Austrian Marshall Plan Foundation Research Report

Evolutionary Anthropology - Paleogenomics and Archaeochemistry

Suzanne Freilich M.Sc.

Supervisors: Prof. Maanasa Raghavan, University of Chicago
Privatdozent Dr. Helmut Schaschl, University of Vienna

Abstract

Significant advances in next generation sequencing technologies, improved success rates of DNA sampling from ancient specimens and reduced processing costs have intensified research into the genomics of prehistoric human populations. Particular attention has recently been given to large-scale studies of Eurasia, which have revealed how migrations of early farmers, spreading agriculture, and admixture with other populations, led to significant changes in the genetics and population structure of Europeans during and after the Neolithic transition [1,2,3,4]. Relatively little is currently known, however, about the genetic ancestry and demographic shifts that shaped South Asian population history, despite it being host to a very large and culturally diverse population [5, 6]. Questions remain about the migration routes of early settlers and patterns of admixture between different groups, which led to the complex structuring seen today. This region is now garnering more attention in ancient DNA studies that take advantage of high-throughput, next generation sequencing technologies and ever-improving bioinformatics analyses to understand the complex population history of this region [6, 7]. This research endeavours to use bioinformatics tools to investigate the population history of South Asia from ancient genomes sampled in recent years, and close the gap in our understanding of human prehistory in this region.

Table of Contents

List of Tables

Table of Figures

Introduction

The last decade has witnessed significant progress in the technologies used for sequencing DNA. This had had the knock-on effect of much better rates of sequencing success at lower costs and faster turnaround, leading to an explosion of new data that is sourced from ancient specimens, which are notably more difficult to sequence than present-day DNA. Thanks to these advances, the field of genomics that explores human prehistory has grown immensely. In particular, many studies have been conducted about Eurasia in order to help answer long-held questions such as whether technology such as agriculture was introduced into Europe via the transfer of ideas or the migration of early farmers from Anatolia, who then mixed with local hunter-gatherers and brought with them a sedentary way of life. The latter has been upheld by ancient DNA studies, which evidences a change in population structure accompanied by sedentary lifestyles and agriculture as part of the Neolithic transition in Europe [1,2,3,4]. Further significant population changes also occurred at the end of the Neolithic, when pastoralists from the Pontic-Caspian steppes migrated westwards into Europe, again mixing with the local populace. Studies have shown a significant amount of steppe ancestry present in the DNA of people in Europe from the late Neolithic onwards, and this can be attributed to these migrations [1,2,3].

While we now know a lot about the genetic composition of prehistoric Europe, data from South Asia is sparse. Relatively little is known to date about South Asian genetic ancestry and the demographic shifts that shaped population history in this part of the world. Hence, questions remain about the migration routes of early settlers and patterns of admixture between different groups, which led to the complex structuring seen in today's large and diverse South Asian populations [5, 6].

A significant factor affecting the availability of ancient DNA data from South Asia in comparison to other regions such as Europe is the challenging climatic conditions. The preservation of DNA over time is much poorer in hot and humid climates, and the increased degradation of the DNA makes extraction of useful data from it extremely challenging. Despite this, a number of recent studies have been conducted that have successfully generated new datasets for the Bronze Age until historical

periods, covering a vast geographical range within India and across numerous ancient cultures [8,9,10]. The goal of this project is to apply statistical and bioinformatics tools to these large, new datasets as part of a new pilot project about South Asia, in order to fill this geographic gap and reconstruct population migrations and admixture in the region. With the advancement of high throughput genome-wide aDNA techniques, the aim is to reconstruct lineages and assess evidence of genetic structures and identify temporal continuity or change. This study will help fill the paucity of knowledge we have about South Asia, and will elucidate demographic influences that shaped population prehistory, and help explain the patterns of population genetic structure seen today. To achieve this goal, datasets from whole genome sequencing data will be analysed and evaluated.

DNA and population histories

Similarities between individuals or populations can be suggestive of being closely related biologically as a result of a more recent population separation, or due to interbreeding. In the past, phenotypic variability was believed to be a strong indicator of biological diversity, and such differences were used to define relatedness between groups. However, reliance upon the phenotype alone can be misleading for many reasons. A phenotype is the outcome of both environmental influences and your DNA, or genotype. By looking at genotypes, we can often obtain a much more direct indication of the biological affinities of different populations. Many present-day human populations have been genotyped, and these genotypes correlate to some degree with geography. Observing ancient genotypes has become possible with the development of ancient DNA technologies both in the wet lab and in terms of specialised bioinformatics tools. Analysing ancient DNA directly sampled from archaeological specimens allows us to analyse the genotypes of people in the past before more recent genetic changes have taken place due to such processes as interbreeding. This allows us to then make inferences about migration histories and interactions between different groups of people in the past that may have contributed to present-day genotypes, as well as make discoveries about genotypes that may no longer be represented in today's populations.

Over the past 30 years, this type of research has revolutionized the field of archaeology, bringing forth an understanding of past migrations and interactions that could not have been achieved with any other method. The first study to report the successful extraction and sequencing of DNA from an ancient organism came in 1984 [11]. In that case, the DNA was extracted from an extinct South African equid, the quagga, similar to a zebra that had died about 150 years prior. This study was followed by several more by Svante Pääbo and colleagues in Germany, showing that this could also be done in human mummies that dated as far back as several thousand years [12]. Despite some early set-backs, the field has developed effective strategies for dealing with issues such as contamination and degradation of ancient DNA samples that can hamper efforts (discussed in more detail below), enabling researchers to recover and analyse very small quantities of surviving DNA from many thousands of years ago.

Measuring variability in DNA

Over time DNA collects harmless errors called mutations, or more accurately, single nucleotide polymorphisms (SNPs, pronounced ‘snips’). These are heritable and the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may be the replacement of the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. There are roughly 10 million SNPs in the human genome (three billion base pairs). Most commonly, these variations are found in the DNA between genes and so have no effect on health or development. Geneticists can compare DNA sequences among individuals or populations; those who are closely related will share more of the same SNPs than those who are more distantly related. In this way, the accumulation of SNPs stored in DNA is a record of the genetic history of a population. People who share the same mutations typically share a common ancestor. Such genetic information about individuals and populations histories can be garnered from three different sources of DNA. Most informative is the nuclear DNA, which is contained in our autosomal chromosomes. The Y chromosome is only inherited from father to son however, and so this DNA contains information about patrilineal ancestry passed down the

male line. In addition, mitochondrial DNA is a separate circular genome inherited by children only from their mother, and so can be used to understand a populations' maternal ancestry.

Ethical considerations when working with ancient DNA

The analysis of ancient DNA requires the handling and destructive sampling of human remains, usually bone and teeth, and this necessarily warrants ethical consideration. A review of the literature reveals a vacuum with regards treatment of archaeological samples that are used as objects of scientific investigation. Some institutes have produced their own voluntary guidelines, which encourage consideration of the appropriateness of the proposed level of invasiveness weighed against scientific progress and societal impact [13]. Such a code of ethics and other published guidelines about destructive sampling are recommended as a benchmark for ancient DNA research [14].

Analyses derived from non-destructive sampling methods cannot offer comparable insights into an individual or population's genetic history. The ability to directly look at the human past whilst developing improved techniques are worthwhile scientific outcomes of high impact, which will contribute significantly to scientific progress. As a steward of human remains, scientists endeavour to treat them with respect, and mitigate against unnecessary invasiveness and negative impacts on future research in the following ways; a pilot study can be carried out before applying the methods on a larger scale; the selection of a sampling location on the specimen should avoid areas important for sex or age identification or anatomical measurements, and should not have a visual impact if the remains are considered for future exhibit display; sampling will be fully documented for future reference, and if necessary, digital scans and replicas can be produced for documentation before destruction; one should take care to sample a suitable amount that allows for additional future experiments to prevent the need for re-sampling.

Posthumous harm to body and personal identity is another issue in terms of an individual's right to an undisturbed grave [15]. There is necessarily a lack of knowledge surrounding the individual's consent to scientific research, and therefore attempts should be made to investigate the identity of

the person living within their socio-cultural context. Understanding these individuals in the wider context of their society is an integral part of this type of research and they are not presented as purely biological data. It is also important to consider where specimens came from originally and if proper agreements were in place for their export to another country's museum, for example, and where possible, inclusion of the country of origin should be sought as collaborations on the study, as well as repatriation after research has concluded.

Furthermore, sex and gender dimensions are also a consideration when designing an ancient DNA study. While the availability of specimens may be more reliant on the state of their preservation, one should still strive to obtain data for male and female specimens, and answer research questions that encompass both. An early step in data analysis involves molecular sexing of the samples, therefore providing opportunities to look at gender dimensions in a way that has not been achieved before. For example, questions about sex-biased mobility and residency patterns, differences in social status, mortuary practices, nutrition, disease and gender-based activities in subadults and adults alike can be co-analysed with ancestry. Sex-specific haplogroup markers from mitochondrial DNA and the Y-chromosome are thus also important in addition to data garnered from nuclear DNA.

Sources of ancient DNA

DNA survives best in teeth and the dense part of the petrous bone where the cochlea of our inner ear is located. One reason for this is that the inside of the petrous and teeth are less prone to environmental contamination. Therefore these are the most popular places to sample ancient DNA from. However, ancient DNA can also be sampled from a variety of other substrates that can also help us to reconstruct ancient palaeoenvironments, microbiomes, diet and disease. For example, dental calculus, which is fossilised plaque that builds on the tooth surface, contains DNA not only from the host organism, but also from animals and plants that were eaten, and can therefore be an indicator of diet. In addition, it will contain DNA from pathogens and the oral microbiome. Pathogen DNA can also be found in the dentin of teeth, whilst soil is also often used for ancient DNA studies where the DNA from humans, animals, plants and microbes can be preserved. Archaeobotanical

remains such as charred seeds or barley grains, stone and ceramics, hair, coprolites and shells are further examples of sources of ancient DNA [16]. In this study, largely the petrous bone as well as teeth were used for DNA sampling, and occasionally long bones.

Working with ancient DNA: challenges and solutions

Ancient DNA is a lot more challenging to work with than modern DNA, and therefore a whole active sub field of research has developed around it that requires rigorous, time consuming, and sometimes expensive experimental and computational methods that are well beyond what is needed when working with modern DNA. Two primary limiting factors on the survival of DNA from ancient periods are time and climatic conditions. DNA degrades over time, breaking into smaller fragments, and this is further accelerated by high temperatures. Many studies therefore focus on analysing ancient DNA from countries that have a temperate climate as the DNA survives better for longer, and specimens from cold climates and permafrost offer the best preservation. Thus ancient DNA from early prehistory is scarce from places such as South Asia that has a hot and humid climate, limiting the number of available studies. However, successfully analysing ancient DNA from specimens from hot climates is increasing with the many technological advances that are being made in the lab and computationally, which offer a much higher chance of successfully sequencing even minute fragments of DNA [17].

Another drawback however, is that when DNA fragments, it also accumulates post-mortem errors which can look like mutations. Now it is possible to correct for these errors enzymatically, which restricts the damage to the ends of the sequences. This can be used to authenticate the DNA as ancient and rescue the majority of the DNA sequence for population genetic analysis, while the ends can be trimmed off.

Another major problem to overcome is that ancient DNA is easily contaminated by other ancient and modern DNA molecules, both human and bacterial, as well as other organisms in the burial matrix. This problem is further exacerbated by the fact that only small quantities of degraded DNA are

preserved in comparison to modern DNA. Because of this, ancient DNA laboratories must follow stringent protocols to prevent contamination from other sources of DNA. UV radiation of specimens, bleaching of surfaces, filtering of air and donning of personal protective clothing is required to prevent exogenous DNA contaminating the ancient samples. During the preparation of samples for sequencing, molecules are barcoded so that they can be identified as the only DNA endogenous to the sample. Computational methods are also used after sequencing to align the reads to the human genome while discarding any other contaminant reads, and those that do not show the typical signs of damage mentioned earlier, which would indicate that it is not of ancient origin. However, ancient contamination can still remain, and the amount can be estimated with methods that harness the more abundant mitochondrial DNA. If contamination estimates are high, those samples or reads can be excluded from downstream analysis.

Where quantities of ancient DNA are very low, modified sequencing methods can be used. For example, a single stranded method has been developed to capture very short fragments found in highly-degraded samples which can recover more DNA that would otherwise be lost. In addition, hybridisation-capture methods can be used to only amplify a specific set of SNPs that are commonly used in population genetic studies rather than the whole genome, increasing the amount of data available as well as the reliability of results.

These strategies have been applied in the preparation of the samples used in this study and will be discussed further below.

Materials and methods

Description of datasets used for this study

Figure 1 Map showing the location of samples for the study. Showing the Swat and Chitral regions of Pakistan and surrounding countries where the prehistoric and present-day samples in this study come from..Google. (n.d.) [Google Maps image of South Asia]. Retrieved 21 April, 2020, from https://www.google.com/maps/

This project investigates the genomics of over one hundred samples from northern Pakistan which come from Late Bronze Age-Iron Age (referred to here as Iron Age or IA) and historical settlements (referred to here as Medieval and Historical or H) located in the Swat and Chitral districts (see map in Figure 1). Skeletons were recovered from twelve different sites in either single or double burials [10]. Fifty of the 132 individuals were radiocarbon dated, which places the assemblage within the date range of between about 1200 BCE to 1700 CE (see Table 1). Twelve individuals were identified as first degree relatives of other individuals and were excluded from population genetic analysis. Other individuals filtered from the dataset for this study include four that were low coverage, and three identified as genetic outliers, leaving 113 samples.

Analysis group	Date range	No. samples after filtering	Male/Female
Pakistan_IA (Iron Age)	1000-800 BCE	89	46/43
Pakistan_H (Historical)	750 BCE -1400 CE	19	8/11
Pakistan_Medieval	700-1650 CE	5	5/0

Table 1 Summary of ancient samples included in this study from northern Pakistan, ref. [10].

Site ID	No. individuals
Aligrama2_IA	3
Arkotkila_IA	1
Barikot_IA	4
Butkara_IA	4
Gogdara_IA	2
Katelai_IA	30
Khyber.Pakhtunkhwa_LBA	1
Loebanr_IA	32
Udegram_IA	12
Aligrama_H	3
Barikot_H	3
Butkara_H	3
Saidu_Sharif_H	10
Barikot_Medieval	1
Parwak_Medieval	1
RajaGira_Medieval	1
Singoor_Medieval	1
Udegram_Ghaznavid_Medieval	1

Table 2 Northern Pakistan groups broken down according to archaeological site, ref. [10].

Additional, newly-sequenced ancient samples from the region were co-analysed to inform about ancestry, and this includes samples from Iran and southern Central Asia, sampled from modern-day Turkmenistan, Uzbekistan, Tajikistan, Afghanistan and Krygyzstan (ranging from 12000 – 1 BCE from the Mesolithic to Iron Age), as well as further samples from Copper and Bronze Age Steppe pastoralists dating to 3400 to 800 BCE and Siberian hunter-gatherers from 6400 to 3900 BCE [10]. These were sampled from forest and steppe regions in Kazakhstan and Russia. In order to fully investigate the genetic history of South Asia, this dataset was merged with a dataset of present-day South Asian populations composed of a large array of different ethnogeographic groups from [18], as well as further sets of previously published present-day and ancient genomes from across Eurasia following [10].

Wet lab methods used for processing the ancient samples in the lab

Following [10], DNA was sampled from the petrous bone and prepared in ancient DNA clean rooms. The bone was either drilled or sandblasted and milled into a powder. DNA was then extracted using standard protocols that first uses enzymes to digest the collagen, and then the DNA is bound to silica beads, washed and eluted for storage. The DNA extract was then built into double-stranded libraries using Uracil-DNA Glycosylase (UDG) treatment which reduces the cytosine to thymine damage normally found in degraded ancient DNA. This treatment preserved the damage at the ends of the DNA fragments to allow for verification later that the molecules are indeed ancient in origin. Before sequencing, the resulting DNA libraries were enriched both for mitochondrial DNA and for approximately 1.2 million positions in the nuclear DNA that are commonly targeted for population genetic analysis (known as the 1240K panel). After enrichment, libraries were tagged with indexes which provides them with a unique identifier. This means the libraries could be disaggregated according to sample after pooling together for sequencing. The enriched DNA libraries were then sequenced on an Illumina NextSeq500 machine.

Bioinformatics pre-processing and quality control

The first step following sequencing of the ancient DNA samples involved demultiplexing the pooled libraries based on their unique index identifiers. Following this, adapters were removed, and libraries from the same sample were merged with a minimum overlap of 15 base pairs. BWA [19] and SAMtools [20] were used to align the sequences to the human reference genome using the hg19 GRCh37 build. Filters were applied to the sequences which included a minimum mapping quality of 10 and a minimum base quality of 20, while the last two nucleotides of each sequences were trimmed in order to avoid including artefacts caused by ancient DNA damage in subsequent data analysis. One sequence was randomly selected to represent every individual at each SNP position to produce pseudo-haploid genotypes for this low coverage ancient data.

Next, the authenticity of the ancient DNA was assessed by measuring how much damage was present in the first nucleotide of each sequenced read. Libraries were considered possibly

contaminated/not of ancient origin if a UDG-treated library had a rate of less than 3% cytosine to thymine substitution in the first nucleotide. If the library was non-UDG-treated then 10% was used as the cut-off.

Contamination was estimated using contaMix [21] which assesses polymorphism in mitochondrial DNA, which is expected to only carry one haplotype. In addition, as males have only one copy of the X chromosome, ANGSD [22] was used to assess contamination by estimating polymorphism on the X chromosome in males.

Samples that were covered by less than 15,000 SNP positions were excluded from downstream analysis, as well as any first degree relatives. Genetic groupings represented by only one individual were also discarded.

Merging of datasets

Present-day datasets that were merged with this and other ancient datasets mentioned above [see 10], were genotyped at approximately 600,000 SNP positions of the Affymetrix Human Origins array, therefore, after merging, only intersecting SNPs were retained and used for downstream analyses.

Past population structure and migration

In the first work package, population genetic analyses were conducted in order to explore the genetic variation present within South Asian prehistoric populations of northern Pakistan from the Iron Age to historical periods, and their affinities to other relevant ancient and present-day populations were investigated. The prehistoric Pakistan populations were then modelled as a mixture of distally-related source ancestries which would have resulted from admixture and migration events.

Bioinformatics pipeline for population genetic analysis

The initial step in this section is exploration and visualisation of genetic structuring using dimension reduction techniques such as Principal Components Analysis (PCA) [23], UMAP (Uniform Manifold Approximation and Projection) [24] and admixture analysis [25]. This allows comparison of the South Asian samples to other ancient and modern samples to understand their geno-geographic affinities and ancestry components. Following this, patterns that may have become apparent in the visualisation are tested using formal *f* statistics to statistically test genetic affinities between populations. This is achieved by using outgroup f_3 statistics and f_4 statistics [26]. These tests then guide the modelling of the populations of interest as mixtures between probable source ancestries [26]. These steps are explained in more detail below.

Principal Components Analysis (PCA)

PCA is a useful dimension reduction technique which allows us to visualise geno-geographic relationships between individuals or groups of individuals. PCA enables us to increase the interpretability of large, complex datasets whilst retaining the statistical information, or variability in the data as much as possible, by creating new, uncorrelated variables that maximise variance present in the original data. It uses autosomal genotype data as input and gives results as eigenvectors, or principal components, and eigenvalues. The amount of genetic variance explained by a particular eigenvector is equivalent to an eigenvalue divided by the sum of all eigenvalues.

The first two principal components account for most of the observed variance, thus, usually the first two PCs/eigenvectors are the ones plotted as the x and y axis. Similarity between individuals is indicated by the relative genetic distance between them, and an individual's location in the PCA space corresponds to some degree with geographical location. Thus relationships of the new genomes to other ancient and present-day populations can be visualised in this way on the PCA axes and helps to inform further downstream analyses regarding geographic and genetic relationships.

Since ancient genomes are usually low coverage, a principal components analysis is generated by constructing the first two eigenvectors from high quality (high coverage) present-day human genotype datasets (for example genotyped on the Affymetrix Human Origins array that contain about 600,000 SNPs), while the low coverage ancient genomes that have passed quality thresholds, such as having a minimum number of autosomal SNPs of at least 10,000 SNPs, will be projected onto these first two eigenvectors together with other published ancient genomes.

Options implemented in PCA

lsq project - dealing with missing data in ancient genomes

Because the ancient samples have missing SNP data due to poor preservation, they cannot be used to calculate the principal components, which requires high coverage genotype data. Instead the ancient data is projected onto the first two principal components. The default method of orthogonal projection does not perform well when there is a lot of missing data because it uses allele frequency averages in the populations used for the PCA to fill in the missing data, and this will be inaccurate. In order to achieve more accurate results, an alternative method for performing projections was developed which uses least squares equations, and this is implemented with an option called lsq project [26].

shrinkmode

Because some samples are used to compute the PC eigenvectors while others are projected onto it, another problem arises in that the samples used for calculating the PCs have the effect of stretching the axes, so that samples that are genetically the same will appear to be placed differently in PCA space if one is used for calculating the eigenvectors and the other is projected. The use of the shrinkmode option corrects for this problem. This option uses a lot of computational resources however, and an alternative option called autoshrink can be implemented which has similar, albeit not identical results [26].

Uniform Manifold Approximation and Projection (UMAP)

In addition to performing traditional PCA, a more recent dimension reduction technique has been developed, called UMAP, which is similar to t-SNE, which stands for t-distributed stochastic neighbour embedding. All these dimension reduction methods aim to take high dimensional data and create a low dimension representation of it while preserving the structure. However, PCA is linear, while UMAP can accommodate non-linear structure in high dimensional data as well. Two points that are near to each other in high dimensional space will have increased probability of being close in low dimensional space too. UMAP can also preserve global structure and balance it well with local structure. In particular, it can be useful for more clearly separating apart local clusters of similar groups from each other, which can increase resolution among similar genetic clusters. While with PCA only the top two PCs are usually considered, all PCs computed from PCA can be used as input to UMAP, therefore retaining more information about variability for each sample. UMAP works by creating a radius around each point and connects points when they overlap. If the radius is too small, many isolated clusters will form, whereas if the radius is too big, everything will be connected to each other. A radius is chosen by considering how far each point is to its nth nearest neighbour. As the radius gets bigger, the likelihood of connection decreases. Balancing global and local structure is controlled with the options n_neighbors and min_dist.

Options implemented in UMAP

n_neighbors

This parameter indicates the number of nearest neighbours used for making the high-dimensional graph in the first step of the projection, which influences the balance between global and local structure. If you use low values, this means you are reducing the number of nearby points that UMAP considers and it leads to a more local structure being emphasised. If high numbers are used, this will focus more on a global structure and finer details will be lost, as more neighbouring points are connected together. Low values of n_neighbors can also create false clustering when in fact it may be random noise [24].

min_dist

This parameter affects how tightly points are clustered together in low-dimensional space in the second step of the projection, and refers to the minimum distance between points. Points are more tightly clustered with low values, whereas higher values will create a looser clustering and points are more spread out, thus placing more focus on the broad structure. [24]

While distance between points or clusters of points in PCA space indicates genetic distance, in UMAP, this is not the case, as projecting to lower dimensions creates distortion. In addition, the size of clusters in relation to others also do not reflect true differences in sizes. While PCA is deterministic, UMAP uses a stochastic algorithm in that the projection results will differ slightly each time it is run, and it may be worth running it a few times with different parameters to understand the data better.

Admixture analysis

Admixture is a term to describe the process whereby the offspring of interbreeding populations will exhibit a mix of alleles from their ancestral populations. It is helpful to visualise the resulting population stratification based on differing allele frequencies, and understand similarities and differences between populations and individuals. In addition it helps elucidate past mixing and migration events. Therefore, estimation of individual admixture coefficients will be performed using software that implements maximum likelihood estimation methods, in order to infer ancestral fractions present in each sample based on the autosomal SNP genotype datasets [25]. An unsupervised clustering approach will be used whereby a reference dataset comprised of a large number of high coverage, worldwide present-day samples, together with ancient samples of interest will be co-analysed, and each individual plotted as a stacked bar chart showing their estimated ancestral components, such as hunter-gatherer and farmer ancestry. The number of ancestral components present in an individual is represented by K, and the best fitting K can be chosen by performing cross-validation, where the K with the lowest cross-validation error will indicate it is the most suitable K. An unsupervised approach means that one does not specify the ancestries of the

reference data. Ancestry fractions, Q, and the allele frequencies of inferred ancestral populations, P, are output by the software, and the Q estimates are used to plot stacked bars for each sample with different colours indicating different ancestral components. To mitigate against linkage disequilibrium, which is higher in more recently admixed populations, the dataset is thinned before running the software. This is achieved by pruning the dataset by removing SNPs that have an observed sample correlation coefficient with another SNP above a certain threshold within a sliding window of a certain number of SNPs. For results to be informative, samples used in the analysis should not be from closely related individuals, and should meet the threshold for a minimum number of SNPs.

Formal *f* statistics

While the aforementioned analyses can provide indications of individual ancestral fractions and population structure, formal statistics need to be performed to formally test affinities between populations with allele frequency correlations.

Outgroup f_3 statistics

Outgroup f_3 statistics formally tests inferred population relationships by assessing allele frequency correlations. This is a three-population test where shared drift is measured between the population of interest and other reference ancient or present-day populations since their split from a common ancestor (an outgroup) in the form:

$$f_3\text{(X, test; outgroup)}$$

where X refers to the ancient genome or population of interest, in this case the ancient Pakistan populations; test represents a number of other ancient and present-day populations you wish to compare X to, and the outgroup is a population that is distantly-related to both X and test, in this case an African population, Mbuti. The results of this test, including standard errors, shows the populations who have the closest affinity to the ancient Pakistan groups.

f_4 statistics

Once the top matches for the ancient Pakistan populations are obtained from the previous test, information about the direction of gene flow between different populations can be derived by performing a further formal statistic called f_4 statistics, of the form:

f_4 (outgroup, X; test_A, test_B)

This is a four-population test very similar to D-statistics, where a negative score indicates that X, the population of interest, has more affinity to test_A, which is one of the reference populations. Conversely, a positive value indicates that gene flow occurred between X and test_B, another of the reference populations. In other words, this tests for symmetry between pairs of populations, and indicates whether pairs of populations form a clade in comparison to another population. A resulting Z score which is equal to or more than 3 or -3 is considered significant.

Inferring admixture weights with qpAdm

Next, the admixture history of the ancient Pakistan populations are statistically modelled using the genome-wide allele frequency data. This method quantifies the proportion of ancestry derived from a set of two or more likely source populations. A set of source, or *left*, populations are used to model the target population in conjunction with a set of reference, or *right* populations that they are differentially related to. The selection of left and right populations needs to be done carefully as it must be ensured that no gene flow could have occurred after the admixture event between the reference and source populations. This method is based in f_4-statistics, whereby a matrix of *f*-statistics is produced for all pairs of left and right populations in the form:

$f_4(\text{Left}_i, \text{Left}_j; \text{Right}_k, \text{Right}_l)$

Because SNP coverage in the ancient samples can be low, the option *allsnps* can be implemented so that all intersecting SNPs are used for each *f*-statistic rather than intersecting SNPs of all the left

and right populations overall, thus increasing coverage for each statistic calculated. The result provides a chi-squared p-value indicating goodness of fit of the model. The p-value is obtained using a likelihood ratio test, which tests whether any further ancestral populations are needed when including the target population in the list of source populations. Admixture models with a p-value above 0.05 are considered as feasible models that fit the population history, and a percentage of each source population is given.

Admixture graph modelling with Treemix

For the second work package, this method builds a model tree of population relationships based on the *f* statistics, where observed populations, ancestral populations and genetic drift that separates them from each other is depicted, showing population merges and splits [27]. Timing and number of past admixture events can also be estimated. The results help to understand where the Iron Age Pakistan populations fit in relation to other populations in the region as well as non-South Asian populations, and the demographic shifts that contributed to the patterns of genetic structure and variation seen today.

Results

PCA

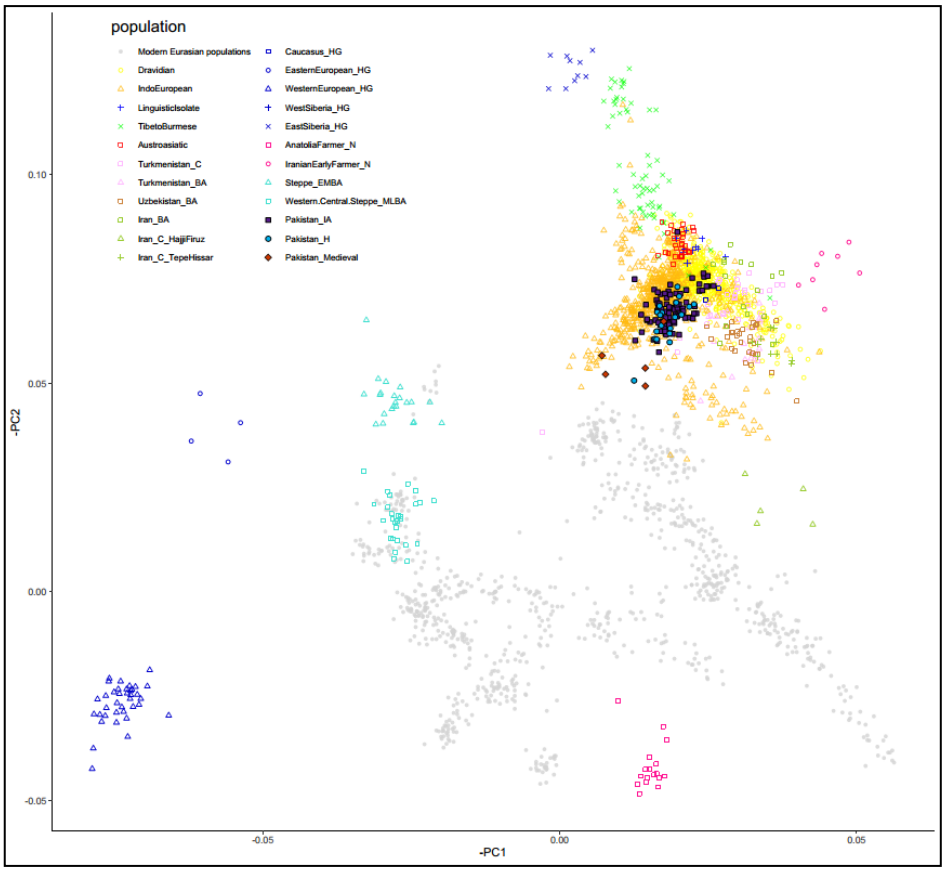

Figure 2 Eurasian PCA. Eurasian PCA highlighting present-day South Asian populations coloured by language group, and ancient populations as points. Pakistan populations are indicated with filled points.

PCA was run with default parameters using smarpca version 16000, which is included in EIGENSOFT package ver.6.0.1 [23]. Additional options of lsqproject:YES, shrinkmode:YES and numoutlieriter:0 were used in order to project ancient individuals and the modern-day South Asian populations onto PCA space constructed from the first two PCs of a Eurasian dataset made from 1340 individuals belonging to 991 present-day populations genotyped for approximately 600,000 SNPs of the Affymetrix Human Origins array [10, references therein, 18] . Related and low coverage individuals were filtered.

The ancient Pakistan populations were labelled as three macro groups, Pakistan_IA, Pakistan_H and Pakistan_Medieval based on time period, whilst five main linguistic groups from present-day South Asian populations from [18] that overlap in PCA space were plotted; Dravidian, Indo-European, LinguisticIsolate, TibetoBurmese and AustroAsiatic. Other relevant ancient populations were grouped according to country or region and main time periods such as hunter-gatherer (HG), Neolithic (N), Chalcolithic (C), Bronze Age (BA – EMBA denoting Early to Middle Bronze Age and MLBA denoting Middle to Late Bronze Age) and clustering on the PCA, with Western_Central_Steppe _MLBA comprising samples from Kazakhstan and Russia.

The PCA is anchored by Western European hunter gatherer populations in the lower left in PCA space, Neolithic Anatolian farmer populations in the lower right of the PCA, Bronze Age steppe populations further up along PC2, Eastern Siberian hunter gatherers at the top, and Neolithic Iranian early farmers to the right of the PCA. The ancient Pakistan samples from the Iron Age and historical period cluster tightly, appearing left along PC1 from Iranian early farmers, and there is a large overlap of the Pakistan samples from the three different groups with present-day South Asian populations from Indo-European and Dravidian linguistic groups. The location of the cluster of ancient Pakistan samples suggests mixing of Iranian-related ancestry and Steppe-related ancestry. Four of the five Pakistan_Medieval samples fall lower down along PC2 and slightly further left along PC1 from the other Pakistan samples. Despite the small sample size, this could suggest there is a shift towards increased steppe-related ancestry.

UMAP

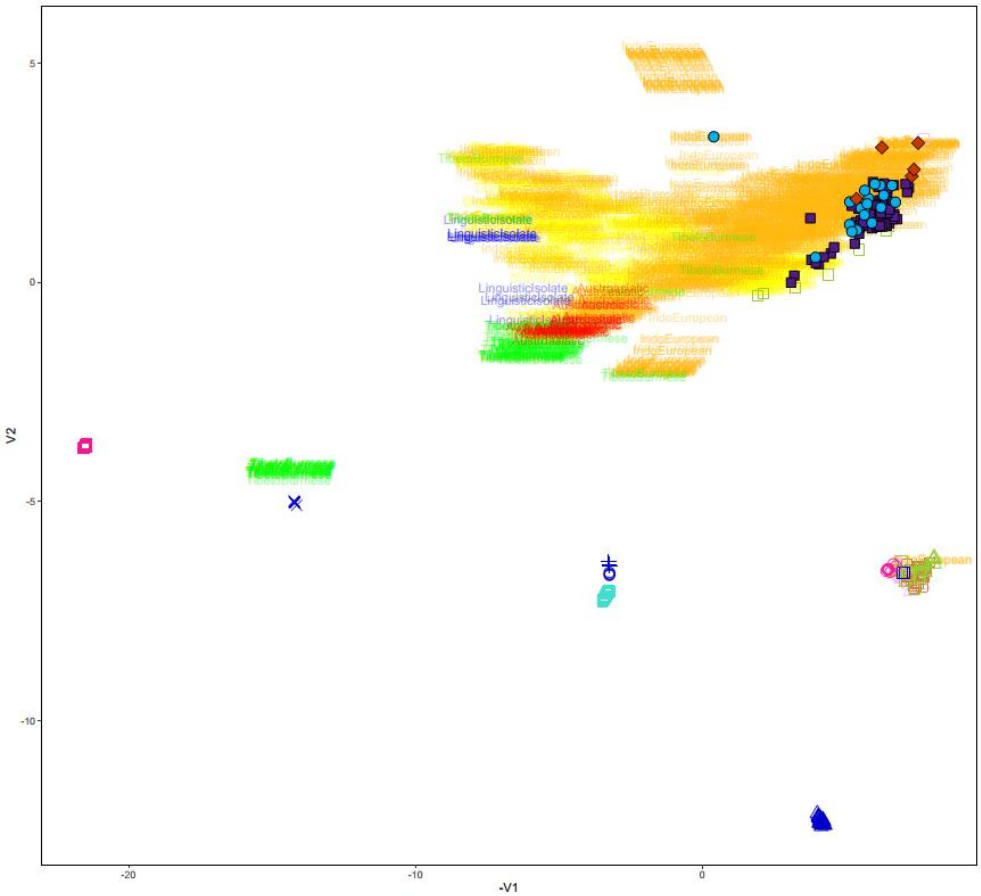

Figure 3 UMAP. UMAP showing present-day South Asian populations as text, coloured by language group, and ancient populations as points, following the legend of Figure 1.

UMAP [24] ver.0.2.3.1 was implemented in RStudio [28] ver.1.2.1335 using the 10 PC coordinates computed for the PCA as input and default parameters (n_neighbors = 15 and min_dist = 0.1). For this projection only the ancient samples and present-day South Asian samples were included. In UMAP it is important to remember distances between clusters do not represent genetic distance and orientation of the plot may differ from the PCA. Far right along axis V1 in the lower quadrant there is a tight cluster of samples from different time periods from Iran and Caucasus hunter gatherers whilst the other major clustering is similar to the PCA, with ancient Pakistan populations falling over present-day South Asian Dravidian and Indo-European groups and some Iran_BA samples. Again most of the Pakistan Medieval samples are more loosely clustered and slightly separated from the older samples.

Admixture

Admixture clustering was performed using ADMIXTURE [25] ver.1.3.0 on a panel of 3,356 present-day and ancient genomes comprised of present-day South Asians from [18], together with present-day and ancient Eurasians and South Asians from a number of other datasets as per referenced in [18]. Plink2 [29] was used to first filter the dataset for variants with a minor allele frequency below 0.01. Next, pairwise differences were calculated and the dataset pruned for linkage disequilibrium, leaving 261,249 SNPs. ADMXITURE was then run in an unsupervised mode from K4 to K15 with three replicates. CV errors were calculated and the best K chosen based on the lowest CV error that also provided the best resolution for differentiating ancestry in hunter-gatherer and Neolithic farmer populations.

Figure 4. Admixture components of present-day and ancient populations at K=8. Linguistic groups are labelled in boxes for the modern South Asian populations according to [18].

The admixture plot shows ancestry from Western European hunter gatherers maximised in blue, Anatolian Neolithic farmers maximised in light green, Iranian Neolithic early farmers in dark green, Eastern Siberian hunter gatherers in pink, and South Asians in yellow. The plot indicates the Pakistan samples are a mix of Neolithic Iranian and Anatolian, Steppe, western hunter gatherer and South Asian ancestries.

Outgroup f_3 statistics

Outgroup f_3 statistics was performed using the qp3pop package ver.435 included in ADMIXTOOLS [26]. For this analysis the Pakistan macro groups were tested against other ancient and present-day populations in the dataset with the African population Mbuti used as the outgroup. The top 30 populations with the highest f_3 values are plotted below. Present-day populations are coloured according to linguistic group, and ancient populations according to time period.

Consistent with the previous analyses, the results indicate shared drift is highest with Chalcolithic and Bronze Age populations that share Iranian and Steppe ancestry. The Bronze Age groups of Turkmenistan and Uzbekistan share Iranian farmer related ancestry, while shared drift with populations sharing steppe-related ancestry, represented by Russian MLBA groups, appear to increase in later periods.

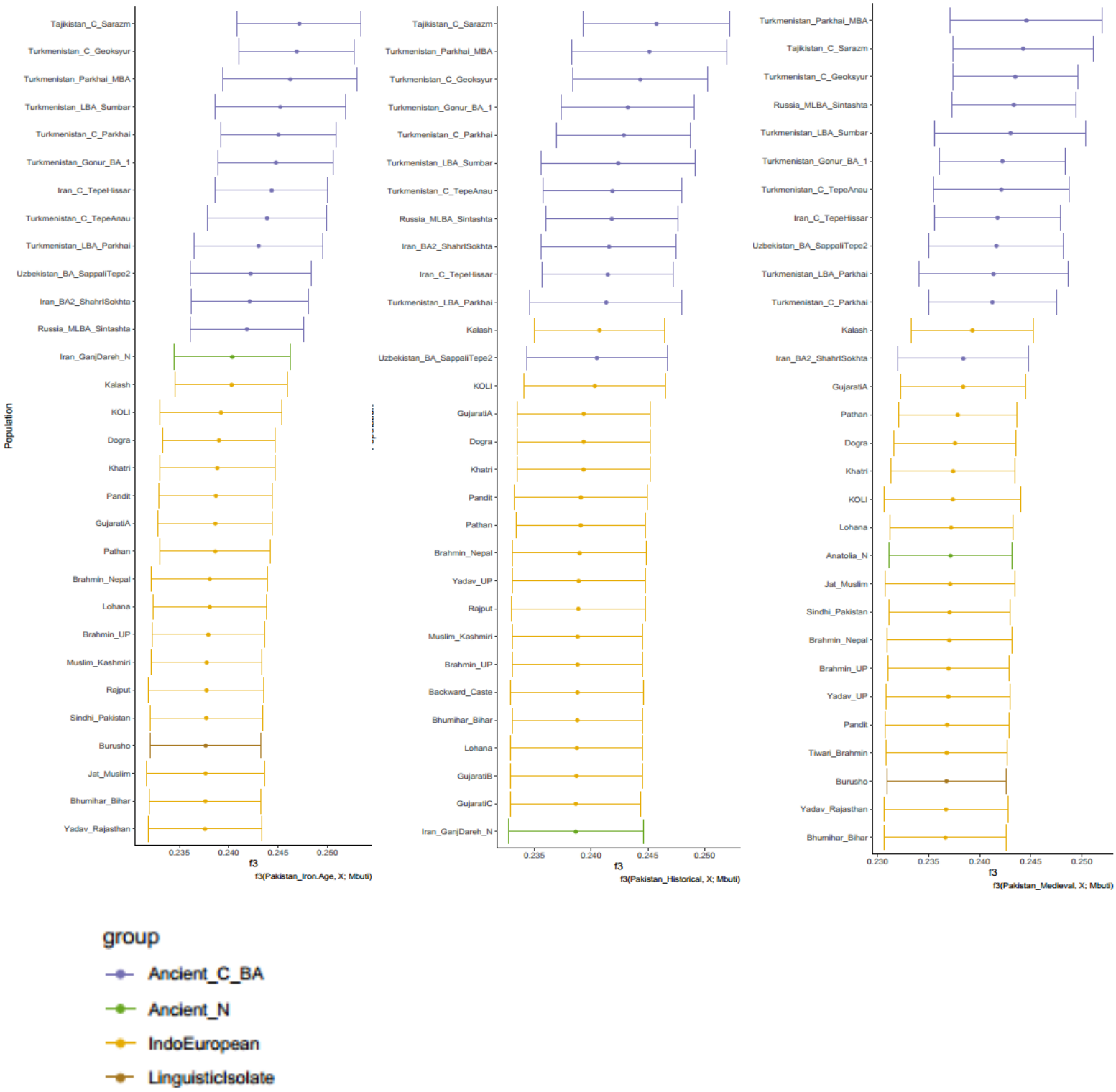

Figure 5. Outgroup f_3 statistics for Pakistan Iron Age, Historical and Medieval periods. The top thirty populations with the highest f_3 values are plotted. Present-day populations are coloured according to linguistic group, and ancient populations according to time period. Error bars are three standard errors.

f_4 statistics

f_4 statistics were performed using the qpDstat package included in ADMIXTOOLS [26]. The parameter f4mode was used to test cladality of pairs of populations in relation to another comparison population, with a minimum of 129,000 SNPs. Results for the test f_4(Mota, X; Central_ Steppe_MLBA, IranianEarlyFarmer_N) show that the Medieval group share more alleles with Bronze Age steppe pastoralists than the Iron Age populations (Z = -4.8 as opposed to Z = -2.4). The Medieval group also shows slightly higher affinity to South Asian hunter gatherers than the Iron Age group when paired with Iranian early farmers (Z = -7.2 vs Z = -9.7).

qpAdm

The software qpAdm was used in ADMIXTOOLS [26] to estimate ancestry proportions in the test populations Pakistan_IA, Pakistan_H and Pakistan_Medieval in a five-way distal model, which means five distantly-related source populations were used: Western Siberian hunter-gatherers, Western European hunter gatherers, Andamanese hunter-gatherers, Neolithic Iranian Early Farmers and Neolithic Anatolian Farmers. These were tested with a set of ten reference individuals or populations that includes Siberian, Caucasus and Near East hunter gatherers and early farmers, an Ethiopian hunter gatherer and upper Palaeolithic Europeans. Parameters details:YES and allsnps:YES were used. Before running qpAdm, qpWave was run to test the suitability of the reference groups chosen and confirms they are able to differentiate the source populations by testing the number of waves of admixture or migration from the right to the left populations.

A good fit for the data was obtained using five ancestral components that are distally related, comprising hunter-gatherer and farmer ancestry. In line with previous results, this shows that the later historical sites from Pakistan have more Andamanese hunter-gatherer ancestry than the Iron Age sites (20% in the Iron Age and 27% in the historical era (Figure 6). This points to more admixture with other populations harbouring such ancestry which can be found further southeast.

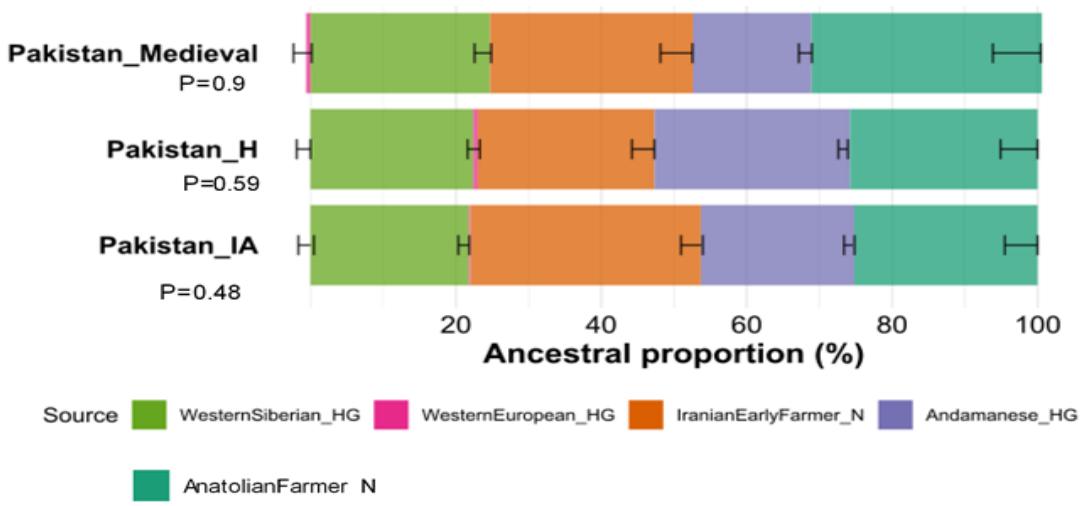

Figure 6 qpAdm stacked bar plots showing five-way distal admixture models for ancient Pakistan. Standard error bars shown.

Population genetic inferences in relation to other past and present-day populations.

The second work package attempts to evaluate and interpret the newly analysed datasets in terms of a broader demographic context, by building an admixture graph with Treemix [27] (see Materials and Methods).

Results

Treemix

Treemix [27] ver.1.13 was used to produce a graph from group-wise stratified allele frequencies after filtering on minor allele frequencies under 0.01 and pruning alleles not covered by any individuals, resulting in 415,900 SNPs and 14 groups. In addition to the Pakistan_IA group, the other groups used to construct the graph comprised Neanderthal and Denisovan genomes from about 800,000 to 400,000 years before present, Yamnaya (steppe pastoralist-related ancestry), Iran-farmer-related BMAC (Turkmenistan and Uzbekistan BA groups) and Indus Periphery Cline (largely Iran) genomes from about 5,000 to 3,000 years before present, and present-day genomes, which include an African outgroup Mbuti, together with Australasians, East Asians and South Asians from Papuan, Onge, Dai, Nicobarese, Juang, Rajput and Palliyar populations.

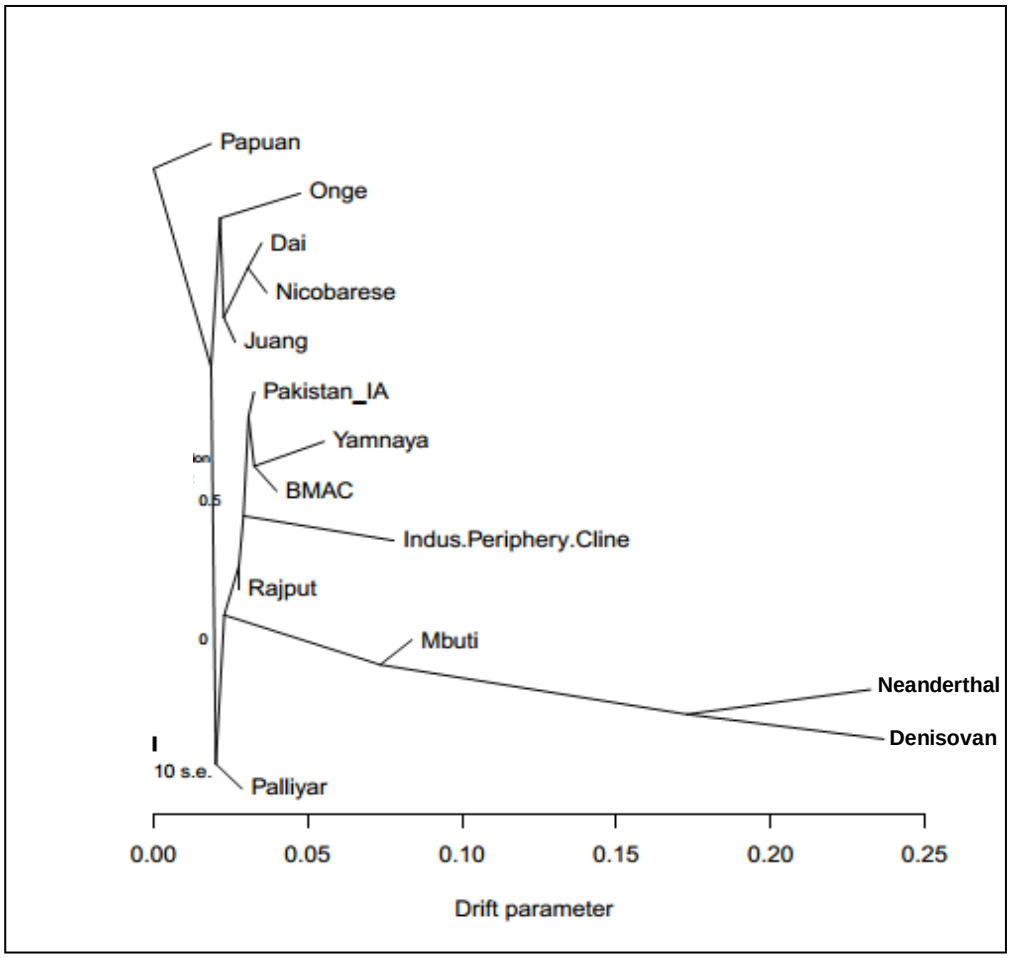

Figure 7. Admixture graph using Treemix

The admixture graph shows the close relationship of Pakistan Iron Age populations to steppe pastoralist and Iranian farmer-related ancestry (BMAC) as seen in the previous analyses, and displays the deep divergence of South Asians from other East Asians and Australasians populations.

Discussion and Conclusion

The results of the first work package provides strong statistical evidence to answer the question whether and to what extent migrations and admixture events led to population change and influenced genetic structure and variation among prehistoric South Asian populations. We were able to characterise the diversity seen in this region at population structure level from the Iron Age to Medieval times, and understand the processes of population admixture that led to the variation we see today. Building on this, the second work package furthers our understanding of past and present-day diversity. The myriad patterns of genetic, social, cultural and linguistic structures and

variation seen today in South Asia is tightly intertwined with past movements of people and their exchange of genes, events only encapsulated in ancient DNA molecules.

Important insights have been gained regarding the impact of Steppe pastoralist-related ancestry in South Asia, as we see that this ancestry is also found in ancient Pakistan by the Iron Age as well as in West Eurasia, where it has been more widely documented following steppe migrations westwards starting at the end of the Neolithic. The source of this ancestry in South Asia is likely from Central_Steppe_MLBA [18], which we have seen contributes to the ancestry of the ancient populations of Pakistan. Another significant source of ancestry is South Asian Andaman in origin. Indeed, present-day South Asians have been modelled as a mixture of Central_Steppe_MLBA with Iranian-farmer-related ancestry deriving from populations of the Indus Periphery Cline and Andamanese hunter gatherers in different proportions [18]. These sources of ancestry in modern-day South Asians have been, through dating of admixture events, confirmed to originate from the second millennium BCE [18]. To conclude, while this sheds new, important light on the past population genetic processes in South Asia, undoubtedly more samples will become available from a range of prehistoric time periods and regions of South Asia. When this happens, it will be possible to greatly refine and expand on these insights, and fill the gaps that remain in our knowledge about South Asian population history.

References

[1] Haak, Wolfgang, et al. "Massive migration from the steppe was a source for Indo-European languages in Europe." Nature 522.7555 (2015): 207.

[2] Mathieson, Iain, et al., “Genome-wide patterns of selection in 230 ancient Eurasians.” Nature 528.7583, pp. 499-503, 2015.

[3] Mathieson, Iain, et al., “The Genomic History of Southeastern Europe.” Nature 555.7695, 197, 2018.

[4] Fu, Qiaomei, et al., “Complete mitochondrial genomes reveal Neolithic expansion into Europe.” PLoS One 7.3, p. e32473, 2012.

[5] Lipson, Mark, et al. "Ancient genomes document multiple waves of migration in Southeast Asian prehistory." *Science* 361.6397 (2018): 92-95.

[6] Metspalu, Mait, Mayukh Mondal, and Gyaneshwer Chaubey. "The genetic makings of South Asia." Current opinion in genetics & development 53 (2018): 128-133.

[7] Basu, Analabha, Neeta Sarkar-Roy, and Partha P. Majumder. "Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure." Proceedings of the National Academy of Sciences 113.6 (2016): 1594-1599.

[8] Reich, David, et al. "Reconstructing Indian population history." Nature 461.7263 (2009): 489.

[9] Moorjani, Priya, et al. "Genetic evidence for recent population mixture in India." The American Journal of Human Genetics 93.3 (2013): 422-438.

[10] Narasimhan, Vagheesh M., et al. "The formation of human populations in South and Central Asia." Science 365.6457 (2019): eaat7487.

[11] Higuchi, Russell, et al. "DNA sequences from the quagga, an extinct member of the horse family." Nature 312.5991 (1984): 282-284.

[12] Pääbo, Svante. "Molecular genetic investigations of ancient human remains." Cold Spring Harbor symposia on quantitative biology. Vol. 51. Cold Spring Harbor Laboratory Press, 1986.

[13] Bouwman, A., et al. Code of Ethics—new principles for an ethical base of research on human remains. Institute of Evolutionary Medicine, University of Zurich, 2nd Edition (2014). Available at: http://evolutionarymedicine.ch/coe/

[14] Mays, Simon, et al. Science and the dead: a guideline for the destructive sampling of archaeological human remains for scientific analysis. English Heritage Publishing with the Advisory Panel on the Archaeology of Burials in England, 2013.

[15] Kreissl Lonfat, Bettina M., Ina Maria Kaufmann, and Frank Rühli. "A code of ethics for evidence-based research with ancient human remains." The Anatomical Record 298.6 (2015): 1175-1181.

[16] Green, Eleanor Joan, and Camilla F. Speller. "Novel substrates as sources of ancient DNA: prospects and hurdles." Genes 8.7 (2017): 180.

[17] Hofreiter, Michael, et al. "The future of ancient DNA: Technical advances and conceptual shifts." BioEssays 37.3 (2015): 284-293.

[18] Nakatsuka, Nathan, et al. "The promise of discovering population-specific disease-associated genes in South Asia." Nature genetics 49.9 (2017): 1403.

[19] Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.6

[20] Li, Heng, et al. "The sequence alignment/map format and SAMtools." Bioinformatics 25.16 (2009): 2078-2079.

[21] Fu, Qiaomei, et al. "A revised timescale for human evolution based on ancient mitochondrial genomes." Current biology23.7 (2013): 553-559.

[22] Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. "ANGSD: analysis of next generation sequencing data." BMC bioinformatics 15.1 (2014): 356.

[23] Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38.8 (2006): 904.

[24] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." arXiv preprint arXiv:1802.03426 (2018).

[25] Alexander, David H., John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals." Genome research 19.9 (2009): 1655-1664.

[26] Patterson, Nick, et al. "Ancient admixture in human history." Genetics 192.3 (2012): 1065-1093.

[27] Pickrell and Pritchard (2012). Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics.

[28] Team, R. Core. "R: A language and environment for statistical computing." (2013): 201.

[29] Chang, Christopher C., et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets." Gigascience 4.1 (2015): s13742-015. Available at www.cog-genomics.org/plink/2.0/.