential quadratic programming algorithm. It also uses a method called quasi-Newton acceleration to speed up the algorithm. Overall, ADMIXTURE is faster and more efficient than other algorithms like EM and MCMC

The ADMIXTURE (Alexander, Novembre, and Lange 2009) software requires a dataset containing only unrelated individuals, so the related individuals were extracted from the merged dataset, besides being pruned. It was used in this study dataset to assess the possible ancestries of the Kyrgyzstan population. Furthermore, populations containing large ROHs were also excluded in order to perform this step to avoid biases.

Therefore, according to the ROH results already obtained from previous steps, populations containing ROH of more than 25.000 Kb on average were removed. The analysis and identification of the populations with great amounts of ROH were done through a Python script.

In order to run ADMIXTURE, a predetermined number of believed ancestral populations (K) is needed as input. The program was run with K values ranging from 2 to 8, and the one that best fits the data will have the lowest cross-validation error, which in this case is K=7. To visualize the ancestry of the dataset, the ancestral component file was plotted using R. The package "pophelper" was used in this process (Francis 2017). This R package was specifically developed to interpret the results of population structure analysis programs, such as ADMIXTURE.

To run projected ADMIXTURE, first, it is necessary to split the dataset into a reference and a study one. As the name suggests, the reference dataset will act as a reference for estimating the ancestral estimates, which then the study populations will be projected into it. The reference populations in this study were chosen geographically. The criteria were one representative population from each country present in the dataset. In case some of the ancestral components were left off after the election, a representative population containing most of it was chosen.

After choosing the reference dataset, ADMIXTURE should be run so the allele frequencies for the ancestral populations can be obtained (.P file). Finally, the study dataset can undergo an ADMIXTURE run so it is projected into the ancestral components of the reference dataset. The .Q files obtained were later plotted using the same R package mentioned earlier (pophelper).

Moreover, it was performed a quick cluster analysis to identify whether Kyrgyzstan really could be treated as three different groups. To achieve that, the R package MClust (Scrucca et al. 2016) was used on the Kyrgyzstan .Q files obtained from ADMIXTURE. Then, boxplots were plotted to better understand and visualize the distribution of the ancestral components in Kyrgyz people.

3.4.3. Clustering Kyrgyzstan

As confirmed by the PCA and ADMIXTURE analysis, Kyrgyzstan can be divided into four different clusters. To further understand and confirm the clusters found in the Kyrgyzstan samples, it was used the R package MClust (Scrucca et al. 2016). This is a contributed R package that uses finite normal mixture modeling for model-based clustering, among other functionalities. In order to cluster the dataset, it uses the method of hierarchical clustering. MClust was performed on the admixture .Q files, which gave us three main clusters. Next, the MClust clusters and the clusters seen in the PCA were compared using Python to confirm with the individuals from each cluster match in both analyses. Once the IDs were properly divided, PLINK was used to update them with the command "--update-ids".

3.4.5. Effective Population Size

Effective population (Ne) size is essentially the number of individuals in an idealized population that can represent the whole real population's genetic diversity. This idealized population is assumed to have random mating and random allele frequency variation that is similar to the true population. Commonly, the census population size is around three times greater than the Ne (Frankham 1995), and it can variate throughout time and loci. Furthermore, the Ne also affects the intensity and pace of natural selection. Therefore, studying the history of a population's Ne can reveal bottlenecks and population growth, essential to describing a population demographic landscape.

IBDNe was the software used to obtain the Kyrgyzstan Ne (S. R. Browning and Browning 2015). This program can estimate the Ne through long Identity by Descent (IBD) segments using a non-parametric approach (does not require assumptions about the data distribution). IBD is a term used to describe when a segment of the DNA is identical in two or more individuals due to ancestral inheritance. The program uses longer IBD segments than a threshold because it gives higher power and a smaller probability of false positives. However, these longer IBD segments reflect only the recent Ne and cannot be used to estimate ancient demographic history. It estimates the Ne for each time to the most recent common ancestor (TMRCA) using the IBD segments' length and quantity assigned to them. Then, the expected distribution of TMRCA is calculated in units of generations.

The IBD segments were first obtained using IBDseq (B. L. Browning and Browning 2013b) for each chromosome of each Kyrgyzstan cluster, and then IBDne was run, also individually for each cluster. The PLINK GRCh37 genetic map was used for both programs (Frazer et al. 2007).

3.4.6. Ancestry-Specific Recent Effective Population Size

In order to estimate the ancestry-specific recent effective population size of the main Kyrgyz group, Kyrgyz A, it was used a combination of the Snakemake (Mölder et al. 2021) pipeline for AS-IBDNe by Henn Lab (available at https://github.com/hennlab/AS-IBDNe) and the original AS-IBDNe pipeline described as in Browning et al. 2018. This pipeline uses the programs BEAGLE and BEAGLE Utilities (B. L. Browning et al. 2021), Refined IBDNe (B. L. Browning and Browning 2013a), and RFMix v.2 (Maples et al. 2013).

To summarize, the AS-IBDNe pipeline consists in first phasing the dataset containing both your study population and the reference populations. In the original pipeline, it was required to identify the IBD segments in order to run RFMix, however, the new version used in this study requires only a .vcf file for both the reference and study populations' phased data. Since RFMix requires at least two reference populations, the ones used for this study were the Finnish population (FIN) and the Japanese population (JPT) from the 1000 Genome Project reference panel. The 1000 Genome Project was chosen as the reference panel in order to increase the number of SNPs. The FIN population represents North European ancestry and the JPT represents the East Asian ancestry of the Kyrgyz A group.

The program BEAGLE was used to phase the data, which can be defined as inferring which parent the haplotypes were inherited from, information that is not usually provided in the SNPs array. Therefore, a statistical analysis must be conducted in order to estimate it, and this information is required for several other analyses, such as RFmix and some IBD identification. BEAGLE is a progressive two-stage phasing method using a sliding marker window, and it is a fast, accurate, and memory-efficient program. The program was run for each chromosome individually, with the use of a HapMap genetic map in PLINK format for GRCh37, and a reference .vcf file from the 1000 Genome Phase 3 reference panel (Auton et al. 2015).

RFMix uses a machine learning (ML) approach, specifically a random forest discriminative ML algorithm combined with a conditional random field model of the linear chromosome. In this way, the program is able to identify segments within an admixture population that have different ancestries.

Then, Refined IBD is used to detect the IBD segments, which applies an algorithm based on a dictionary of haplotypes called GERMLINE (Gusev et al. 2009). In this way, shared haplotypes longer than a determined threshold are identified, and next Refined IBD performs a refinement step to estimate the evidence for IBD. Following this step, it is needed to remove short gaps and breaks in the IBD segments caused by genotype or haplotype phase errors, which is achieved by running the script “Merge IBD segments” provided on the Refined IBD website.

In the Snakemake pipeline, the next step is a preprocessing step in which the files obtained from both RFMix and the Refined IBD with the gaps filled are combined and reformatted in order to run the AS-IBDNe. Finally, the program IBDNe was run with the formatted and pre-processed files obtained in the previous step, obtaining the Ne specifically for each ancestral population. The results were then plotted in R to provide better visualization of the data.

3.5. F Statistics

F-statistics are tests that use allele frequency correlations to study the relationship between populations, quantifying their genetic similarity level. The F3 can be used to either test a two-way admixture or to assess the shared drift between two populations compared with an outgroup population.

The formula for the F3 is defined as the product of the differences in allele frequencies between pairs of populations. It can be described as F3(A, B; C)=(c-a)(c-b), where A, B, and C are three populations; a, b, and c are allele frequencies of their corresponding populations. If the result is negative, it means that the allele frequencies of population C are intermediate to both A and B, which indicates that it is an admixture of A and B. However, it is not possible to firmly reject the possibility of C being admixed if the F3 score is not negative.

F4 statistics is similar to F3 populations, however, it measures the gene flow between four populations. It is defined as F4(A,B;C,D)=(a-b)(c-d). If there is any gene flow between these four populations, it is expected that the result would be zero, since the allele frequencies between them should be completely independent from each other. If the F4 score is significantly positive, it indicates a gene flow between either A and C, or B and D. On the other hand, if it is significantly negative, the gene flow probably occurred between A and D, or B and C. Usually, the population A is an outgroup, some population that is strongly divergent from the others, in order to guarantee that there is no gene flow between A and C or A and D. Thus, it is possible to focus and study only the gene flow between B and D or B and C. Both F3 and F4 statistics were performed using Popstats python script (Skoglund et al. 2015) with the transposed files of the Allen Ancient dataset merged with the Kyrgyzstan dataset. Each Kyrgyzstan cluster (A, B1, B2, and C) was assessed individually. The Popstats also provide the standard error and the Z score. The Z score is a key measure that indicates the significance of the F statistics. It is calculated as the F3/F4 score divided by the standard errors, and it basically measures how much the F score deviates from zero in units of standard error. Generally, a Z score smaller or equal to -3 (Zscore ≤ -3) rejects the null hypothesis, which then assures that the F3 score is actually significant. Therefore, a -3 Z score would indicate a significant gene flow between the populations.

3.6. Dating Admixture

Alder is a software used to assess population admixture which can also infer admixture proportions and dates (Loh et al. 2013). It is a weighted LD-based statistic that can detect signals of admixture through the decay of LD in SNP data of three populations (a test population and two reference populations).

Some events that can cause LD are genetic drift, selection, and population structure. Furthermore, nearby loci that were co-inherited from an ancestral population can interact between them, causing longer-range admixture LD (ALD). Usually in human populations, assuming they are a well-balanced and homogeneous population, the LD is close to imperceptible in distances greater than a hundred kilobases. On the other hand, ALD from past admixture events is still detectable. However, recombination degrades both LD and ALD, which generates a signal that can be measured in order to obtain a precise estimate of the admixture time. Alder was used on the merged dataset (reference populations + Kyrgyzstan) in Eigenstrat file format. The software was run individually for each cluster of Kyrgyzstan (A, B1, B2, and C), and the source populations used were Russian, Han, and Iranian.

3.7. Selection

To assess signals of ongoing selection within the different clusters of Kyrgyzstan, the program Selscan was used (Szpiech and Hernandez 2014; Szpiech 2022). This program implements several haplotype-based statistics, including Integrated Haplotype Score (iHS) and number of segregating sites by length (nSL) which were the ones used in this study.

The iHS statistics were first proposed by Voight et al. (2006) and were applied to perform a genome-wide scan to identify signals of recent positive selection of variants not yet fixated in a single population. In the end, the authors were able to create a genome-wide map of human selective sweeps. The equation that Selscan uses is an adaptation of the original iHS formula, with the same purpose of tracking the decay of haplotype homozygosity in the derived and ancestral haplotypes. The nSL statistics can be used to detect signals of positive selection in a single population based on the increase of haplotype homozygosity (Ferrer-Admetlla et al. 2014). It is a method similar to the iHS statistics, but it does not use a genetic map to be performed, since nSL essentially calculates the length of a homozygous haplotype segment between a pair of haplotypes in a number of mutations. It was proven by the authors that most of the time, nSL is even more robust than other methods.

The program requires separate runs for each population and chromosome. It can be performed on either phased or unphased data, however, it is recommended to phase the data before to achieve more accurate results.

In this study, both iHS and nSL tests were performed in phased and unphased data. The dataset was phased using the program Beagle. Once the data was properly phased, both iHS and nSL were calculated using Selscan, which gives the unstandardized scores. The iHS statistics require a genetic map, specifically, the HapMap in PLINK format was used. Then, all the scores were normalized using the tool “norm” provided by Selscan in order to obtain the standardized scores.

Finally, the normalized absolute values were plotted using the R package “qqman” (Turner 2018) to produce Manhattan plots. The top 100 hits of all statistics performed in all populations were annotated using Ensembl (Cunningham et al. 2022).

4. RESULTS

4.1. Principal Components Analysis

The principal components were extracted from the modern samples contained in the merged dataset using PLINK 1.9, and then plotted using R (Fig. 2). It was performed both unsupervised PCA (Fig. 2A) and projected PCA (Fig. 2B). Key populations were highlighted using different colors and symbols, whereas all the other populations were left as empty light grey circles. The populations highlighted were populations that potentially could influence Kyrgyzstan or populations historically, geographically, and genetically similar to Kyrgyz people. The axes were rotated in order to better fit the actual geographic coordinates. The PCAs have no great differences between them besides the presence of less noise in the projected PCA.

Kyrgyzstan individuals (purple-filled squares) in both PCAs are spread horizontally, almost crossing the whole Asian continent, from Russia to Central Asia to China. The majority of Kyrgyz individuals are co-located with other Central Asia populations, such as Uzbek and Hazara, which are also admixed populations. Some Kyrgyz people are also closer to Tibetan and Chinese people, and on the other side, some are clustered more closely to Russians. Therefore, it is already possible to see that the Kyrgyzstan population is spread throughout the PCA in different groups.

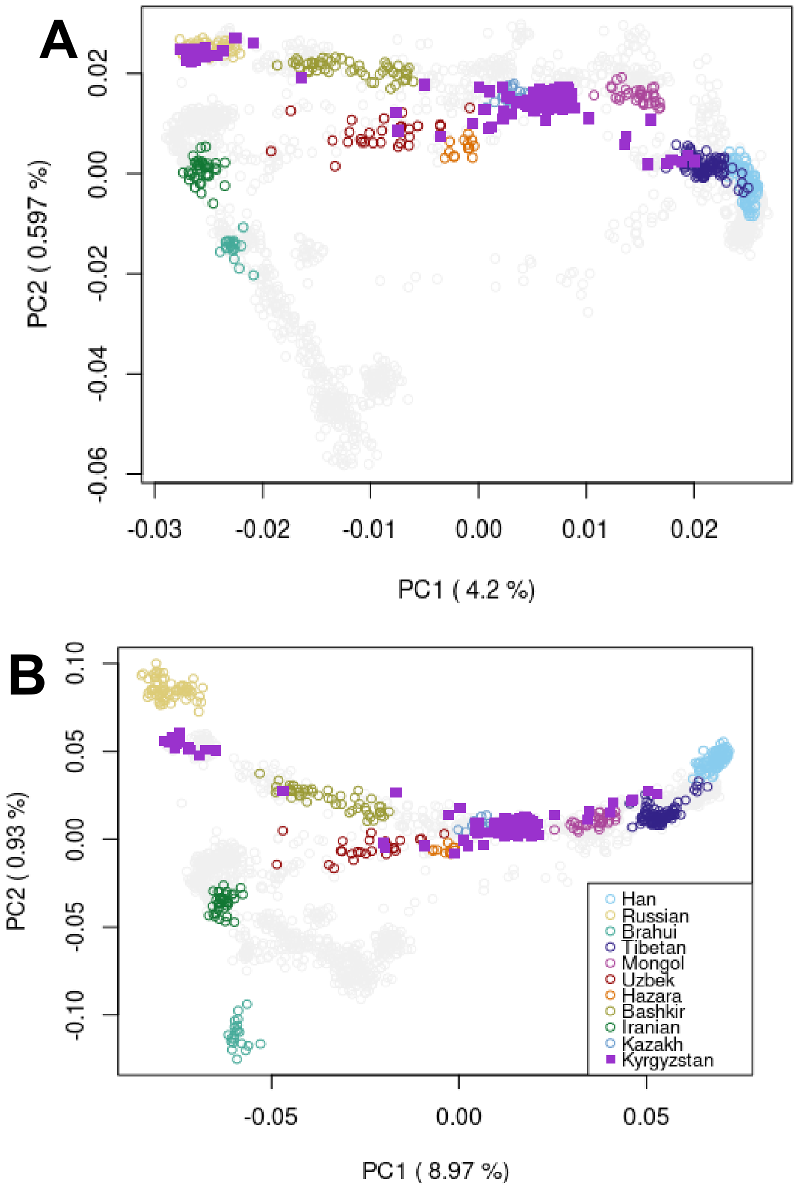


Figure 2. Principal Component Analysis (PCA). PCA of populations around Central Asia filtered for MAF < 0.05, and pruned. **A.** Unsupervised PCA of all present-day individuals from populations around Central Asia. **B.** Projected PCA of unrelated present-day individuals from populations around Central Asia.

4.2. ADMIXTURE

The ADMIXTURE software estimates the ancestry components of populations. It was run with K values from 2 to 8, K=6 for the unsupervised ADMIXTURE was considered the best solution, according to the cross-validation procedure. The results (Fig. 3) show that Kyrgyzstan has all six ancestry components in significant amounts, with each component represented as a different color. This indicates that Kyrgyzstan probably has an admixed population, with European (Russian and Iranian), North and East Asia (Buryat, Tibetan, and Han), and also Pakistan components (Balochi). The pattern of Kyrgyzstan ancestry components is similar to other Central Asian populations, however, it is still unique. Furthermore, it is visible that Kyrgyzstan presents three different patterns of ancestry components within its population. One of the patterns has a predominantly Russian component; another one has more Han and Tibetan components than the average; and the majority presents well-balanced proportions of the six ancestry components, but more predominance of a Buryat/Northeast Asian component.

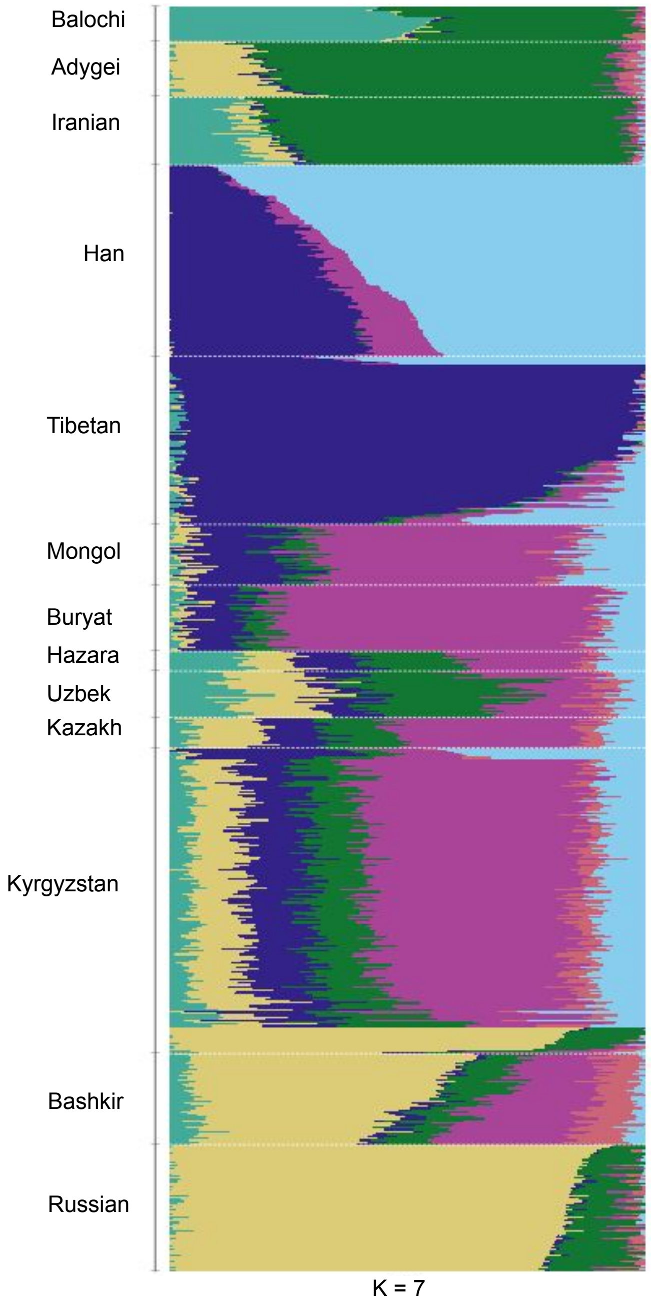


Figure 3. Unsupervised ADMIXTURE results. Plotted using R package “pophelper”. The ADMIXTURE software requires a predetermined number of believed ancestry components (K). The K=7 was the number of ancestral components that had the smaller cross-validation error. Kyrgyzstan contains all seven ancestral components in different proportions, with a greater predominance of the magenta component, which corresponds to a East Asian specific component.

The projected ADMIXTURE requires reference populations to first estimate the ancestry components. One non-admixed population from each country around Central Asia was picked as the reference population. The projected ADMIXTURE (Fig. 4) is similar to the unsupervised one, however, it can more accurately describe the proportion of the ancestry components in each population. The projected ADMIXTURE shows better the proportion of North and East Asian components. Specifically, the significant amount of Han Chinese component in light and dark blue, and the East Asia component in magenta. The same three different patterns within Kyrgyz people described before are maintained also in the projected ADMIXTURE.

4.3. Clustering Kyrgyzstan

After the results of both PCA and ADMIXTURE analysis, it was clear that the Kyrgyz people are divided into three different groups that probably have different demographic histories. Clustering analysis using the MClust R package was performed in order to confirm it and later properly divide the population. The MClust was applied to the ADMIXTURE results to find the three clusters, later named Kyrgyzstan A, B, and C. Additionally, a boxplot was plotted to better visualize the proportion of the ancestry components (AC) in each cluster (Fig. 5).

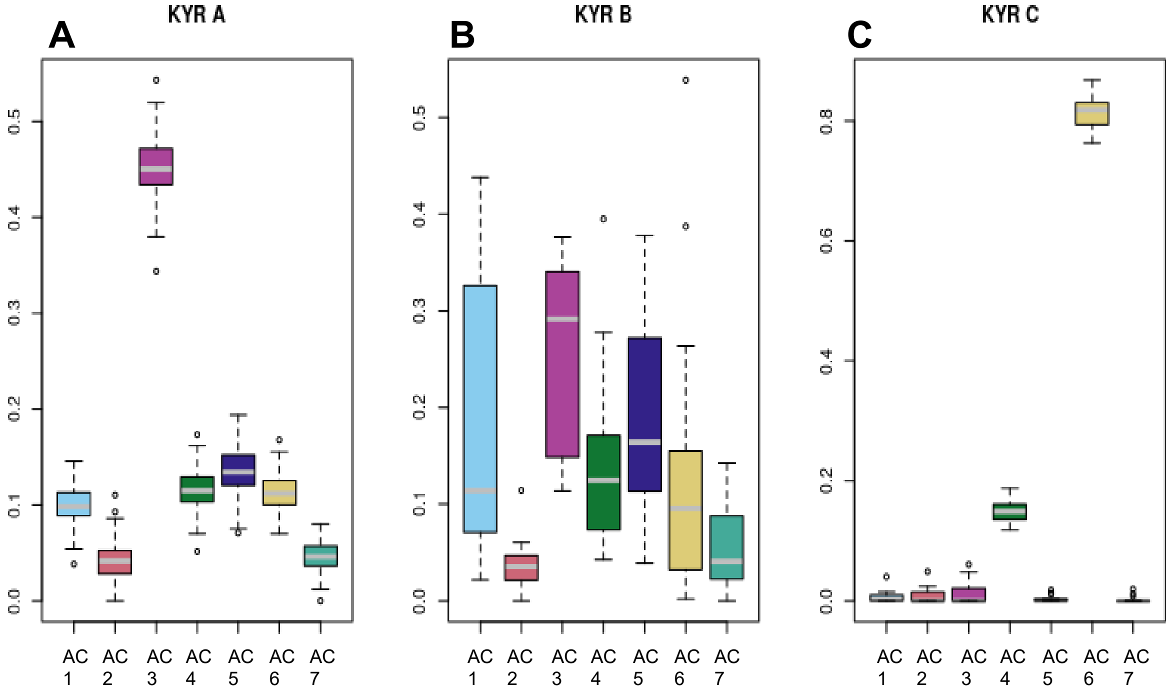


Figure 5. Boxplots of the three Kyrgyz clusters. The boxplots were obtained using MClust. The boxplots shows the proportion of each ancestral component (AC) of the ADMIXTURE results. **A.** Boxplot of cluster "Kyrgyz A", composed by 143 individuals, with a great predominance of AC3 which corresponds to a East Asian component. **B.** Boxplot of cluster "Kyrgyz B", composed by 28 individuals, with no great predominance of a specific AC. **C.** Boxplot of cluster "Kyrgyz C", composed by 13 individuals, with an evident predominance of AC6, which corresponds to a Russian component.

The majority of the Kyrgyz population (143 individuals), named "Kyrgyz A", is clustered as in Fig. 5A. It has a predominance of East Asian (in magenta, AC3, 40%) and Han Chinese-specific components (in light and dark blue, AC1 and AC5 respectively), with no significant differences between the proportions of the other ancestry components. Then, 28 individuals (Kyrgyz B) cluster as in Fig. 5B with well-balanced proportions of all the ancestry components, with a slightly greater amount of all East Asian components. Lastly, Kyrgyz C (Fig. 5C), composed of 13 individuals has indisputable amounts of Russian components (in beige, AC6), with some significant amounts of Iranian components (in green, AC4).

Considering the equitable distribution of the different components in Kyrgyzstan B and analysing the PCA results, the cluster was then divided once again into two groups: Kyrgyzstan B1 and Kyrgyzstan B2, with 18 and 10 individuals, respectively. These two groups were then split according to their distribution in the PCA, whether they were positioned more towards Europe (in the case of Kyrgyzstan B1) or East Asia (Kyrgyzstan B2).

To better visualize the groups and if they are distributed accordingly, a new PCA was performed once the individuals were relabelled (Fig. 6).

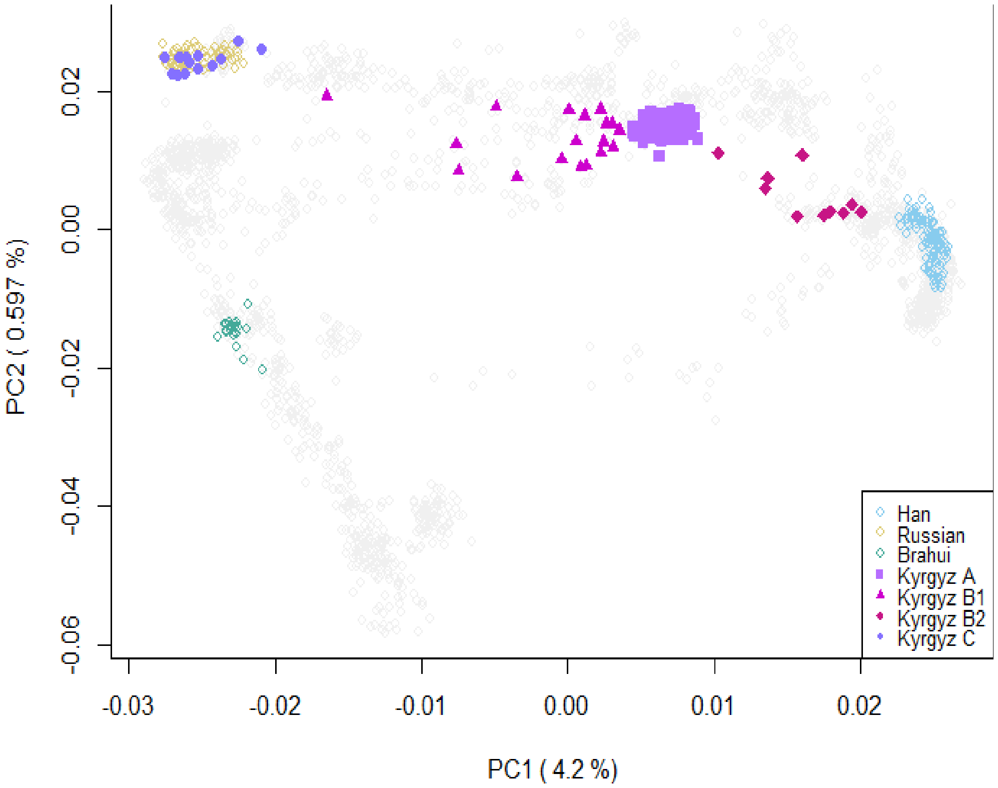


Figure 6. Principal Components Analysis (PCA) with Kyrgyz clusters. Unsupervised Principal Components Analysis (PCA) previously performed, but Kyrgyz individuals were relabeled according to their cluster. Russian, Han Chinese, and Brahui populations are represented in the figure for reference. Kyrgyz A is represented as purple-filled square;, Kyrgyz B1 as magenta triangles; Kyrgyz B2 as dark pink-filled diamonds; and Kyrgyz C as blue violet-filled circles.

4.4. Runs of Homozygosity and Inbreeding

The analysis of runs of Homozygosity was performed using PLINK using the parameters mentioned in the Methods section. The Kyrgyzstan groups present similar values between them, with the exception of Kyrgyz B2, which have a distribution of ROH distinctly smaller (Fig. 7). The groups Kyrgyz A and Kyrgyz C seem to have a distribution more similar to the distribution of European populations such as Russian and Iranian. On the other hand Kyrgyz B1-2, due to their higher level of admixture have lower homozygosity due to ROH. We should note that ROH patterns are linked to recent events such as consanguineous marriages.

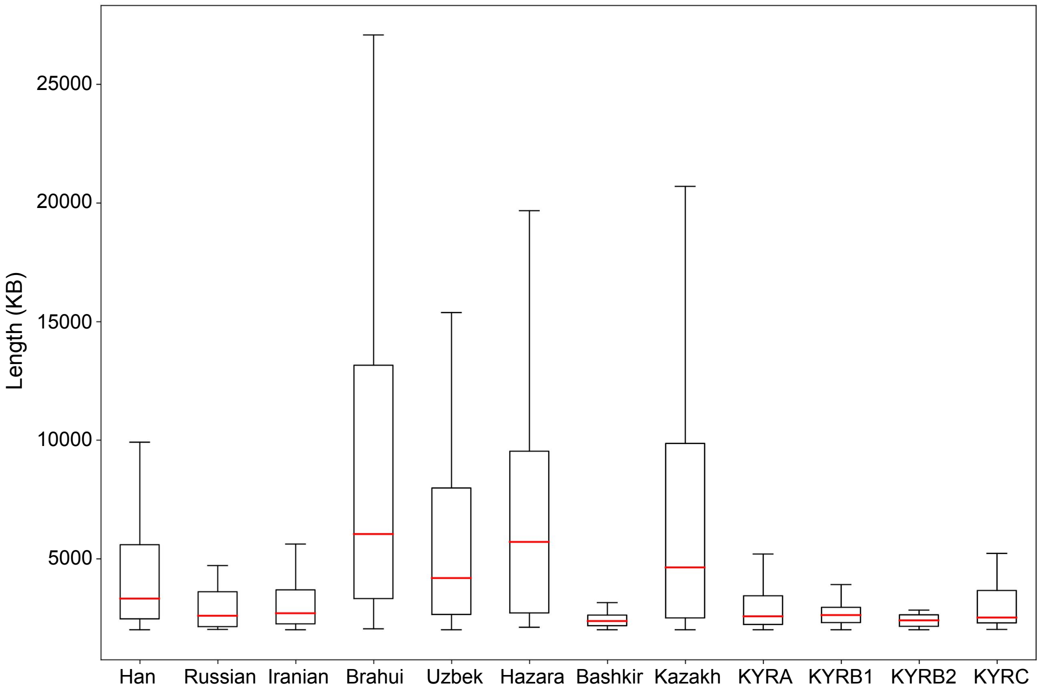


Figure 7. Runs of Homozygosity (ROH) across populations around Central Asia. ROH obtained with PLINK 1.9 for different populations around Central Asia, by length in KB. Outliers are hidden.

F inbreeding or Fhat2 represents the excess of homozygosity in the different populations. In Figure 8, it is visible that the more homogeneous (according to the PCA) clusters, such as Kyrgyz A and Kyrgyz C are the ones with the highest Fhat2. The high level of Fhat2 in Kyrgyz populations could be linked to the small population size of the founders.

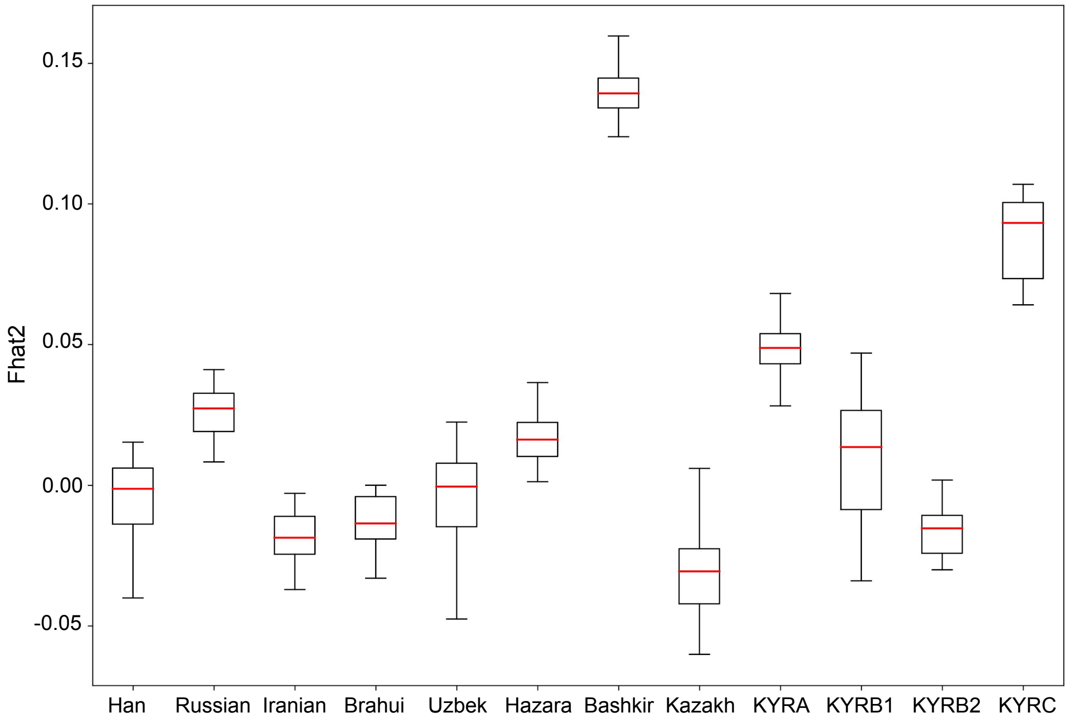


Figure 8. F inbreeding for populations around Central Asia. F inbreeding obtained with PLINK 1.9. The score used to plot was Fhat2, which is an estimate related to the excess of homozygosity.

4.5. F Statistics

The F statistics were performed using the PopStats Python script. First, F3 statistics were done to further confirm whether the Kyrgyz population is indeed a three-way admixture. The test was run with each Kyrgyzstan group (A, B1, B2, and C), which were always put in the third position, changing only the possible source populations. In order for the results to be considered significant, the Z score must be either greater than 3 or smaller than -3. Furthermore, it is considered an admixture when the Z score is smaller than -3 and the F3 score is negative. In the plots (Fig. 9-12), when the Z score obtained was significant, the plotted circles for the F3 score are filled with the color red, when not significant, the F3 score is represented as empty circles.

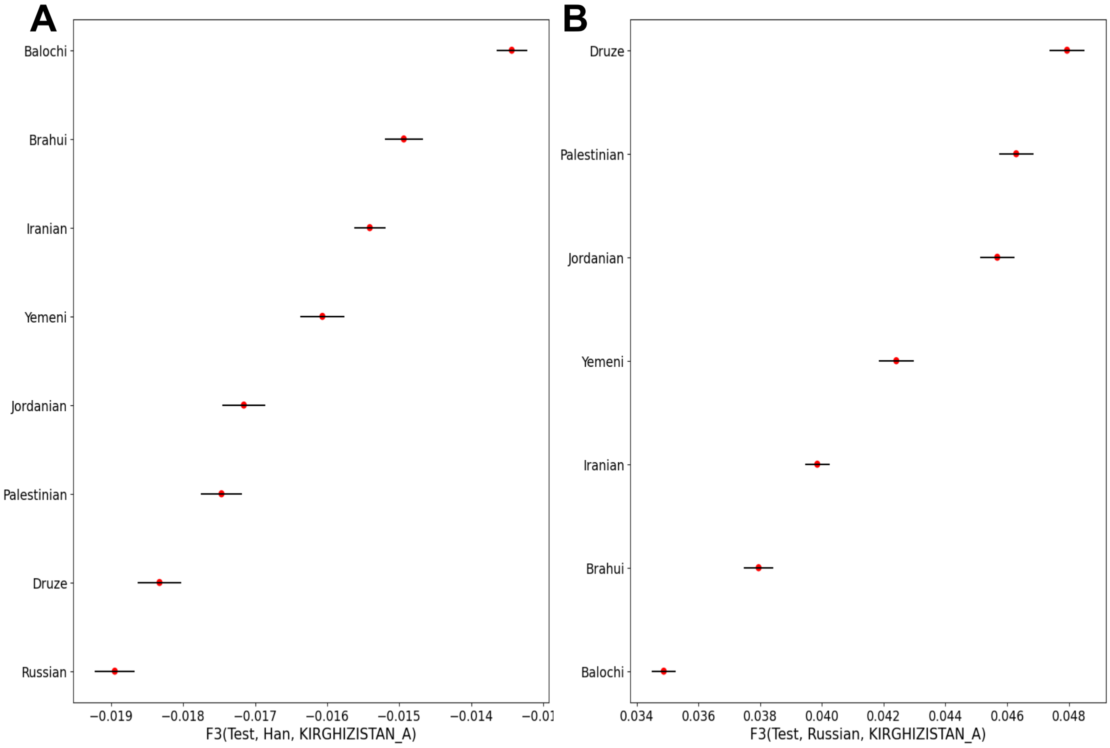


Figure 9. F3 results of Kyrgyz A. A. F3 results obtained with PopStats to test whether Kyrgyz A is an admixture between Han and test populations. Red-filled circles represent significant Z score, horizontal lines represent standard errors. **B.** F3 results obtained with PopStats to test whether Kyrgyz A is an admixture between Russian and test populations .Red-filled circles represent significant Z score, horizontal lines represent standard errors

Kyrgyzstan A had significant F3 scores for admixture between Han Chinese and Russians (Fig. 9). It was also significant for F3(Han, Iranian; Kyr A), F3(Han, Balochi; Kyr A), and also F3(Han, Brahui; Kyr A). Additionally, F3 was also significant when Middle East populations and Han were used as test populations. When the F3 analysis was performed as F3(Test population, Russians, Kyr A), all Z and F3 scores obtained were significant, but positive, indicating that Kyrgyz A is not an admixture of Russians and any other population, except with Han Chinese.

Similar results were obtained with both Kyrgyzstan B1 and B2 (Fig.10 and Fig. 11), being all significant. The F3(Test population, Han, Kyrgyz B1/2) to all test populations had negative scores, indicating that Kyrgyz B1 and Kyrgyz B2 are an admixture between Han and all the test populations. On the contrary, all F3 scores for F3(Test population, Russian, Kyrgyz B1/2) were negative, indicating that they are not an admixture of any test population and Russian.

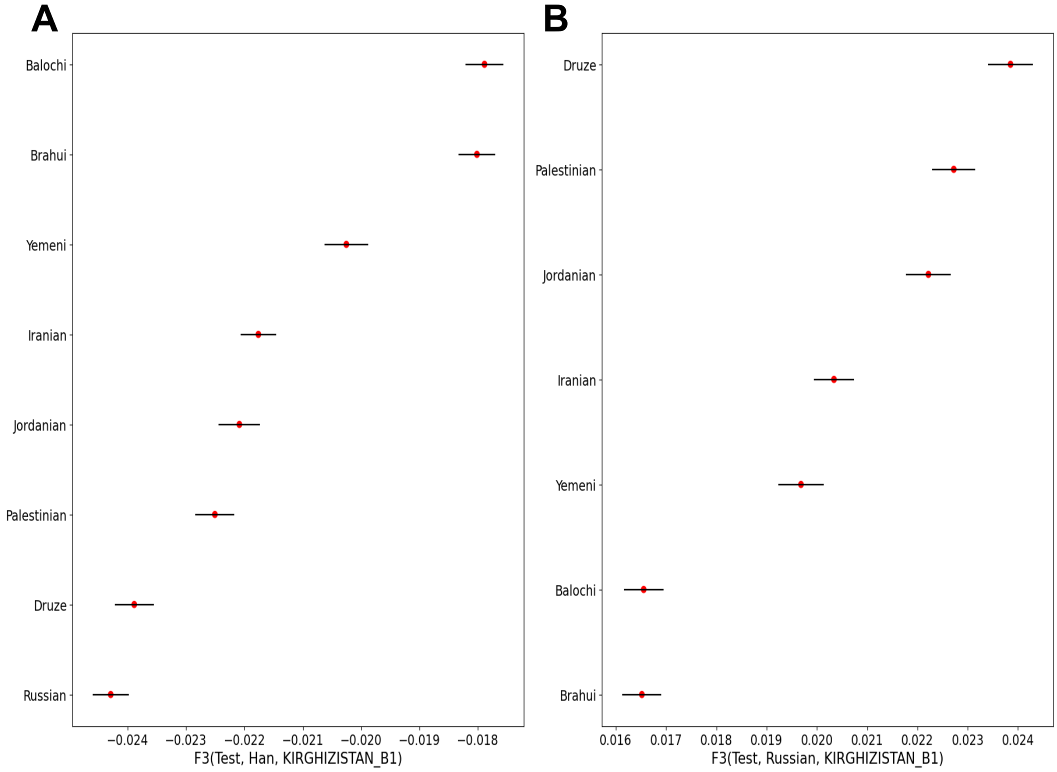


Figure 10. F3 results of Kyrgyz B1. A. F3 results obtained with PopStats to test whether Kyrgyz B1 is an admixture between Han and test populations. Red-filled circles represent a significant Z score, horizontal lines represent standard errors. **B.** F3 results obtained with PopStats to test whether Kyrgyz B1 is an admixture between Russian and test populations. Red-filled circles represent significant Z score, horizontal lines represent standard errors.

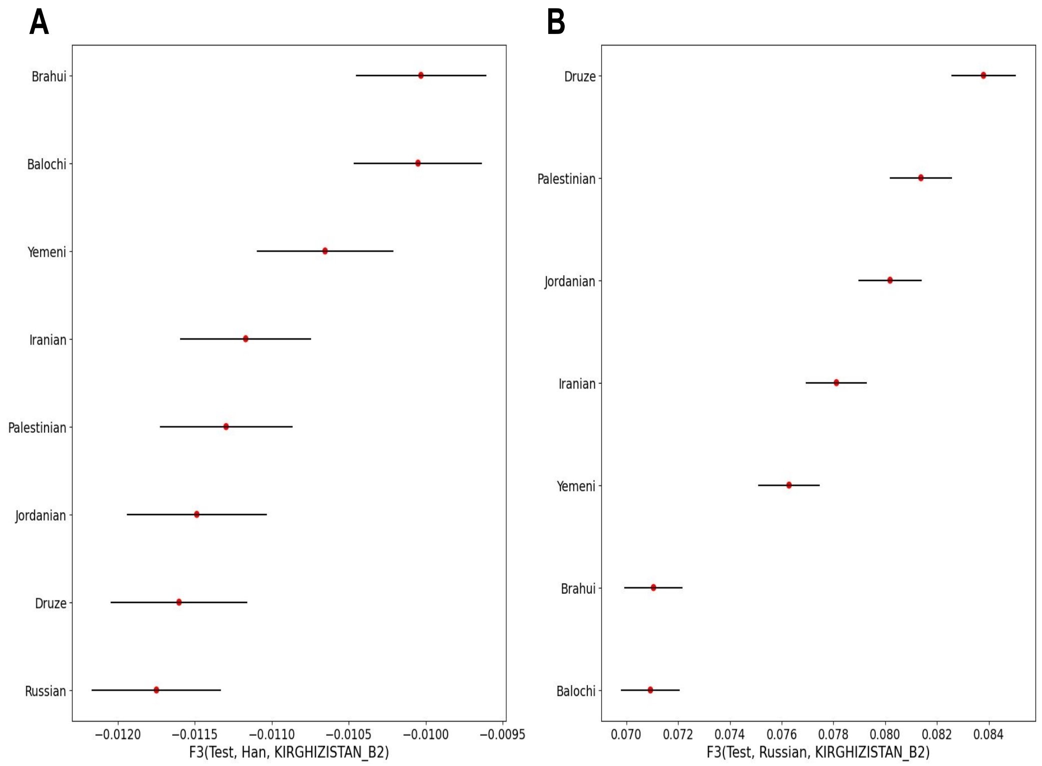


Figure 11. F3 results of Kyrgyz B2. A. F3 results obtained with PopStats to test whether Kyrgyz B2 is an admixture between Han and test populations. Red-filled circles represent a significant Z score, horizontal lines represent standard errors. **B.** F3 results obtained with PopStats to test whether Kyrgyz B2 is an admixture between Russian and test populations. Red-filled circles represent significant Z score, horizontal lines represent standard errors.

On the other hand, the F3(Test population, Han, Kyrgyz C) were significant (Fig. 12A), except when tested with Han, but with positive values. Therefore, Kyrgyz C is not an admixture of any test population and Han, except for the admixture of Russian and Han, in which the Z score was not significant, being not possible to affirm if it is or not an admixture. All the Z scores for F3(Test population, Russian, Kyrgyz C) obtained were not significant (Fig. 12B), therefore it is not possible to confirm whether Kyrgyz C is an admixed population of any of the test populations and Russian population.

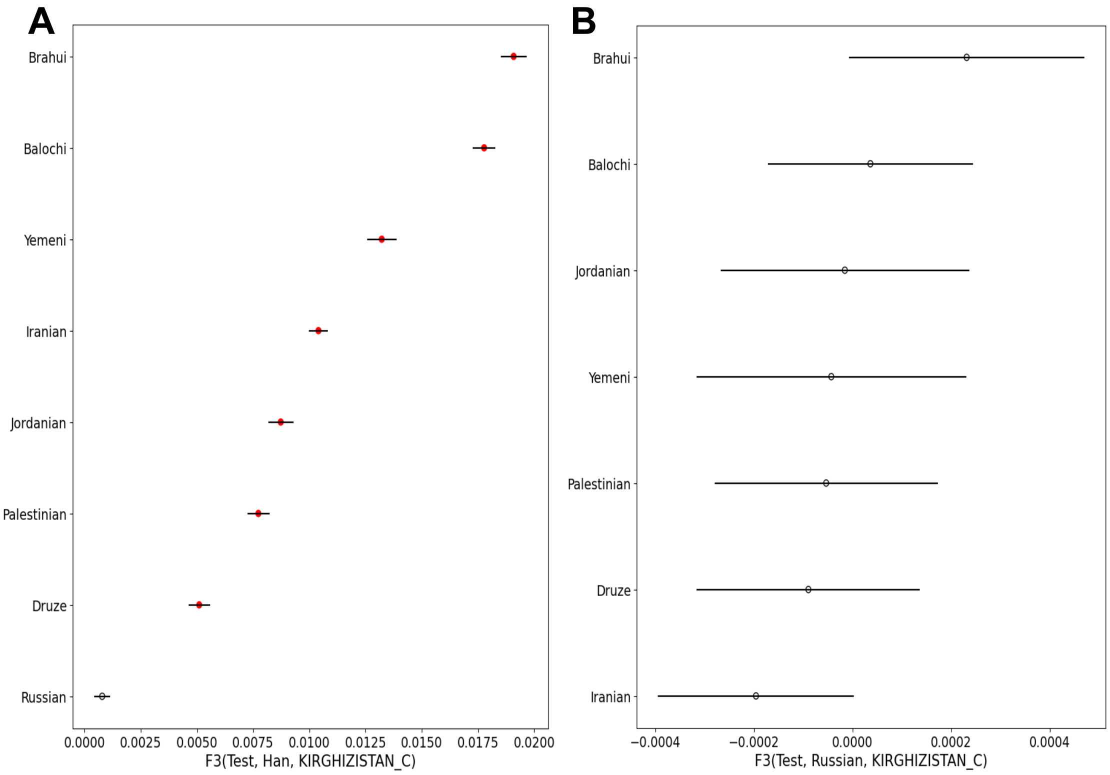


Figure 12. F3 results of Kyrgyz C. A. F3 results obtained with PopStats to test whether Kyrgyz C is an admixture between Han and test populations. Red-filled circles represent a significant Z score, horizontal lines represent standard errors, and empty circles mean not significant Z score. **B.** F3 results obtained with PopStats to test whether Kyrgyz C is an admixture between Russian and test populations. Empty circles represent not significant Z score, horizontal lines represent standard errors.

These results confirm that Kyrgyzstan A and B are an admixture of European, Pakistan, Middle East, and East Asian components. On the other hand, Kyrgyzstan C had no significant F3 or Z score for any of these tests, therefore it is not possible to conclude whether Kyr C is an admixed population or not.

The F4 statistics uses a fourth group as an outgroup, which in this case Mbuti population was used as an outgroup, and it gives information about which test population is closer to the target population. This test was used to better understand which population is most closely related to Kyrgyz ancestry. Regarding East Asian populations, both Japanese and Han

Chinese are equally adequate source populations for Kyrgyz A and B (Table 1). The same is valid for Brahui and Balochi, both having no significant differences between them, indicating that they all could be reliable sources of Pakistan's component into Kyrgyzstan. Russian is distinctly the source population of European gene flow into Kyrgyz.

Target	Population A	Population B	F4 score	Standard error	Z Score
Kyrgyz A	Han	Russian	0.01096	0.00030	36.76981
		Iranian	0.01363	0.00031	44.55388
		Brahui	0.01350	0.00030	45.20247
		Japanese	-0.00030	0.00009	-3.18974
	Dai	Japanese	-0.00171	0.00017	-10.35133
	Russian	Iranian	0.00267	0.00013	20.13018
		Brahui	0.00254	0.00015	17.04706
		French	0.00133	0.00008	16.51555
	French	Iranian	0.00134	0.00013	10.61571
	Brahui	Iranian	0.00013	0.00013	0.99440
	Balochi	Brahui	0.00032	0.00013	2.47021
Kyrgyz B1	Han	Russian	0.00472	0.00030	15.80949
		Iranian	0.00762	0.00030	25.50952
		Brahui	0.00796	0.00029	27.28557
		Japanese	-0.00014	0.00009	-1.47612
	Dai	Japanese	-0.00123	0.00016	-7.62738
	Russian	Iranian	0.00289	0.00013	22.09094
		Han	-0.00472	0.00030	-15.80949
		Brahui	0.00324	0.00015	21.77723
		French	0.00111	0.00008	13.85860
	French	Iranian	0.00179	0.00013	14.12409
	Brahui	Iranian	-0.00035	0.00013	-2.70911
	Balochi	Brahui	0.00024	0.00013	1.92199

Kyrgyz B2	Han	Russian	0.01836	0.00036	50.96498
		Iranian	0.02058	0.00037	55.88798
		Brahui	0.02001	0.00035	56.57956
		Japanese	0.00031	0.00010	2.95589
	Dai	Japanese	-0.00117	0.00019	-6.27920
	Russian	Iranian	0.00222	0.00014	15.32114
		Han	-0.01836	0.00036	-50.96498
		Brahui	0.00165	0.00017	9.91176
		French	0.00145	0.00009	16.07837
	French	Iranian	0.00077	0.00014	5.65103
	Brahui	Iranian	0.00056	0.00014	3.95633
	Balochi	Brahui	0.00029	0.00014	2.08947
Kyrgyz C	Han	Russian	-0.01315	0.00032	-41.52088
		Iranian	-0.00780	0.00031	-25.35078
		Brahui	-0.00605	0.00031	-19.54356
		Japanese	-0.00003	0.00010	-0.28990
	Dai	Japanese	-0.00040	0.00017	-2.30981
	Russian	Iranian	0.00535	0.00014	38.20628
		Han	0.01315	0.00032	41.52088
		Brahui	0.00711	0.00016	44.27724
		French	0.00083	0.00008	9.94605
	French	Iranian	0.00452	0.00013	34.45885
	Brahui	Iranian	-0.00176	0.00013	-13.59907
	Balochi	Brahui	0.00026	0.00013	1.93995

Table 1. F4 results of all Kyrgyz groups. F4 results were obtained with Popstats of all Kyrgyz groups, using Mbuti as Outgroup. Negative values represent gene flow between Target and population B, positive values represent gene flow between Target and population A.

4.6. Dating Admixture

Once confirmed that Kyrgyzstan is an admixed population, dating the time of this admixture is equivalently important. The time of admixture can be obtained using the program Alder as already described in the Materials and Methods section. The Alder test was successful in assessing the admixture of Kyrgyzstan A between Han and Iranian, with a p-value of 1.8e-118. The estimated admixture time for Kyrgyzstan A was around 22.2 human generations ago, which is around the year 1328 CE. Considering a human generation is around 28 years, then the admixture most likely took place around 621 years ago, being a recent admixture. The program discarded the Russian population due to long-range LD correlation, probably due to genetic similarity between source and target population.

The results for Kyrgyzstan B1 and B2 were both successful for Han and Russian, and then for Han and Iranian (Table 2). The p-value for Han and Russian admixture is 2.1e-06 in the Kyrgyzstan B1 population and 1.7e-11

for Kyrgyzstan B2. The admixture time between these source populations for Kyrgyzstan B1 and B2, respectively, is 15.12 (~423 years), and 18.88 generations (~528 years) ago. The p-value for Kyrgyz B1 admixture between Han and Iranian was 6.9e-13, with an admixture time dated 17.5 generations ago (~492 years). Kyrgyz B2 had a p-value of 2.3e-10 and a dating of 18.93 generations (~560 years) ago for the same source populations.

Finally, Kyrgyzstan C had no successful tests when estimating the admixture time, which is expected since no significant evidence of admixture was found when performing F statistics.

	Kyrgyz A	Kyrgyz B1			Kyrgyz B2		Kyrgyz C
Test Status	Success	Success	Success	Success	Success	Success	Failure
p-value	2.3e-111	9.50E-08	5.20E-20	0.038	1.70E-11	2.30E-10	0.2
Test population	Kyrgyz A	Kyrgyz B1	Kyrgyz B1	Kyrgyz B1	Kyrgyz B2	Kyrgyz B2	Kyrgyz B2
ref A	Han	Han	Han	Russian	Han	Han	Russian
ref B	Iranian	Russian	Iranian	Iranian	Russian	Iranian	Iranian
Z score	22.47	5.53	9.28	2.49	6.88	6.51	-
Admixture time	~1307-1349 CE	~1450-1603 CE	~1407-1512 CE	~929-1513 CE	~1344-1498 CE	~1338-1501CE	-

Table 2. Alder results of Kyrgyz A, B1, B2, and C. The results obtained with Alder indicate whether the test population is an admixture between the two reference populations and the estimated time period of when then the given admixture probably occurred.

4.7. Effective Population Size

Effective population size (Ne) is an important aspect to assess since it can give relevant information about a population's demographic history, such as bottlenecks and a population’s boom growth. It was used in the IBDne program to estimate the Ne of Kyrgyzstan throughout time using IBD segments. This analysis was not performed with the other Kyrgyz groups due to the small number of individuals.

Kyrgyz A showed a high and steady number of individuals (around 60 thousand) 100 generations ago (Fig. 13). After 100 generations, Kyrgyz A shows an accentuated decline in the Ne. It starts to stabilize around 20 generations ago, which is also approximately the admixture time given by Alder, indicated by the red line in Figure 11. After this decrease, the Ne stabilizes around 10 thousand individuals.

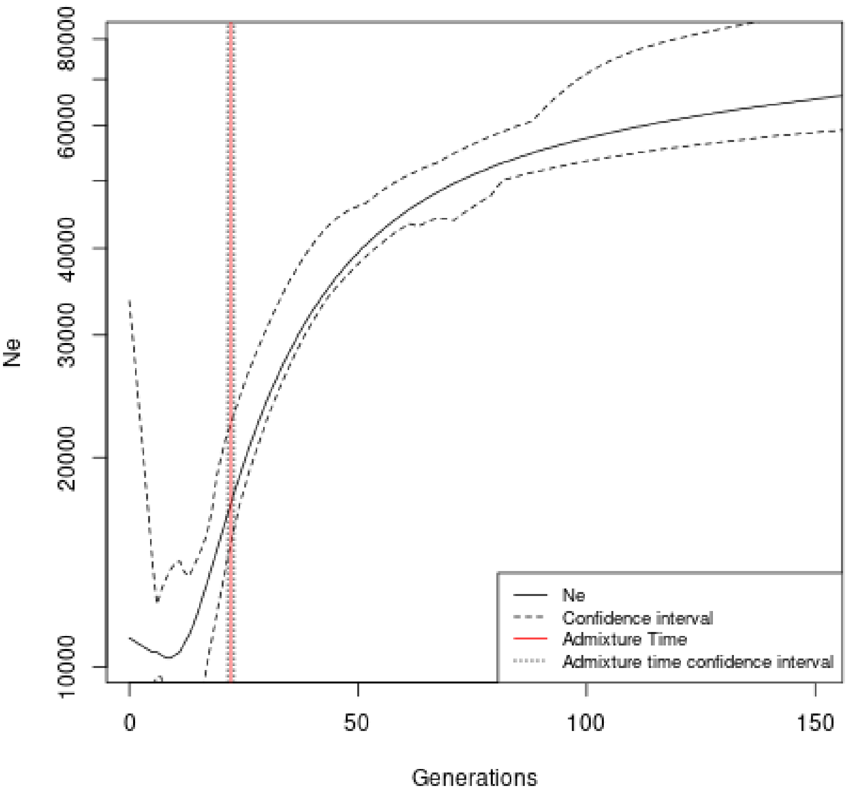


Figure 13. IBDNe result for Kyrgyz A. The IBDNe results show the variation of the effective population size (Ne) over time (in human generations). The dashed line represents the confidence interval for the Ne. The red line corresponds to the admixture time obtained with Alder and the dotted line is its respective confidence interval.

4.8. Ancestry-Specific Recent Effective Population Size

Ancestry-specific recent effective population size (AS-IBDNe) was performed using a Snakemake pipeline created by Henn Lab, which is a Snakemake adaptation of the original AS-IBDNe pipeline. It was then obtained the Ne for both European and East Asian ancestry, which were plotted using R (Fig. 14B and 14C). The admixture time obtained using Alder is represented in Figure 13 as a red vertical line, with its upper and lower bound represented as dotted red lines. The upper and lower bound of the Ne is represented as the filled polygon around the Ne main line.

It is visible that both the European-specific and East Asian-specific Ne are more constant over time compared to the Kyrgyz A obtained from the classic IBDNe, with a Ne of approximately 100,000 in 80 generations ago (~290 BCE). This is higher than what was seen for Kyrgyz A, which its Ne obtained was around 50,000 at the same number of generations. No steep decline is observed in both ancestry-specific graphs, with only a smoother decline co-localized with the line of admixture time in the plot, reaching its lowest Ne of around 50,000 after 20 generations. On the other hand, Kyrgyz A reaches its lowest Ne also after 20 generations (close to the admixture time estimated), but with a Ne of 10,000 individuals, which is considerably lower compared to the other ancestry-specific Nes.

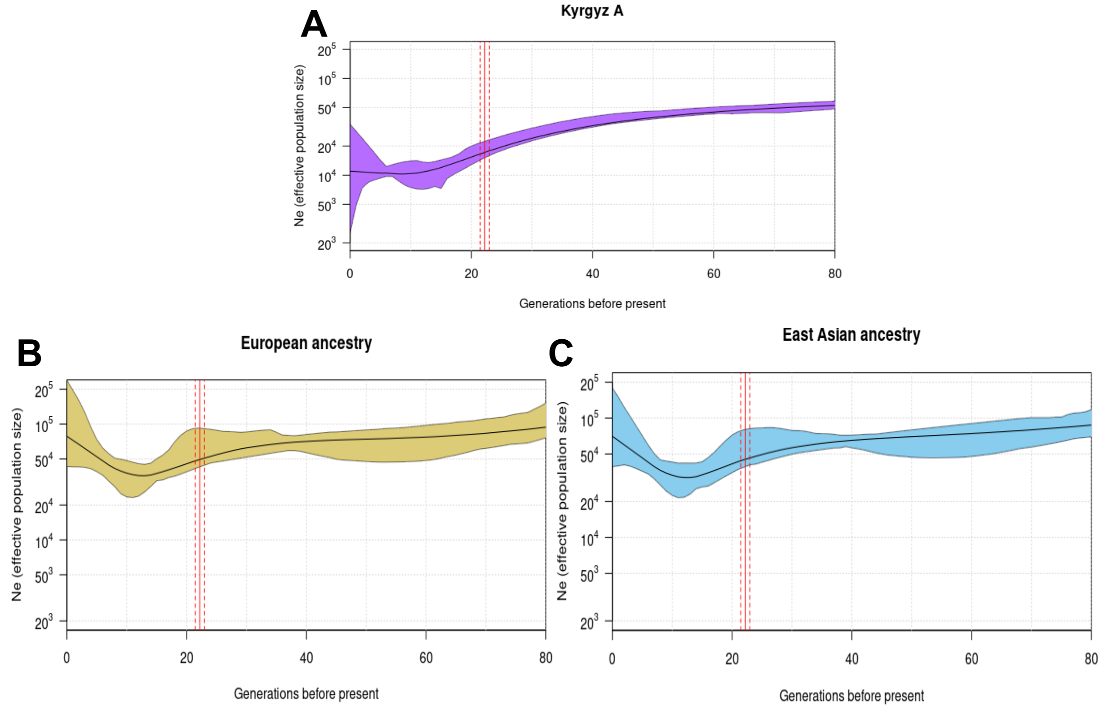


Figure 14. Effective population size (Ne) over time in generations before present. Admixture time is indicated as red line, with the dotted red lines representing the upper and lower bound of the confidence interval. The upper and lower bound of the Ne are represented as the colored-filled area. **A.** AS-IBDNe result of Kyrgyz A indicating Ne over time. **B.** AS-IBDNe result for the European ancestry Ne over time. **C.** AS-IBDNe result for the Ne of East Asian ancestry over time.

4.9. Selection

Two different analyses were performed using Selscant to detect signals of ongoing selection in all clusters of Kyrgyzstan, iHS, and nSL, in unphased and phased data. The corresponding scores were normalized using the tool “norm” provided by Selscan and then the results were plotted in R as a Manhattan plot to better visualize them (Figures 15-18). It is visible in all Kyrgyz groups that the results with phased data seem to present less noise than when performed with unphased data, however, both approaches are adequate. Moreover, it was seen that the nSL plots have a more dispersed distribution than the iHS plots. Specifically, Kyrgyz B1, B2, and C Manhattan plots, for both iHS and nSL, show a more equal distribution of the top hits, while the Kyrgyz A top hits (Fig. 15) seem to be more concentrated in specific chromosomes.

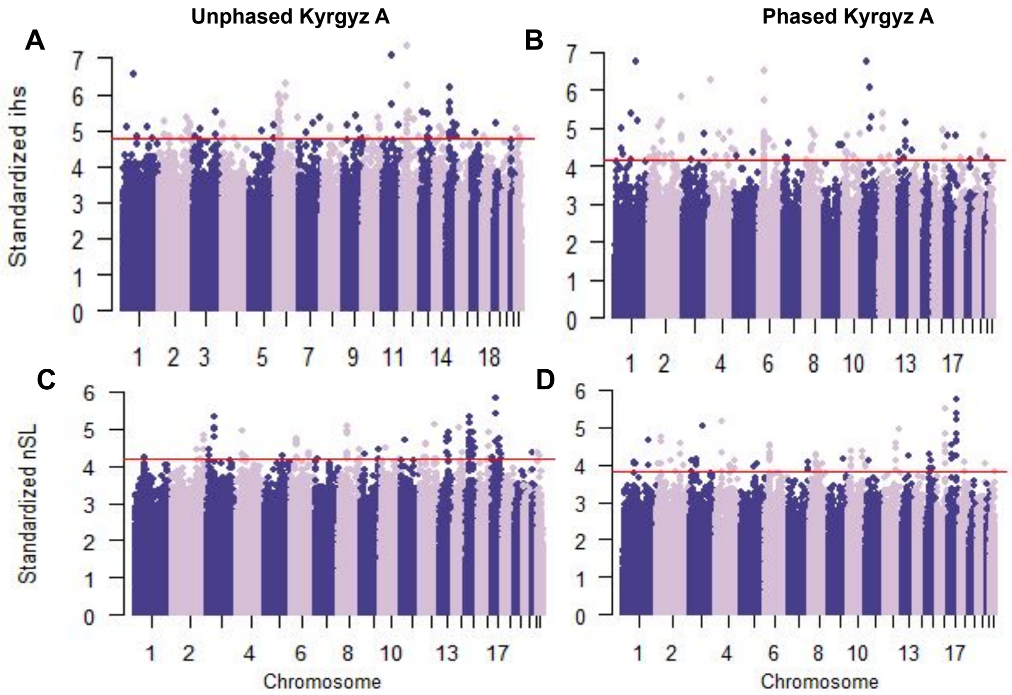


Figure 15. Manhattan plots of selection analysis in Kyrgyz A. Manhattan plots of results of both statistics, iHS (A and B) and nSL (C and D). The plots in the left correspond to the statistics performed in unphased data, and the right plots with phased data.

However, in the Kyrgyz B1 nSL graphs (Fig. 16B and 16D), a slight tendency towards chromosomes 16 and 17 are seen, and also it seems that chromosome 6 tends to have more top hits of iHS score in all clusters. Specifically in the Kyrgyz A plot of the iHS approach (Fig. 15A and Fig. 15C), a great number of SNPs from chromosome 6 seems to be under ongoing selection, with only two SNPs coming from other chromosomes in the top hits.

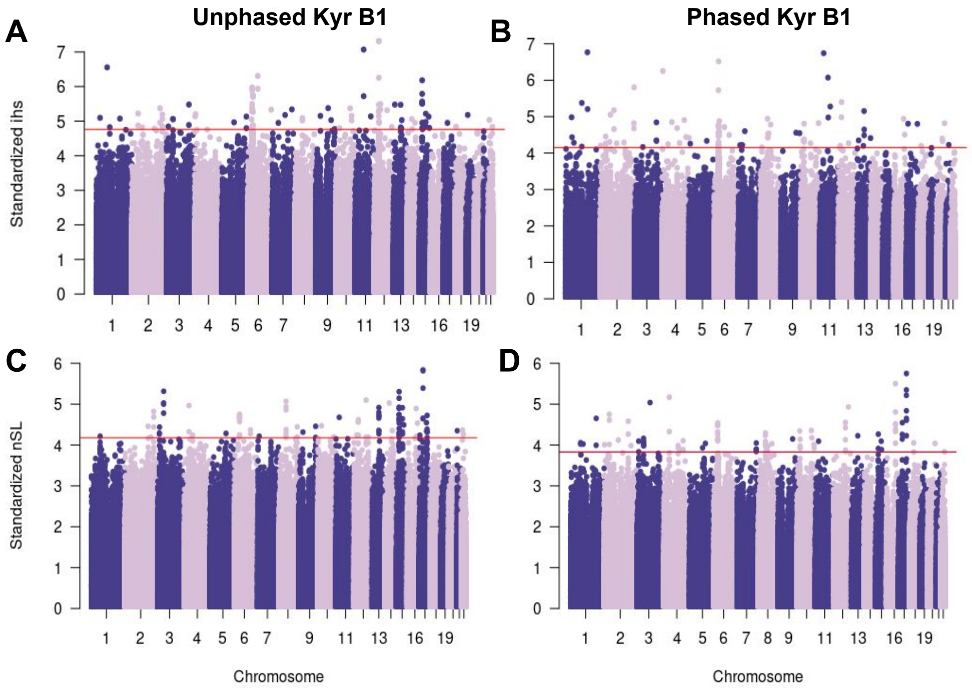


Figure 16. Manhattan plots of selection analysis in Kyrgyz B1. Manhattan plots of results of both statistics, iHS (A and B) and nSL (C and D). The plots in the left correspond to the statistics performed in unphased data, and the right plots with phased data.

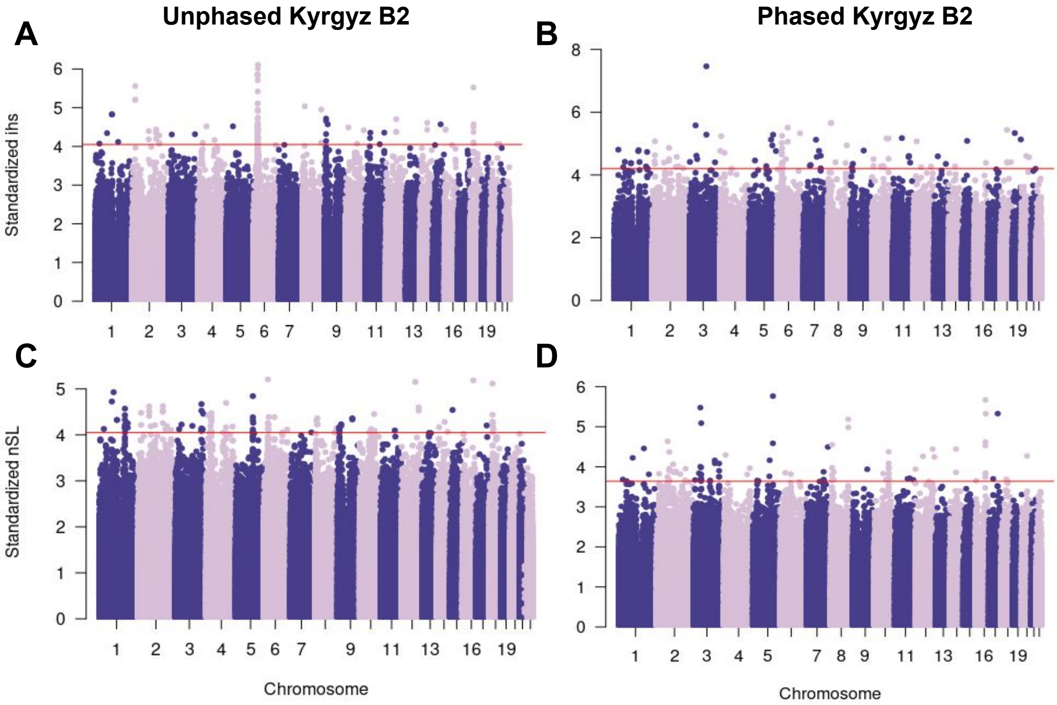


Figure 17. Manhattan plots of selection analysis in Kyrgyz B2. Manhattan plots of results of both statistics, iHS (A and B) and nSL (C and D). The plots in the left correspond to the statistics performed in unphased data, and the right plots with phased data.

The first top hits of the iHS approach in all Kyrgyz clusters had absolute values between ~7.5 and 6.7, while the lowest top hits were between ~4.4 and 4.15. The nSL absolute values are in general a bit lower than the iHS score, with the first top hits between 5.8 and 4.7 approximately. The 100th top hits had values between 3.8 and 3.6.

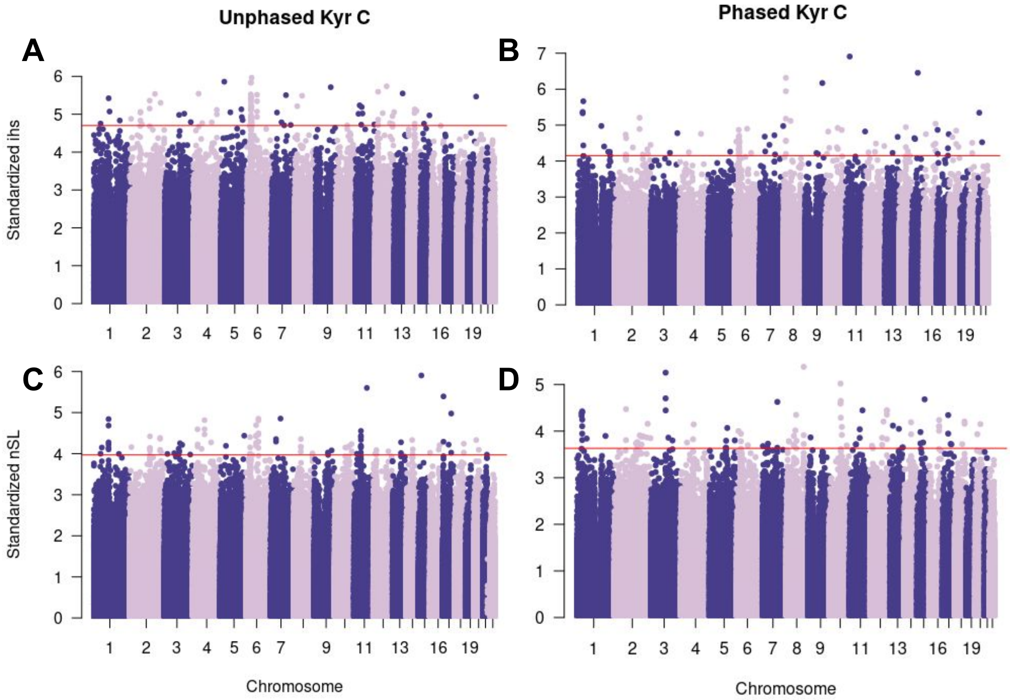


Figure 18. Manhattan plots of selection analysis in Kyrgyz C. Manhattan plots of results of both statistics, iHS (A and B) and nSL (C and D). The plots in the left correspond to the statistics performed in unphased data, and the right plots with phased data.

Next, the top 100 hits of each analysis of each Kyrgyz cluster were annotated using Ensembl. The top hits are different not only for each Kyrgyz cluster but also for each statistical method used. Out of 400 annotated hits, 26 variants have a missense coding consequence, which corresponds approximately to 6,5% of the annotated hits. All Kyrgyz groups present variants with missense coding consequences, and some of them have phenotypes associated with them. Most of these phenotypes are related to specific diseases, for example, the SNP rs3749971 in Kyrgyz B1 is linked to Lupus Erythematosus Systemic (Supplementary Table 1), which could give some hints on the putative selective pressure on that specific gene or marker.

Interestingly, some of the top 100 annotated variants from each Kyrgyz group are linked to genes related to the Behçet syndrome (Table 3), popularly known as the Silk Road disease. In total, there are 24 genes associated with the top hits in both selection statistical methods related to this syndrome. In total there are 525 SNPs related to Behçet syndrome in the dataset used to perform the selection analyses, of which 56 are among the top 100 hits.

Genes	Total SNPs	Top 100	Statistics	Group
CDSN	11	1	iHs	Kyrgyz A
C6orf15	15	1	iHs	
CCHCR1	18	1	iHs	
DDR1	13	1	iHs	
HCG27	17	1	iHs	
LTA	7	1	iHs	
HCG22	18	1	iHs	
MICB	14	2	iHs	
NCR3	4	1	iHs	
HLA-S	4	2	iHs	
POU5F1	18	2	iHs	
HCP5	43	7	iHs	
DPCR1	15	3	iHs	
MCCD1	8	2	iHs	
PSORS1C1	31	3	iHs	
MICA	18	4	iHs	
DHFRP2	7	4	iHs	
DHFRP2	7	1	nSL	
POU5F1	18	1	nSL	
HCP5	42	1	iHs	Kyrgyz B1
MICB	14	1	iHs	

CCHCR1	18	1	iHs	
DHFRP2	7	2	iHs	
HLA-L	14	1	iHs	Kyrgyz B2
HLA-DOB	22	1	iHs	
TAP2	22	1	iHs	
DHFRP2	7	1	iHs	Kyrgyz C
HLA-L	15	1	iHs	
ZFP57	10	1	iHs	
TRIM31	21	1	iHs	
C6orf15	15	2	iHs	
CDSN	11	2	iHs	
UBAC2	21	1	nSL	

Table 3. Genes linked to Behçet disease under selection in Kyrgyz groups. Table with genes associated with Behçet disease of the top 100 variants of both selection statistics (iHS and nSL) that have a missense consequence. The column "Total SNPs" is the total number of SNPs associated to the corresponding gene that are present in the dataset on which the selection statistics were performed. The column "Top 100" is the number of SNPs associated with the corresponding gene that are present in the top 100 hits.

5. DISCUSSION

5.1. Population Structure of Kyrgyzstan

According to the PCA results (Fig. 2), the Kyrgyzstan population is spread across the first PC axis, where the majority of individuals cluster in a central position between European and East Asian populations. Previous studies showed that populations in Central Asia that once were crossed by the Silk Road present genetic characteristics intermediate between Europe and East Asia, most likely to an extensive admixture of these populations (Comas et al. 1998). Moreover, this main cluster is in close proximity to other Central Asia populations like Kazakh.

This suggests that Kyrgyz people have a close relationship with other Central Asia populations, which is coherent with the literature (Guo et al. 2018). However, Kyrgyz people are separated into four groups according to the PCA: one cluster together with other Central Asia populations (as previously mentioned); one cluster more towards other East Asian populations such as Tibetans, Mongols, and Chinese; another cluster intermediate to European and Central Asia populations; and a final cluster closely grouped with the Russian population. This indicates that Kyrgyz people most likely have different demographic and genetic histories within the population. Furthermore, this result is consistent with the local anthropological history, since Kyrgyzstan was involved in major events that occurred in Central Asia that probably shaped the population structure of Kyrgyz people. For example, the part of the population that is spread towards Russians and the other spread more around Mongols could be explained by the events of the Soviet Union and the Mongol invasion, respectively. However, it is intriguing that Kyrgyz people are clearly separated in the PCA, which indicates that they are not only admixed but probably also present very distinct patterns and portions of ancestry components.

The results seen in PCA are also supported by the ADMIXTURE results, which show that the Kyrgyzstan population has a complex combination of ancestral components. However, only three different patterns in Kyrgyz people are visible in the ADMIXTURE results, not having the same resolution as the PCA results. This is most likely due to the fact that ADMIXTURE is focused on a specific set of K and it is also linked to the dataset used. The four clusters would probably be seen if Kyrgyzstan was run individually. Another factor that influences it is the sample size, which is considerably small for some Kyrgyz groups.

Additionally, the ADMIXTURE results showed that Kyrgyz people in general present a great amount of East Asian component, not specifically from Han Chinese, but also from other specific populations such as an ancestral component that seems to be more related to Buryat populations. It is also visible that European components are present within the whole population, even if in different portions, especially the Russian component. Some ancestral component associated with Iranians is also present within the Kyrgyzstan population, as well as a component seen in Pakistan populations such as Brahui and Balochi. As mentioned before, both the Russian and East Asian ancestral components present in Kyrgyz people can be easily explained by historical events. On the other hand, the historical reason for the other components is not that clear. One event that could explain these different ancestral components is the emergence of the ancient Silk Road, as mentioned earlier, studies have shown that Silk Road populations feature intricate admixture and migration patterns (Mezzavilla et al. 2014). Therefore, it was expected that Kyrgyzstan may have received gene flow from populations around Central Asia due to its closeness to what once was one of the main roads of the ancient Silk Road.

Considering the results of PCA and ADMIXTURE, it was clear that the Kyrgyz population could not be assessed as a single and homogenous group, since it would not reflect its actual genetic diversity, besides ignoring the distinct demographic history. For this reason, Kyrgyzstan was divided into four groups: A, B1, B2, and C. This separation was done after analyzing the Q matrix obtained using ADMIXTURE with the R package MClust and the PCA results. MClust was able to successfully identify three clusters within the Kyrgyz people (Fig. 5), which A the one in which most of the individuals were grouped (143 individuals), which corresponded to the main cluster grouped around other Central Asia populations in the PCA analysis, and also the major portion of the ADMIXTURE analysis. Kyrgyz C is the group that has the most percentage of Russian ancestral components and is also clearly isolated in the PCA analysis and with a unique pattern in the ADMIXTURE results. Kyrgyz B represents part of the population with the most admixed pattern with similar proportions of all ancestral components identified with the ADMIXTURE program (Fig. 3 and Fig. 4). However, the PCA results show that this group could be further divided into groups since part of Kyrgyz B is clearly spread more towards European populations and the other part, East Asian populations. MClust struggled in further dividing this group probably due to the intricate admixture pattern this individual presents, with almost equal parts of each ancestral component. Therefore, the Kyrgyz B individuals that present greater proportions of either European or East Asian components were manually identified and separated into Kyrgyz B1 and Kyrgyz B2, respectively.

The ROH results obtained from PLINK 1.9 showed that the Kyrgyz groups are lower than Souther and Central Asia populations, therefore this indicates that recent inbreeding and recent consanguineous marriages are unlikely to have occurred within these groups. However, the inbreeding coefficient obtained, Fhat2, is higher than other surrounding populations like Uzbek and Kazakh, which could be explained by a small Ne of the founder population with no recent consanguinity.

Once the Kyrgyz dataset was divided, the next step was assessing the effective population size, which does not correspond to the actual number of individuals in a population, but the number of individuals needed to explain the whole population's genetic diversity. It is usually a smaller number than the census (Frankham 1995), however, it can give powerful insights into bottlenecks and population growth over time. Only Kyrgyz A and B1 had successful results in this analysis because the program failed to find IBD segments within Kyrgyz B2 and Kyrgyz C. This is a common issue when dealing with groups with a smaller number of individuals in this cluster since the probability of these individuals sharing IBD segments is smaller.

The initial analysis using the program IBDNe showed a high number of individuals in Kyrgyz A before 100 generations (Fig. 11), around 60 thousand people, followed by a great decline, stabilizing in a lower value. Kyrgyz B1 had similar results, sharing the same pattern as Kyrgyz A, but with a considerably higher Ne. This value for Kyrgyz B1 is probably extremely high due to the small sample number (18 individuals), which decreases the accuracy of the analysis. The Ne and the steep decline are greater than expected for both Kyrgyzstan groups, and this is probably due to noise from the ancestral populations.

Recently admixed populations usually contain longer IBD segments that are derived from the ancestral source populations in different proportions. Therefore, in contrast with non-admixed populations which the whole genome can be used to characterize the individual’s population, the genome of a recently admixed population presents fragments from the different ancestral populations, and inferring them can increase the accuracy of admixture-based analyses (Sankararaman et al. 2008). Therefore, these ancestral IBD segments are also detected by programs such as IBDNe, no distinction is made between ancestral IBD segments and the target population-specific segments. In simpler words, the results shown are similar to a sum of both the Ne of the ancestral population and the target population, which produces a higher number of Ne. For that reason, in order to assess the true value of Ne of Kyrgyz people, these ancestral IBD segments need to be identified and filtered out from the analysis. The results would suggest that the population went through a bottleneck, indicated by the great decline in the graph, however, this could be just due to the artifacts produced by the admixture, as mentioned above.

The AS-IBDNe pipeline was performed in order to obtain the ancestry-specific Ne of both the European and East Asian ancestry in the Kyrgyz A group. The results obtained are coherent with what was expected, which is a smoother curve for both the ancestral populations (represented in this study by modern-day populations). Both of them do not present the steep decline seen in the IBDNe result for Kyrgyz A, instead a modest decline is observed which is co-localized with the estimated admixture time that occurred in the Kyrgyz A population. Therefore, it is possible to conclude that the accentuated decline seen in IBDNe is indeed due to the admixture, causing the observed pattern.

In conclusion, the IBDNe result of the estimation of Ne for Kyrgyz A (Fig. 11) does not reflect the true Ne of the group, since the admixture can bias the values. The AS-IBDNe results give more insights into what could truly be the Ne for Kyrgyz A before the admixture, since it accounts for the IBD segments specific to each ancestral population. Specifically, the AS-IBDNe results show that Kyrgyz A most likely did not go through a bottleneck and that the decline seen in the graph is probably due to the lack of growth of the population after the admixture took place. Thus, the ancestral-specific Ne should be taken into consideration when analyzing the IBDNe results.

5.2. Admixture in Kyrgyzstan

The results obtained from both PCA and ADMIXTURE are good indicators of gene flow and admixture that occurred in Kyrgyzstan. However, to further confirm it, F statistics, an allele-frequency-based statistic, were performed. First, F3 was used, which uses two reference populations and the target population, which can inform whether the target is an admixture of the two reference populations based on the differences in allele frequency between them.

The F3 results showed that Kyrgyzstan A, B1, and B2 are an admixture between European and East Asian ancestral populations, having significant F3 and Z scores for an F3 between Han and Russian.

Moreover, significant results were also obtained with Han and Pakistan populations; Han and Iranian. There were also significant results for F3 between Han and Middle East populations, however, the Middle East populations also present some gene flow from other ancestral populations. This could be just an indication that both Kyrgyzstan and the Middle East populations received gene flow from the same ancestral populations, and not that Kyrgyz are directly admixed with them.

The results indicate that these three Kyrgyz clusters probably have a greater predominance of Han and Russian components. On the other hand, Kyrgyz C had no significant values to any of the F3 tests performed, therefore this group probably is not admixed. A possible explanation could be that Kyrgyz C corresponds to a group that maintained isolation due to cultural and/or linguistic reasons, for example, they could still have a strong Russian recent ancestry and maintained their culture and language. Nonetheless, future studies could address the motives for the isolation of this Kyrgyz group to assess these possibilities and further understand the genetic history of Kyrgyzstan.

This is consistent with the ADMIXTURE results which showed that Kyrgyz people have a mosaic admixture pattern with several different ancestral components. As mentioned before, the gene flow from Pakistan and Iranian populations is most likely due to the emergence of the Silk Road which induced a high transit of people within its region.

The F4 results confirmed once again the admixture of clusters A, B1, B2, and C, and also provided information about which reference population is more closely similar to Kyrgyz. The East Asian component could be accurately assigned to either Han Chinese or Japanese populations as a reference. As for the Pakistani populations, Brahui or Balochi are reliable references. Finally, the Russian population is the European population most closely related to the Kyrgyz people.

These results confirm all the other analyses performed so far, however, it is important to be aware of the fact that these reference populations are only a proxy for the ancestral contributing populations since the modern populations are not the populations that once interacted with the ancestral individuals of Kyrgyzstan.

5.3. Selection in Kyrgyz people

Two different statistical approaches were performed using the program Selscan in order to investigate evidence of ongoing selection within the Kyrgyz groups. One difference seen between the Manhattan plots was that Kyrgyz A had its results more concentrated in some specific chromosomes when compared to the other groups. This could be an indication that Kyrgyz A analyses present less noise due to the number of individuals contained in the group. Since the other Kyrgyz groups present considerably fewer individuals than Kyrgyz A (40% of the overall Kyrgyz dataset), they are more prone to have artifacts, being less accurate.

The top hits of each approach and each Kyrgyz group were then annotated using Ensembl. Around 6,5% of all the annotated SNPs for all Kyrgyz groups present missense coding consequences, some of which are associated with phenotypes, usually diseases. These annotated genes could be interesting targets for future biomedical and evolutionary studies in Kyrgyz and Central Asia populations.

Particularly, some of the annotated top hits are associated with the Behçet Syndrome. Also known as the “Silk Road disease” because of its general distribution within countries where once was the ancient Silk Road (Kaklamani, Vaiopoulos, and Kaklamanis 1998; Sakane et al. 1999), it is not yet fully understood within the literature, including its causes, and it apparently presents more than one mechanism.

There are several reasons that make the Behçet syndrome so challenging to comprehend, such as its geographically different prevalence and lack of laboratory and histopathology-specific components. Therefore, clinical features are preferred to describe it, however, there is also a wide range of symptoms associated. The manifestations include major vascular diseases, eye disease, and gastrointestinal ulcerations, besides central nervous system involvement (Yazici et al. 2018).

The main gene so far associated with this syndrome is HLA-B*5 (de Menthon et al. 2009), however, another study suggests that the gene MICA is perhaps the primary gene associated with Behçet syndrome (Mizuki et al. 1997). In this study, MICA has 18 variants associated with it among the top 100 highest iHS scores (Supplementary figure 1). This signal of positive ongoing selection in Kyrgyz people in the MICA gene is certainly interesting, considering that Kyrgyzstan is located at where one of the major routes of the ancient Silk Road crossed and that, as mentioned before, the Beçmet syndrome associated with this gene is also known for being more prevalent in populations around the Silk Road. In conclusion, this could suggest that this gene probably undergoes a positive selective pressure due to the emergence of such syndrome, which can give promising insights for future studies about Kyrgyzstan and Central Asia medical and evolutionary studies.

6. CONCLUSION

This study aimed to better describe the population structure of Kyrgyz people, analyze possible admixture events, and also assess signals of ongoing selection through genotype data. Using population structure and admixture analysis of the genotype data of Kyrgyzstan, the results obtained showed that the population presents a complex substructure and different demographic and genetic histories being possible to split the population into four groups. This was confirmed by PCA, ADMIXTURE, and clustering analysis, however, further studies should be conducted to better understand the reasons why there is a clear distinction between the Kyrgyz groups.

The admixture analysis with ADMIXTURE, Alder, and F statistics, also showed that the majority of Kyrgyz people are an admixture of mainly Russian and Han Chinese ancestral populations. Other evidence of gene flow between Kyrgyz people and other populations was also found, and it is coherent with anthropological history.

Moreover, the Ne over time of Kyrgyz A was estimated using IBDNe, and then AS-IBDne was also performed to give insights into the true Ne since the ancestry from different ancestral populations interfere with the results. The AS-IBDNe revealed that Kyrgyz people most likely did not present a bottleneck, as the IBDNe suggested, and most likely stayed constant.

Furthermore, the Kyrgyzstan population presents signals of ongoing selection, which were detected by the program Selscan, and the top 100 variants with the highest scores of each statistical method were annotated. Among them, 26 variants present missense coding consequences, some associated with phenotypes. The highlight of the annotated SNP variants is the ones associated with genes related to the Behçet syndrome, a syndrome with great occurrence among populations that once existed on the Silk Road. Nevertheless, future research is needed to more accurately determine the direction of this ongoing selection and other implications of it.

In conclusion, this study grants further understanding of the population genetics of the Kyrgyzstan population, providing valuable information into the demographic and genetic history of Kyrgyz people. Population genetics is definitely a powerful tool when it comes to understanding our diversity and origins. Moreover, assessing the genetic diversity of the human genome can also facilitate and improve our medical knowledge, understanding better the association between genes and diseases or traits.

Lastly, the findings of these observations can potentially improve our knowledge of Central Asia populations' history and genome, opening new opportunities for further studies in the field.

7. APPENDIX

Population	Location
Adygei	Russia, Euroasia
Armenian	Armenia, Eurasia
Azeri	Azerbaijan, Eurasia
Balochi	South Asia
Bashkir	Russia, Europe
Brahui	South Asia
Burusho	South Asia
Druze	Middle East
GujaratiB	India, South Asia
Han	China, East Asia
Hazara	Central Asia
Iranian	Iran, Middle East
Jew Iraqi	Iraq, Middle East
Jordanian	Middle East
Kazakh	Kazakhstan, Central Asia
Mongol	Mongolia, East Asia
Palestinian	Middle East
Russian	Russia, Europe
Syrian	Syria, Middle East
Tamang	South Asia
Tibetan	Tibet, East Asia
Uzbek	Uzbekistan, Central Asia
Yemeni	Yemen, Middle East

Supplementary Table 1. Populations referenced in this study and their corresponding locations.

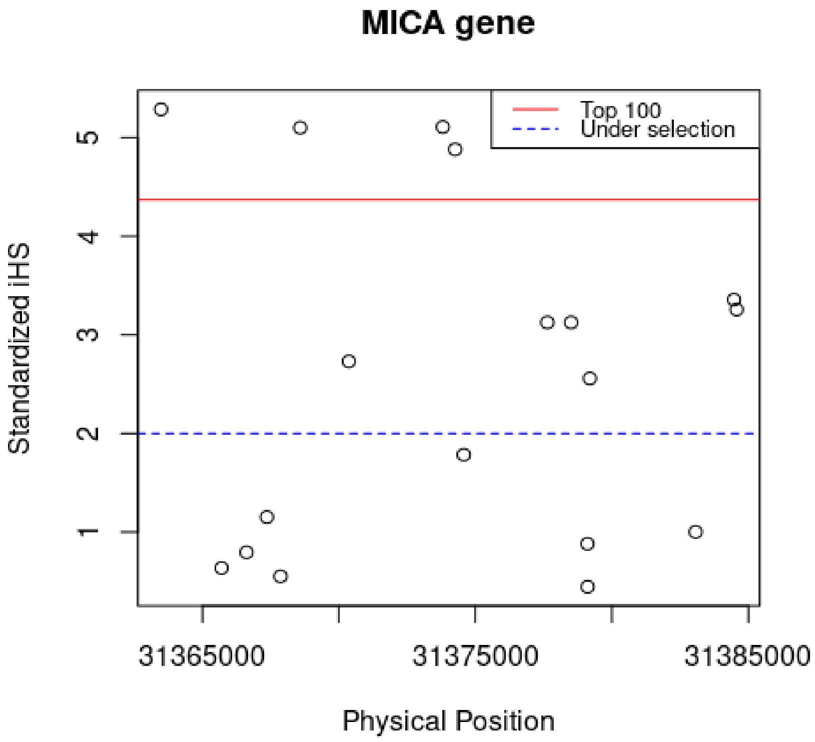
Group	Test	SNP ID	CH	Gene	Phenotype	Freq.	Freq.	Freq.	Freq.	Freq.	Freq.

			R	Symbol		KYR	AFR	AMR	EAS	EUR	SAS
Kyrgyz A	iHS	rs9263726	6	PSORS1C1	-	10.14%	16%	12%	6%	15%	5%
		rs2229092	6	LTA	Susceptibility to Leprosy, Susceptibility to Psoriatic Arthritis, Tumor Necrosis Factor Beta Levels	8.04%	0%	3%	1%	6%	3%
		rs34816476	6	NCR3	Mild Susceptibility To Malaria	2.80%	0%	0%	1%	1%	3%
		rs2736182	6	AIF1	-	10.14%	22%	8%	15%	3%	7%
	nSL	rs2240037	19	ZNF749	-	1.05%	9%	1%	2%	2%	6%
		rs2240038	19	ZNF749	-	1.05%	9%	1%	2%	2%	6%
		rs6008794	22	CELSR1	Neural Tube Defects	4.55%	74%	14%	1%	16%	13%
		rs6008795	22	CELSR1	-	4.55%	48%	12%	0%	16%	13%
Kyrgyz B1	iHS	rs3748800	1	PTPRF	Athelia, Isolated Congenital Breast Hypoplasia/Aplasia	2.78%	0%	0%	1%	0%	0%
		rs1453543	11	OR4D6	-	2.78%	25%	15%	6%	1%	2%
		rs3019198	11	CPSF7	-	2.78%	70%	16%	3%	15%	19%
		rs3749003	2	XIRP2	-	2.78%	0%	1%	8%	0%	1%

		rs9257770	6	OR5V1	-	2.78%	1%	2%	4%	3%	1%
		rs3749971	6	OR12D3	Lupus Erythematosus Systemic, Lupus Erythematosus Systemic, Lupus Erythematosus Systemic, Hemoglobin Levels, Intelligence, Schizophrenia	2.78%	1%	3%	3%	7%	1%
	nSL	rs2228561	3	COL7A1	Epidermolysis Bullosa, Nail Disorder Nonsyndromic Congenital	2.78%	3%	7%	4%	12%	25%
Kyrgyz B2	iHS	rs17258982	1	CR2	Immunodeficiency, Susceptibility to Systemic Lupus Erythematosus, Systemic Lupus Erythematosus	5.00%	6%	6%	0%	7%	5%
		rs16925541	10	TET1	-	5.00%	0%	7%	11%	6%	20%
		rs1453543	11	OR4D6	-	5.00%	25%	15%	6%	1%	2%
		rs4832524	2	KCNS3	-	5.00%	26%	19%	2%	39%	13%

		rs4911290	20	BPIFB3	-	10.00%	35%	48%	76%	39%	53%
		rs818817	3	SLC22A14	-	5.00%	11%	19%	24%	14%	16%
		rs11708527	3	RETNLB	-	5.00%	12%	23%	4%	28%	38%
		rs10075302	5	SLC25A2	-	5.00%	2%	4%	4%	5%	6%
		rs9257770	6	OR5V1	-	5.00%	1%	2%	4%	3%	1%
	nSL	rs2228561	3	COL7A1	Epidermolysis Bullosa, Nail Disorder Nonsyndromic Congenital	5.00%	3%	7%	4%	12%	25%
Kyrgyz C	iHS	rs16954698	16	PKD1L2	Serum Metabolite Levels	3.85%	4%	12%	5%	5%	9%
		rs3218600	6	REV3L	Moebius Syndrome	3.85%	3%	2%	0%	3%	9%

Supplementary Table 2. Ensembl results of all variants with missense coding consequences in each Kyrgyz group and in each test. The table also includes the Minor Allele Frequencies for each variant obtained from the Ensembl website. The phenotypes correspond to either the phenotype associated to the variant or for the gene associated with the variant.



Supplementary figure 1. MICAsnps in Kyrgyz A. Plot of MICA snps present in the Kyrgyz dataset. The red line corresponds to threshold of the top 100 hits, and the blue dashed line is the threshold value to be considered under selection.

8. BIBLIOGRAPHY

Adle, Chahryar, Irfan Habib, and Karl M. Baipakov. 2003. *History of Civilizations of Central Asia: From the Sixteenth to the Mid-Nineteenth Century*. Multiple History Series. Paris: UNESCO publishing.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64. https://doi.org/10.1101/gr.094052.109.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Broman, Karl W., and James L. Weber. 1999. “Long Homozygous Chromosomal Segments in Reference Families from the Centre d’Étude Du Polymorphisme Humain.” *American Journal of Human Genetics* 65 (6): 1493–1500.

Brookes, Anthony J. 1999. “The Essence of SNPs.” *Gene* 234 (2): 177–86. https://doi.org/10.1016/S0378-1119(99)00219-X.

Browning, Brian L, and Sharon R Browning. 2013a. “Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data.” *Genetics* 194 (2): 459–71. https://doi.org/10.1534/genetics.113.150029.

Browning, Brian L., and Sharon R. Browning. 2013b. “Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data.” *The American Journal of Human Genetics* 93 (5): 840–51. https://doi.org/10.1016/j.ajhg.2013.09.014.

Browning, Brian L., Xiaowen Tian, Ying Zhou, and Sharon R. Browning. 2021. “Fast Two-Stage Phasing of Large-Scale Sequence Data.” *The American Journal of Human Genetics* 108 (10): 1880–90. https://doi.org/10.1016/j.ajhg.2021.08.005.

Browning, Sharon R., and Brian L. Browning. 2015. “Accurate Non-Parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent.” *The American Journal of Human Genetics* 97 (3): 404–18. https://doi.org/10.1016/j.ajhg.2015.07.012.

Browning, Sharon R., Brian L. Browning, Martha L. Daviglus, Ramon A. Durazo-Arvizu, Neil Schneiderman, Robert C. Kaplan, and Cathy C. Laurie. 2018. “Ancestry-Specific Recent Effective Population Size in the Americas.” *PLOS Genetics* 14 (5): e1007385. https://doi.org/10.1371/journal.pgen.1007385.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (1): s13742-015-0047–0048. https://doi.org/10.1186/s13742-015-0047-8.

Charlesworth, D., and B. Charlesworth. 1987. “INBREEDING DEPRESSION AND ITS EVOLUTIONARY CONSEQUENCES.”

Annual Review of Ecology and Systematics 18 (1): 237–68. https://doi.org/10.1146/annurev.es.18.110187.001321.

Chen, Pengyu, Xing Zou, Biao Wang, Mengge Wang, and Guanglin He. 2019. "Genetic Admixture History and Forensic Characteristics of Turkic-Speaking Kyrgyz Population via 23 Autosomal STRs." *Annals of Human Biology* 46 (6): 498–501. https://doi.org/10.1080/03014460.2019.1676918.

Comas, David, Francesc Calafell, Eva Mateu, Anna Pérez-Lezaun, Elena Bosch, Rosa Martínez-Arias, Jordi Clarimon, et al. 1998. "Trading Genes along the Silk Road: MtDNA Sequences and the Origin of Central Asian Populations." *The American Journal of Human Genetics* 63 (6): 1824–38. https://doi.org/10.1086/302133.

Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, et al. 2022. "Ensembl 2022." *Nucleic Acids Research* 50 (D1): D988–95. https://doi.org/10.1093/nar/gkab1049.

Curik, Ino, Maja Ferenčaković, and Johann Sölkner. 2014. "Inbreeding and Runs of Homozygosity: A Possible Solution to an Old Problem." *Livestock Science*, Genomics Applied to Livestock Production, 166 (August): 26–34. https://doi.org/10.1016/j.livsci.2014.05.034.

Ferrer-Admetlla, Anna, Mason Liang, Thorfinn Korneliussen, and Rasmus Nielsen. 2014. "On Detecting Incomplete Soft or Hard Selective Sweeps Using Haplotype Structure." *Molecular Biology and Evolution* 31 (5): 1275–91. https://doi.org/10.1093/molbev/msu077.

Francis, R. M. 2017. "pophelper: An R Package and Web App to Analyse and Visualize Population Structure." *Molecular Ecology Resources* 17 (1): 27–32. https://doi.org/10.1111/1755-0998.12509.

Frankham, Richard. 1995. "Effective Population Size/Adult Population Size Ratios in Wildlife: A Review." *Genetics Research* 66 (2): 95–107. https://doi.org/10.1017/S0016672300034455.

Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164): 851–61. https://doi.org/10.1038/nature06258.

Guo, Yuxin, Chong Chen, Xiaoye Jin, Wei Cui, Yuanyuan Wei, Hongdan Wang, Tingting Kong, Yuling Mu, and Bofeng Zhu. 2018. "Autosomal DIPs for Population Genetic Structure and Differentiation Analyses of Chinese Xinjiang Kyrgyz Ethnic Group." *Scientific Reports* 8 (1): 11054. https://doi.org/10.1038/s41598-018-29010-8.

Gusev, Alexander, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. 2009. "Whole Population, Genome-Wide Mapping of Hidden Relatedness." *Genome Research* 19 (2): 318. https://doi.org/10.1101/gr.081398.108.

Jorde, L.B., W.S. Watkins, and M.J. Bamshad. 2001. "Population Genomics: A Bridge from Evolutionary History to Genetic Medicine." *Human Molecular Genetics* 10 (20): 2199–2207.

https://doi.org/10.1093/hmg/10.20.2199.

Kaklamani, Virginia G., George Vaiopoulos, and Phaedon G. Kaklamanis. 1998. “Behçet’s Disease.” *Seminars in Arthritis and Rheumatism* 27 (4): 197–217. https://doi.org/10.1016/S0049-0172(98)80001-2.

Loh, Po-Ru, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. 2013. “Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium.” *Genetics* 193 (4): 1233–54. https://doi.org/10.1534/genetics.112.147330.

Maca-Meyer, Nicole, Ana M. González, José M. Larruga, Carlos Flores, and Vicente M. Cabrera. 2001. “Major Genomic Mitochondrial Lineages Delineate Early Human Expansions.” *BMC Genetics* 2 (1): 13. https://doi.org/10.1186/1471-2156-2-13.

Mallick, Swapan, Adam Micco, Matthew Mah, Harald Ringbauer, Iosif Lazaridis, Iñigo Olalde, Nick Patterson, and David Reich. 2023. “The Allen Ancient DNA Resource (AADR): A Curated Compendium of Ancient Human Genomes.” bioRxiv. https://doi.org/10.1101/2023.04.06.535797.

Mallick, Swapan, and David Reich. 2023. “The Allen Ancient DNA Resource (AADR): A Curated Compendium of Ancient Human Genomes.” Harvard Dataverse. https://doi.org/10.7910/DVN/FFIDCW.

Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. 2010. “Robust Relationship Inference in Genome-Wide Association Studies.” *Bioinformatics* 26 (22): 2867–73. https://doi.org/10.1093/bioinformatics/btq559.

Maples, Brian K., Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. 2013. “RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.” *The American Journal of Human Genetics* 93 (2): 278–88. https://doi.org/10.1016/j.ajhg.2013.06.020.

Menthon, Mathilde de, Michael P. LaValley, Carla Maldini, Loïc Guillevin, and Alfred Mahr. 2009. “HLA–B51/B5 and the Risk of Behçet’s Disease: A Systematic Review and Meta-Analysis of Case–Control Genetic Association Studies.” *Arthritis Care & Research* 61 (10): 1287–96. https://doi.org/10.1002/art.24642.

Mezzavilla, Massimo, Diego Vozzi, Nicola Pirastu, Giorgia Girotto, Pio d’Adamo, Paolo Gasparini, and Vincenza Colonna. 2014. “Genetic Landscape of Populations along the Silk Road: Admixture and Migration Patterns.” *BMC Genetics* 15 (1): 131. https://doi.org/10.1186/s12863-014-0131-6.

Mizuki, Nobuhisa, Masao Ota, Minoru Kimura, Shigeaki Ohno, Hitoshi Ando, Yoshihiko Katsuyama, Masaaki Yamazaki, et al. 1997. “Triplet Repeat Polymorphism in the Transmembrane Region of the MICA Gene: A Strong Association of Six GCT Repetitions with Behçet Disease.” *Proceedings of the National Academy of Sciences* 94 (4): 1298–1303. https://doi.org/10.1073/pnas.94.4.1298.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. 2021. “Sustainable Data Analysis with Snakemake.” F1000Research. https://doi.org/10.12688/f1000research.29032.1.

Motuzaite Matuzeviciute, Giedre, Kubatbek Tabaldiev, Taylor Hermes, Elina Ananyevskaya, Mindaugas Grikpedis, Elise Luneau, Inga Merkyte, and Lynne M. Rouse. 2020. “High-Altitude Agro-Pastoralism in the Kyrgyz Tien Shan: New Excavations of the Chap Farmstead (1065–825 Cal b.c.).” *Journal of Field Archaeology* 45 (1): 29–45. https://doi.org/10.1080/00934690.2019.1672128.

Peng, Min-Sheng, Weifang Xu, Jiao-Jiao Song, Xing Chen, Xierzhatijiang Sulaiman, Liuhong Cai, He-Qun Liu, et al. 2018. “Mitochondrial Genomes Uncover the Maternal History of the Pamir Populations.” *European Journal of Human Genetics* 26 (1): 124–36. https://doi.org/10.1038/s41431-017-0028-8.

Sakane, Tsuyoshi, Mitsuhiro Takeno, Noboru Suzuki, and Goro Inaba. 1999. “Behçet’s Disease.” *New England Journal of Medicine* 341 (17): 1284–91. https://doi.org/10.1056/NEJM199910213411707.

Sankararaman, Sriram, Srinath Sridhar, Gad Kimmel, and Eran Halperin. 2008. “Estimating Local Ancestry in Admixed Populations.” *The American Journal of Human Genetics* 82 (2): 290–303. https://doi.org/10.1016/j.ajhg.2007.09.022.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. “Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models.” *The R Journal* 8 (1): 289–317.

Skoglund, Pontus, Swapan Mallick, Maria Cátira Bortolini, Niru Chennagiri, Tábita Hünemeier, Maria Luiza Petzl-Erler, Francisco Mauro Salzano, Nick Patterson, and David Reich. 2015. “Genetic Evidence for Two Founding Populations of the Americas.” *Nature* 525 (7567): 104–8. https://doi.org/10.1038/nature14895.

Szpiech, Zachary A. 2022. “Selscan 2.0: Scanning for Sweeps in Unphased Data.” bioRxiv. https://doi.org/10.1101/2021.10.22.465497.

Szpiech, Zachary A., and Ryan D. Hernandez. 2014. “Selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection.” *Molecular Biology and Evolution* 31 (10): 2824–27. https://doi.org/10.1093/molbev/msu211.

Taylor, William, Svetlana Shnaider, Aida Abdykanova, Antoine Fages, Frido Welker, Franziska Irmer, Andaine Seguin-Orlando, et al. 2018. “Early Pastoral Economies along the Ancient Silk Road: Biomolecular Evidence from the Alay Valley, Kyrgyzstan.” *PLOS ONE* 13 (10): e0205646. https://doi.org/10.1371/journal.pone.0205646.

Turner, Stephen D. 2018. “Qqman: An R Package for Visualizing GWAS Results Using Q-Q and Manhattan Plots.” *Journal of Open Source Software* 3 (25): 731. https://doi.org/10.21105/joss.00731.

Voight, Benjamin F., Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K. Pritchard. 2006. “A Map of Recent Positive Selection in the Human Genome.” *PLOS Biology* 4 (3): e72. https://doi.org/10.1371/journal.pbio.0040072.

Yao, Yong-Gang, Xue-Mei Lü, Huai-Rong Luo, Wen-Hsiung Li, and Ya-Ping Zhang. 2000. “Gene Admixture in the Silk Road Region of China: Evidence from MtDNA and Melanocortin 1 Receptor Polymorphism.” *Genes & Genetic Systems* 75 (4): 173–78. https://doi.org/10.1266/ggs.75.173.

Yazici, Hasan, Emire Seyahi, Gulen Hatemi, and Yusuf Yazici. 2018. “Behçet Syndrome: A Contemporary View.” *Nature Reviews Rheumatology* 14 (2): 107–19. https://doi.org/10.1038/nrrheum.2017.208.

9. ACKNOWLEDGMENTS

I would like to express my deepest gratitude and appreciation to all those who have contributed to the completion of this master's thesis. Their support, guidance, and encouragement have been instrumental in this journey, and I am truly grateful.

First and foremost, I am indebted to my supervisor Dr. Massimo Mezzavilla for their invaluable guidance, expertise, and unwavering support throughout the research process. Their mentorship, insightful feedback, and dedication have shaped the direction of this thesis and enriched my understanding of the subject.

Many thanks also go to Professor Paolo Gasparini from the University of Trieste, for sharing genetic data from the Silk Road populations, and the research from IRCSS “Burlo Garofolo” from Trieste.

I am grateful to the faculty and all my laboratory colleagues at the University of Padua, especially to Professor Luca Pagani, dr. Leonardo Vallini, and Ina Cheshmedzhieva, for providing a conducive and welcoming academic environment and access to resources essential for my research.

Thank you all for being a part of this journey and for the impact you have had on my academic and personal growth. Your support has been invaluable, and I am truly grateful for the opportunity to have worked on this thesis.