Insights into the genomic histories of diverse human populations using whole-genome sequencing analysis.

Mohamed A. A. M. Almarri

Wellcome Sanger Institute

Darwin College

University of Cambridge

This thesis is submitted for the degree of Doctor of Philosophy

January 2021

Abstract

Despite the progress in sampling many populations, human genomics research is still not fully reflective of the diversity found globally. Understudied populations limit our knowledge of genetic variation and population history, and their inclusion is needed to ensure they benefit from future developments in genomic medicine. In this thesis, I describe extending our understanding of global genetic diversity and population history by two main projects. The first is focused on structural variation in a diverse set of 54 human populations which are part of the Human Genome Diversity Project (HGDP-CEPH) panel. Using whole-genome sequences previously produced at the Wellcome Sanger Institute, I generated a comprehensive catalogue of structural variation identifying a total of 126,018 variants, of which 78% are novel. Some reach high frequency and are private to continental groups or even individual populations, including regionally-restricted runaway duplications and putatively introgressed variants from archaic hominins. By *de novo* assembly of 25 genomes using linked-read sequencing, I discovered 1643 breakpoint-resolved unique insertions, in aggregate accounting for 1.9 Mb of sequence absent from the GRCh38 reference genome, highlighting the limitation of a single human reference genome. In the second project I collected and analysed a dataset of 137 high-coverage physically-phased genome sequences from eight Middle Eastern populations using linked-read sequencing. Focusing on the population history using single nucleotide variants, I found no genetic traces of archeologically documented early expansions out-of-Africa in present-day populations in the region. I show that Arabian populations have the lowest Neanderthal ancestry of all non-African populations tested, which is explained by them having elevated Basal Eurasian ancestry. By comparing Levantines and Arabian historical population sizes, I find a divergence that starts before the Neolithic era, when Levantines expanded while Arabians maintained small populations that could have derived ancestry from local epipaleolithic hunter-gatherers. All populations suffered a bottleneck overlapping the archaeologically-documented aridification events, with Arabians decreasing in size with the onset of the desert climate in Arabia ~6 kya while the Levantine bottleneck overlaps the 4.2 kiloyear aridification event. I also identify an ancestry that is associated with the spread of Semitic languages across the region during the Bronze Age. Finally, I identify novel variants that show evidence of selection, including signals of polygenic selection. This thesis fills an important gap in the study of diverse human populations, although further work is needed to sequence and characterize additional genetically underrepresented groups.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared at the beginning of each chapter. None of the contents in this thesis have been submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other university or similar institution. It does not exceed the prescribed word limit for the Biology Degree Committee.

Mohamed Almarri

January 2021

Acknowledgements

I would like to thank Chris Tyler-Smith and Yali Xue for all their guidance and support throughout my PhD, and for giving me the opportunity to work on exciting projects. During my time at the Sanger Institute, I did not just feel like a member of a research group, but part of a family. I also would like to thank my colleagues at the Human Evolution team for all their support and discussions: Anders Bergstrom, Elena Arciero, Michal Szpak, Javier Prado-Martinez, Pille Hallast, and Marc Haber. I would also like to thank Hilary Martin for her guidance and support over the past year, and also QinQin Huang, Emilie Wigdor, Kartik Chundru for all their helpful discussions. I thank Matthew Hurles and Richard Durbin for all their suggestions and feedback on my work.

I would like to thank Darwin College for making my stay in Cambridge so enjoyable, and for providing a community of friends that supported me over the last four years.

Last but not least, I would like to thank my parents, Ali and Jamila, for all their encouragement and support.

Dedication

I dedicate this work to the memory of my brother, Abdulla, who passed away towards the end of my PhD. The happiest moments I remember always include you.

Table of Contents

Chapter 1: Introduction

1.1 Thesis outline

I begin this thesis with an introductory chapter where I discuss principles in population genetics and methods used for the generation and analysis of genetic variation. I also briefly review current understanding of human population history and adaptation. During my work on the projects described in this thesis, the field of human genomics has been progressing at a rapid pace. I review at the start of each subsequent chapter the state of knowledge at the time of investigation, while in the results and discussion sections I cover recent relevant studies alongside my work. In the second chapter I describe my work on the analysis of a large set of human populations from the Human Genome Diversity Project. The third chapter describes my work on Middle Eastern populations. As the work presented here has been published in scientific journals or pre-prints (Bergstrom *et al.,* 2020; Almarri *et al.,* 2020a; Almarri *et al.,* 2020b), I will briefly cover the methods at the end of each chapter and refer to the publication for more detail. The final chapter discusses future directions.

1.2 Principles of population genetics

A genome of an organism contains the sum of genetic information encoded in deoxyribonucleic acid (DNA) base pairs (bp). At a particular position in a genome, differences, or genetic variants, can often be found when comparing individuals. The most abundant class of genetic variants are single nucleotide variants (SNV), substitutions at single sites. Insertions or deletions of sequences under 50 bp are referred to as indels, while structural variants encompass changes ≥ 50 bp, which include deletions, duplications, inversions, and insertions. A type of structural variant which varies in number of copies is known as a copy number variant (CNV). Tandem repeats are sequences in which a pattern of bases is repeated head-to-tail a variable number of times, and include satellites, minisatellites and microsatellites.

Genetic variants arise through mutations which have differing rates depending on the class of variation. One way of estimating the mutation rate is by directly analysing the number of new mutations, or *de novo* variants, between generations in a parent-child trio (Jónsson et. al.,

2017). Another less direct approach, sometimes called the phylogenetic rate, is performed by comparing genome sequence divergence between species, for example human and chimpanzee, with an estimated split time based on fossil evidence (Scally and Durbin, 2012). A third approach is by calibrating using accurately-dated fossils which have been sequenced to high accuracy, and comparing the number of mutations accumulated since a common ancestor with present-day samples: the 'missing' mutations provide an estimate of the mutation rate (Fu *et al.,* 2014). Autosomal SNVs have an estimated mutation rate of 0.4-0.6 x 10^{-9} per site per year. SNVs on the Y chromosomes and mitochondrial DNA have different rates, 0.7-0.9 x 10^{-9} and 1.8-3.2 x 10^{-8} per site per year respectively (Fu *et al.,* 2014). The mutation rate is not uniform across the genome, as it varies due multiple factors including sequence composition and nucleotide type. Most *de novo* mutations are inherited paternally, and the number increases as a function of the father's age (Jónsson et. al., 2017). This is consistent with the multiple rounds of mitosis during spermatogenesis after puberty, in contrast to ova which do not divide after birth.

Once a mutation arises, it will eventually either be fixed in a population, i.e. reach 100% frequency, or be lost. This future trajectory is affected by multiple factors, including natural selection and genetic drift. If a variant influences the fitness of the carrier, or the number of descendants, positively or negatively, it will be affected by selection. Positive selection acts to increase the frequency of a variant in a population, while negative, also called purifying selection, will decrease it. If the variant has no effect on fitness, a neutral variant, then its frequency will fluctuate randomly between generations and the probability of fixation is equal to its frequency. This is due to genetic drift, the random sampling of alleles from one generation to the next. It is related to the concept of an effective population size (N_e), which refers to the number of individuals contributing genetically to the next generation in a population. When the effective population size is small, genetic drift becomes stronger, while if a population size is large drift becomes weaker. Consequently, a neutral or even a weakly deleterious variant can potentially increase in frequency due to genetic drift. Thus understanding the demographic history of a population is important when attempting to pinpoint variants that have undergone positive selection. The effective population size of humans has likely been variable across time and populations, but has been estimated over the long-term to be around 10,000 individuals (Gronau *et al.,* 2011).

In a sexually reproducing organism, an individual contains many lineages inherited from its ancestors. This is due to the process of recombination, which shuffles the genetic sequences across homologous chromosomes during meiosis. An exception to this is the mitochondria and the non-recombining part of the Y chromosome, which are inherited with no recombination and only differ between generations due to mutations. Recombination does not occur randomly across the genome, with hotspots often associated with particular sequence motifs (Myers *et al.,* 2008). The distance between variants across a chromosome affects whether they will be inherited together or be separated by recombination, with closely physically-located variants more likely to be inherited together and passed on to the next generation. This is referred to as genetic linkage, and is a property of an individual chromosome. The resulting co-segregation of variants, more frequently than expected due to chance, i.e. not being in 'equilibrium', is referred to as linkage disequilibrium (LD) and is a property of a population.

Estimating the rate at which genetic variants accumulate allows us to understand and date when a variant arose. For example, assuming that a variant is neutral and in a stable large population, common variants are relatively old, while rare variants that are more region-specific are generally relatively young. Studying the pattern of shared mutations across populations allows us to date divergence times (Scally and Durbin, 2012). A generation time is needed to scale the estimates in years, and for humans it is estimated to be 26-30 years (Moorjani *et al.,* 2016). Genomes thus form a record of the forces that shaped them, and their investigation can illuminate the evolutionary history of our species.

1.3 Technology for the generation of genome sequences

The study of genetic variation and its association with phenotypic traits predates the discovery of DNA as the carrier of genetic information and the structure of the double helix. Through experimenting with pea plants, Gregor Mendel in the mid-19th century identified that certain traits follow specific inheritance patterns, which are now referred to as Mendel's laws. Subsequent studies analysed genetic variation indirectly by evaluating their effect on the biochemical properties of proteins. The first population-scale study of human genetic variation was reported in 1919, where frequencies of certain blood types were found to vary between populations (Hirschfeld and Hirschfeld 1919). The use of restriction enzymes in the 1970s subsequently allowed the direct study of genetic variation. These enzymes cleave specific sequences, which range mostly from 4-8 bases, by introducing double stranded breaks. This

creates fragments of different sizes which can be studied by gel electrophoresis. The presence of genetic variants within a population, also called polymorphisms, at such restriction sites will influence whether or not they can be cut by a restriction enzyme. The ability to detect and study variation using this method led to restriction fragment length polymorphism analysis. A subsequent improvement was the analysis of microsatellites, which due to their higher mutation rate are highly polymorphic, and the variable number of repeats can, after amplification by the Polymerase Chain Reaction (PCR), be separated by gel electrophoresis.

The genome sequencing era started in the 1970s when Fredrick Sanger and colleagues developed a method to directly read the sequence of DNA using dideoxy chain termination, now commonly called ‘Sanger sequencing’ (Sanger *et al.,* 1977). The method relies on modified deoxynucleotides which lack another hydroxyl group, dideoxynucleotides. Once incorporated into the expanding DNA molecule by DNA polymerase, the expansion will terminate. The four dideoxynucleotides are labelled with fluorescent dyes and after random termination the fragments of different sizes can be separated by gel electrophoresis and the sequence can be read. The method was subsequently automated, and coupled with the development of capillary electrophoresis increased throughput and the size of the molecule that can be read, or sequenced, to around 800 bp. This contiguous sequence of DNA that is read is referred to as a ‘read’ of this length. These developments allowed the first, relatively simple, genomes of viruses and bacteria to be sequenced. The assembly of large and more complicated eukaryotic genomes necessitated the analysis of longer and more repetitive sequences. This challenge led to overlapping sections of a genome being cloned separately and sequenced. These reads were then assembled into an accurate contiguous sequence, ‘contigs’, computationally. Although labour intensive and time consuming, this technique allowed the generation of assemblies such as the human genome. Genome assembly today no longer uses Sanger sequencing, but this ‘first-generation’ method remains widely used due to its accuracy in specific experiments, such as in targeted resequencing of a small region.

Before developments that lead to the widespread use of second-generation sequencing, DNA microarrays or ‘chips’ became, and still are, an important technology to investigate genome variation. Each array contains hundreds of thousands to millions of short probes which hybridize to a sequence that contains a known variant. The relative intensity of fluorescence caused by the hybridization of different alleles to the probe determines the genotype. Microarrays are a cost-effective way to analyse a large number of variants, and are widely applied to look for

associations between a trait and a variant through genome-wide association studies (GWAS). They are also a popular method used to study population history, but suffer from the requirement of needing to know a variant beforehand. This ascertainment issue is amplified due to the history of overrepresentation of European populations in genetic studies. Another limitation is that rare variants, which tend to be more population-specific, will likely not be included on the array.

Multiple different technologies were developed to usher in the era of ‘next-generation’ or second-generation sequencing. These methods were able to generate sequences at a substantially higher throughput relative to Sanger sequencing and at a lower cost, which continues to decrease. The dominant technology emerging from this period was developed by Solexa, and is commonly now referred to as Illumina sequencing. A ‘shotgun’ sequencing approach, this method relies on fragmenting DNA into random sequences which after library preparation are ‘sequenced-by-synthesis’. A camera captures the identity of a fluorescently-labelled reversibly-terminated nucleotide incorporated to a DNA molecule by a DNA polymerase. The high throughput is achieved by paralleling this step on billions of molecules across a flow cell. Initially the read length was only 35 bp (currently it can reach 250 bp), significantly shorter than Sanger sequencing, but reads have ~1% base error rate. Since the method uses no prior knowledge of the order of sequences, coupled with the relatively short length of the reads, these factors severely limit this technology in genome assembly. However, if a high-quality reference genome has been generated, this method offers a powerful way to study variation by comparing reads to the reference (‘mapping’) and identifying positions with differences. This approach can identify variants across the allele frequency spectrum, including ‘singletons’ or ‘doubletons’, variants that appear once or twice in the population studied, respectively.

Despite the wide application of Illumina sequencing today, the short length is an issue for the analysis of repetitive regions, which vary in length depending on the genome investigated. For example, if a read is 150 bp, but is composed of a long repetitive sequence that is found in multiple locations in the genome, this ambiguity means that it cannot be placed. The read needs to traverse the repeat in order to accurately place it in the genome on the basis of the non-repetitive portion. An improvement in Illumina technology is the use of pair-end sequencing, where both ends of a DNA fragment are sequenced. If one pair of the reads can be unambiguously mapped, this can assist in placing the second of the pair. However, this will still

not allow the investigation of large repeats such as segmental duplications, blocks of repetitive sequences over 1 thousand bases, or 1 kb, and have a sequence similarity of over 90%.
An additional challenge in sequencing genomes is ploidy, the number of sets of chromosomes. In diploid genomes, such as humans, a genetic variant can be heterozygous or homozygous. A pair of physically-nearby heterozygous variants can consequently be located on the same chromosome, or be on different chromosomes. This is known as the haplotype phase of the variant. If a read, or a read-pair, spans two heterozygous variants then the phase can be known, but this is limited by the length of the read and the size of the DNA fragment sequenced. Another factor is the heterozygosity of the sample investigated. Humans, on average, have 1 heterozygous variant every 1kb. A pair of 150 bp reads sequenced from both ends of a 500 bp fragment will thus provide limited phasing information. A technology recently developed retains long-range information from short-reads by creating pseudo-long 'linked-reads'. Such reads are generated by performing haplotype-level dilution of DNA fragments into over a million barcoded partitions, which then undergo standard short-read sequencing (Marks *et al.,* 2019). Linked-reads can, with high molecular weight DNA, create phased blocks of over several megabases. The accurate phasing simplifies genotyping, as each haplotype can only have one allele.

The limitations of short reads technology spurred advances that led to third-generation sequencing. Still in active development, the two dominant technologies emerging are single-molecule real-time (SMRT) sequencing from Pacific Biosciences and nanopore sequencing from Oxford Nanopore technologies. In SMRT sequencing, a DNA fragment and a DNA polymerase are immobilized at the bottom of an optical waveguide. The sequence is determined in real-time by detecting fluorescence from the binding of a labelled nucleotide in the active site of the polymerase. Alternatively, nanopore sequencing directly determines sequence by monitoring changes in electrical current using protein nanopores as nucleic acids pass through them. The reads generated by these technologies are typically over 10kb, and can reach over 1 million bases (Mb) in nanopore sequencing. Initial use of long-read technology required high input DNA, and additionally suffered very high error rates (>15%). A hybrid approach was often used where the error-prone long reads were corrected using accurate short reads. At the time of writing of this thesis, the latest developments have improved accuracy, especially in SMRT sequencing, which now can rival short-read accuracy; however, their costs render them impractical for large population studies.

Improvements in DNA extraction methods coupled with advances in the aforementioned sequencing technology have also allowed the study of genetic variation from ancient remains. After early PCR-based studies that were prone to problems with contamination, the ancient DNA field has matured over the past decade to provide unprecedented insights into human history. The generation and analysis of ancient DNA sequences create their own challenges. Since the time of death they start to degrade and fragment, and if successfully extracted, the resulting DNA molecules are usually under 50 bp. The finding that the petrous bone of the inner ear harbours the highest yield of DNA of all bone fragments has significantly improved the recovery of DNA, even allowing studies in relatively warm and humid climates that hinder preservation (Lipson *et al.,* 2020; McColl *et al.,* 2018). Ancient DNA molecules also accumulate damage that appears non-random. In particular, the deamination of cytosine to uracil, which is subsequently read as thymine, results in artefactual cytosine to thymine transitions. This occurs especially towards the ends of fragments. Restricting analysis to transversions or excluding the end few bases can be used to reduce this issue in analysis. Environmental contamination by microbial DNA and in addition by modern human DNA also complicates downstream analysis. If possible, increasing the number of reads sequenced can increase the absolute number with endogenous DNA up to a certain level for analysis. The high divergence of microbial to human DNA can exclude them from impacting analysis, but modern human contamination is a more serious issue due to the similarity of sequences. If contamination is found at a relatively high level, observing the proportion of reads with damage and limiting analysis to these can potentially remove the bias.

1.4 Methods for processing sequencing data

Choosing which way to process reads depends on the aims of the study, the presence of a high quality reference genome and the diversity of the organism investigated. Humans have relatively low diversity, and the current iteration of the reference genome, GRCh38, is the most accurate and complete vertebrate genome generated. For this reason, the most commonly used method for studying human variation is by mapping reads to a mostly linear haploid human reference. Regions of the reference that are accessible to reads and analysis are included within a ‘mappability mask’. The percent of the genome included in this mask depends on the read length of the technology and the stringency of mappability required for analysis. For short reads mapping to the human reference, the ‘strict mask’ encompasses only around 75% of all bases (Bergstrom *et al.,* 2020). Mapping reads has been aided by the development of efficient

methods, which can take into account some differences between the sequenced reads and the reference. Currently one of the most popular software for this process is BWA (Li and Durbin, 2009). After alignment, genotypes can be identified by evaluating the relative number of reads carrying a reference or alternative allele. For SNVs and indels, many software have been developed for this step, including GATK (McKenna *et al.,* 2010) and samtools (Li, 2011). An important metric impacting the accuracy of genotype calling is the sequencing coverage (or 'coverage'), the number of reads covering a position, on average, in the reference. Genotype calling needs to take into account the sequencing errors in reads, as well the noise in coverage at heterozygous sites. The target coverage depends on the aims of the study, with 'high-coverage' sequencing usually referring to around 30x, in which genotypes can be determined with high accuracy across all frequency bins within mappable regions. Early studies which were limited by sequencing costs sequenced a larger number of individuals from a population and leveraged the variation found within the entire dataset to call variants from lower coverage data and reduce expenses, but are more likely to miss rare variants.

In contrast to short forms of variation, different methods are used to identify structural variants, but they still mostly use comparisons with the reference genome. Initial and widely used approaches relied on indirect inferences, such as discordant read-pair mapping (Chen *et al.,* 2009); however, they tended to have high false positive rates and are limited to detecting a subset of SV classes. Improvements in subsequent methods, including ones that rely on coverage and split reads, resulted in substantially more sensitive and specific variant callers. The latest methods can also perform local assembly at putative identified structural variants, allowing accurate breakpoints to be determined (Kosugi *et al.,* 2019).

Despite the success and widespread use of the human reference genome in studying genetic variation, it does have limitations, such as introducing reference bias. Analysis is restricted to regions of the reference genome present and correctly assembled. Although GRCh38 is a high-quality reference, it is still incomplete and contains gaps and misassemblies which are continually updated through patches. Unresolved regions remain, such as ribosomal RNA gene arrays on acrocentric chromosome short arms, megabases of satellites in pericentromeric regions and large segmental duplications over hundreds of kilobases in size. Moreover, the current reference genome is a composite of a few individuals, the majority of which (~66%) is contributed by one person (Ho *et al.,* 2019). Sequences that are more divergent to the reference are less likely to be mapped due to having too many differences. This is especially a problem in

regions with high sequence diversity, such as the major histocompatibility complex. In addition, the reference can also harbour rare variants specific to the donor, including complicated sequence rearrangements.

Another limitation of the reference sequence is it being mostly a linear haploid genome, which cannot reflect the sequence diversity found in human populations. GRCh38 does contain some alternative haplotypes, regions of the genome with different versions of a sequence. However, the inclusion of such regions in a mostly linear reference brings its own challenges which has led many projects to exclude them from analysis. There is a large interest in using genome graphs to circumvent the issues of using a linear genome, and it is currently an area of active development. The use of graphs will increase in the number of reads aligned and allow better representation of the diversity of haplotypes in human populations. This is becoming increasingly important due to many population studies reporting tens of megabases of sequences that are not found in the reference (Sherman *et al.,* 2018). In addition to haplotypes with high divergence to the reference, these missing sequences could also be explained by population-specific sequences or by unresolved repetitive and complicated sequences which have not been placed in the reference. The lack of inclusion of such sequences in studies limits the scope of functional and association analyses, especially in diverse human populations not represented well by the reference.

An alternative way to process sequencing reads is by *de novo* assembly. This creates a consensus sequence from overlapping reads that does not make use of a reference genome or any prior information. Illumina reads can be used for this process, but they result in highly fragmented assemblies with missing sequences, especially for complex repetitive regions. In addition, they require relatively high coverage (~50x) and memory requirements. As a result, standard short-reads are not routinely used for *de novo* assembly of complex genomes, but can be of use for specific cases such as the investigation of structural variants (Li *et al.,* 2015). Linked-reads have the advantage of the accuracy of short-reads with the addition of long range information from the barcodes of the DNA molecule. Thus *de novo* assembly of linked-reads can create large contigs and even larger scaffolds, contigs separated by gaps of known length. Unsurprisingly, long-read technologies generate the highest quality and most complete assemblies. By comparing *de novo* assemblies to the reference, complicated structural rearrangements can be identified and included in population analysis. *De novo* assemblies can

even bypass the reference and the biases it introduces by comparing assemblies with each other.

1.5 Methods for analysing population histories using genetic data

A large number of methods have been developed to study population history using genetic data. They differ in which features of data are used, their complexity, the number of parameters set beforehand and what type of inferences are produced. Popular methods commonly used for initial exploration utilise information from allele frequencies at genotyped sites, and generally assume that similar allele frequencies are a result of shared ancestry. Principal component analysis (PCA) and model-based clustering are two of the most popular methods in this category. PCA is a dimensionality-reduction method which differentiates individuals along axes of variation, principle components (PCs), in an unsupervised manner (Patterson *et al.,* 2006). Different PCs can be investigated to explore the variation and structure of the dataset, and can also identify processing artefacts such as batch effects. An advantage of PCA is that it is relatively fast, and has essentially no parameters to set beforehand, resulting in it being reproducible. Recent developments in dimensionality-reduction techniques have led to more powerful clustering algorithms being used to explore genetic data. One is uniform manifold approximation and projection (UMAP), which is able to generate a two dimensional representation of the data which can be easily visualized (McInnes *et al.,* 2018). However, the interpretation of the clustering is more complicated than PCA, due to hyperparameters that need to be tuned by the user which affect the local and global clustering, as well as being a stochastic algorithm which will produce different results in every run. Model-based clustering implemented in algorithms such as STRUCTURE, require a pre-defined number of ancestral components (K) defined by the user to partition the variation observed in the dataset (Pritchard *et al.,* 2000). Usually, many different Ks are tested and all the results are taken to form a conclusion, although a certain K can form a better fit to the data. Extensions have allowed temporally-aware model-based clustering, which takes into account that different samples have lived at different times (Joseph and Pe'er, 2019). These methods work with variants that are approximately in linkage equilibrium, and as a result pruning of correlated variants is generally required beforehand.

A classic and important statistic to measure population differentiation is F_{ST}, or the fixation index, which measures the variation shared within and between populations. In addition to calculating average differentiation between populations, F_{ST} is also used to identify regions that show high

differentiation, relative to the average, between populations. An extension is the population branch statistic (PBS), which can identify regions that show differentiation in a specific population, potentially as a consequence of positive selection (Yi *et al.,* 2010). Another extension led to the development of *f-statistics*, a family of drift-based statistics that use allele frequency correlations (Patterson *et al.,* 2012). They have become powerful tools to explore hypotheses involving population history, and in their simplest form test whether populations are related to each other in a simple tree-like topology. They can be used to test whether a target population is descended from a specified set of ancestral populations and determine their relative contributions. They can also create models of population splits and admixtures that best represent the data. As these methods only use allele frequencies, they can be applied to any type of genetic data generated from different technologies, but increase in statistical power with increasing number of variants.

Alternative methods exploit the density of variants in genetic linkage through haplotype-based tests. They generally provide more power to investigate population structure in comparison to single marker tests. The Chromopainter/fineSTRUCTURE method produces a haplotype matrix of shared segments that can be used to statistically group samples into populations (Lawson *et al.,* 2012). Other methods, such as RFMix (Maples *et al.,* 2013), use haplotype information to perform local ancestry deconvolution of sequences in admixed samples by comparing to a set of reference populations. As these methods are haplotype-based, they require knowing the haplotype phase. One common way of phasing is to compare variants in a dataset with a large external set of haplotypes, a reference panel, where the most likely phase is determined. Alternatively, if a dataset is large, phasing can be determined through analysing variation found within the dataset. These two approaches are known as statistical phasing, and come with a rate of error which increases with decreasing frequency of the variant, as the number of times a variant is present within a panel is correlated with phase determination confidence. This can be exemplified by the extreme case of singletons within a dataset, which cannot be statistically phased. Experimental phasing of variants is an alternative approach to produce more accurate haplotype phasing, and can be performed using linked-read technology. As linked-reads retain the long range information from the original DNA molecule, variants across the allele frequency spectrum are accurately experimentally-phased.

The previous sets of methods analyse variable sites in a dataset which can be generated by genotyping arrays. The continued decreasing costs of sequencing have recently generated a

large number of whole genomes. These have led to the development of new methods that use information from both variable and non-variable sites in genome sequences to explicitly model their relationships through coalescence and recombination events. These models attempt to approximately reconstruct the ancestral recombination graph (ARG), a detailed description of the data that can be utilised to study population history. As human genomes are composed of multiple lineages that descend from many different common ancestors, this information can provide insights into ancestral population sizes. This is because the coalescence rate between haplotypes is dependent on the effective population size. Extending this to chromosomes from different populations, comparing the relative rate of within to between population coalescence can estimate population divergence, or split times. PSMC (Li and Durbin, 2011) and its later iteration MSMC (Schiffels and Durbin, 2014) have become widely used methods to study historical effective population sizes and separation history of humans and other species. The former is limited to the analysis of a single genome, providing estimates from ~20,000 years ago and older in humans, while the latter extended the analysis to multiple genomes (4 genomes or 8 haplotypes), increasing resolution in recent times to around 2,000-4,000 years ago. However, MSMC requires phased data, and poor phasing has been shown to lead to substantially distorted population histories (Terhorst *et al.,* 2017). The small number of genomes restricts the analysis to particular uses and time periods. A recently developed method, Relate, can create genome-wide genealogies that can scale to thousands of samples (Speidel *et al.,* 2019). The higher number of genomes handled by Relate increases the resolution of recent human population history up to and even within the last millennium, which is of interest as it overlaps with events in written history. This approximation of the ARG also allows a more powerful method to study the evolutionary history of a variant, in comparison to tests that only use summary statistics such as the allele frequency. Analysing local genealogies offers an approach to detect variants under selection, both strong and weak, as they will result in a burst of coalescent events over a short period of time. This can be extended by studying multiple variants simultaneously and investigating polygenic selection, where many variants each have a small effect on a trait.

1.6 A brief summary of current knowledge on human evolutionary history.

Anatomically modern humans are the only living species from the genus *Homo*, and together with their closest living relatives chimpanzees, bonobos, gorillas and orangutans are classified as hominids, or great apes. The availability of high-quality genomes from different hominids and

assumptions about the mutation rate have allowed their historical divergence times to be estimated, with humans splitting from chimpanzees around 6 million years ago (mya) and at an older time from gorillas and orangutans, ~9 and ~14 mya respectively (Scally and Durbin, 2012). The closest extinct groups to modern humans are Neanderthals and Denisovans, populations who are thought to have gone extinct ~40 kya. We know a lot more about the anatomy and lifestyle of Neanderthals in comparison to Denisovans, due to the large number of fossils identified from the former in Europe and the Middle East. The latter is more of a mystery, with only a finger bone, a few teeth and a mandible attributed to them in Eastern Eurasia (Chen *et al*., 2019). High-quality whole genomes from both groups have shed light into their historical relationships, and for Denisovans even identified their existence as they were previously unknown (Meyer *et al.,* 2012; Prufer *et al.,* 2014). Analyses of the sequences have shown that modern humans diverged from both archaic groups at around a similar time, 600 kya, while they diverged from each other more recently ~450 kya (Prufer *et al.,* 2014). These old split times suggests that all three groups inhabited and evolved at different locations with limited contact and gene flow for most of their existence. Denisovans also appear to carry sequences that may have been introduced through admixture with an early divergent hominin, potentially *Homo erectus* (Prufer *et al.,* 2014). The genomes also showed that both Neanderthals and Denisovans had relatively low genetic diversity resulting from long-term low effective population sizes, potentially impacting their health by increasing the burden of deleterious variants (Prufer *et al.,* 2014). In addition, a study using autosomal data suggested gene flow from an early diverged modern human population into Neanderthal groups (Kuhlwilm *et al*., 2016). Moreover, population relationships inferred by autosomal data do not mirror uniparentally-inherited chromosomes, as the mitochondria and Y-chromosome of Neanderthals appear closer to modern humans, instead of Denisovans, suggesting a possible replacement (Posth *et al.,* 2017; Petr *et al*., 2020).

1.6.1 Africa

There is a consensus that anatomically modern humans evolved in Africa. Early studies on mitochondrial DNA have shown that African populations have the highest genetic diversity of all human populations, non-Africans have a subset of this variation, and the phylogenetic tree is rooted in Africa (Cann *et al.,* 1987). This conclusion has been repeatedly confirmed using different classes of genetic variation in modern and ancient datasets. Additionally, a decrease in genetic diversity is found with increasing distance from Africa, consistent with a serial founder effect where a subset of the variation is carried in repeated expansions (Li *et al.,* 2008;

Jakobsson *et al.,* 2008). The oldest fossil assigned as modern human, although with some primitive features, dates to around 300 kya in Morocco, Northwest Africa (Richer *et al.,* 2017). Findings in Ethiopia, Eastern Africa, identified ~200 kya modern human remains (McDougall *et al.,* 2005). Using ancient autosomal genetic data, the earliest divergence between modern human populations has been estimated to be around 350-260 kya (Schlebusch *et al.,* 2017). In addition, ancient DNA from Africa have suggested complex diversifications of lineages within the continent, with a deeply divergent lineage contributing more ancestry to some West African populations (Skoglund *et al.,* 2017). Divergent uniparental chromosomes of present-day modern humans appear at different locations: The most divergent Y chromosomes have been identified in West and Central Africa (Mendez *et al.,* 2013), while the most basal mitochondrial sequences have been identified in Southern and South-eastern Africa (Chan *et al.,* 2019). Modern-day DNA has not been able to convincingly pinpoint the location where modern humans evolved within Africa; however, advances from multidisciplinary studies suggest that different human characteristics may have evolved at the same time across Africa, rather than in one single location (Scerri *et al.,* 2018). The degree of gene flow and population structure between these historical populations is not fully understood, as modern events have likely collapsed this ancient structure. However, an increasing number of studies are identifying divergent lineages present within modern-day Africans, suggesting that unsampled and unknown archaic populations have been present and their traces survive through admixture (Plagnol and Wall, 2006; Hammer *et al.,* 2011; Durvasula and Sankararaman, 2020). However, as no genome sequences of such archaic populations have been identified, which is not helped with the hot and humid climate present across much of the continent negatively affecting the preservation of DNA, these inferences of archaic populations are not widely-accepted. Through modern-day population comparisons, the deepest genetic splits within Africa appear between 100 and 200 kya (Mallick *et al.,* 2016). The population consistently appearing the most divergent in these comparisons is the Khoi-San speaking population of Southern Africa, who also appear to have had the highest historical effective population size of all studied human populations and the greatest genetic diversity. The next most divergent populations are rainforest hunter-gatherer populations of Central Africa, such as the Mbuti and Biaka.

An important recent event that reduced the older population structure in the continent is the complex expansion of Bantu-speaking agriculturists from Western Africa into Central, East and Southern Africa starting ~4 kya. In addition, East African populations have notable Eurasian ancestry believed to have been a result of Middle Eastern farmer movements and admixture

around 3 kya (Pagani *et al.,* 2012; Pickrell *et al.,* 2013; Skoglund *et al.,* 2017), which may be reflected in some of them today speaking Semitic languages, although Cushitic speakers have similar levels of Eurasian ancestry to Ethiosemitic-speaking groups (Pagani *et al.,* 2012). Northern African populations appear distinct from sub-Saharan groups and closer to Middle Eastern populations. Most of their ancestry is thought to come from back-migrations from the Middle East; however, it appears that Northern Africa and the Middle East have had historical and genetic connectivity over the past 15 kya (van de Loosdrecht *et al.,* 2018).

1.6.2 Out of Africa and archaic admixture

Most genetic studies suggest that human populations expanded out of Africa around 50-80 kya, a period in which a divergence between Africans and non-Africans is observed (Schiffels and Durbin, 2014). As genetics cannot identify the geographic location of the separation, this divergence pattern does not necessarily mean that it occurred through a population movement out of Africa. The separation could have started within Africa during a period of restricted gene flow between populations. A notable bottleneck is observed in all non-African populations during this period, resulting in much less diversity and higher levels of linkage disequilibrium in non-Africans in comparison to African groups (Li *et al.,* 2008; Jakobsson *et al.,* 2008; Li and Durbin, 2011; Schiffels and Durbin, 2014). The timing of this genetic expansion contrasts with the identification of modern human fossils outside Africa dating to over >80 kya years ago. A jaw bone identified in the Levant, a finger found in North Western Arabia and teeth located in Southern China have been dated to at least 177 kya, ~85 kya and more than 80 kya respectively (Hershkovitz *et al.,* 2018; Groucutt *et al.,* 2018; Liu *et al.,* 2015). In addition to fossils, tools and footprints attributed to modern humans have also been identified at similar old periods in Arabia around ~125 kya (Armitage *et al.,* 2011; Stewart *et al.,* 2020). Assuming the dates are correct, which in some cases are debated (Sharp and Paces, 2018), it is becoming clear that there were probably multiple human expansions out of Africa, predating the main event estimated by genetic studies. These two ideas can be reconciled by proposing that these early expansions became extinct or migrated back to Africa, and in either case did not leave a genetic trace in modern human populations. Nevertheless, there has been interest in identifying populations that may carry, even a small amount, of descent from an earlier expansion out of Africa, and indeed some reports have suggested so (Pagani *et al.,* 2016), although this is currently not widely-accepted. A complication for these types of analysis is that the availability of Neanderthal and Denisovan genome sequences have conclusively shown that modern humans encountered and admixed with archaic hominins in Eurasia during their expansion. Thus the

out-of-Africa event needs clarification; indeed if the ancestors of Neanderthals and Denisovans expanded out of Africa, they could be thought of an early expansion of humans that left some descent in modern humans due to admixture. The out-of-Africa event described in the rest of this thesis refers to the event 50-80 kya, described above. The overlap between the disappearance of Neanderthals and Denisovans around the same time as humans likely encountered them have been suggested to be related, either due to direct conflict or by indirect competition for the same resources. It could also be caused by modern humans being better at adapting to changes in climate, or having a lower burden of deleterious mutations due to their higher effective population size.

Admixture with archaic hominins has resulted in archaic sequences from Neanderthals and Denisovans surviving in modern human populations today (Reich *et al.,* 2010; Green *et al.,* 2010). It has been estimated that contemporary non-African populations have around 2% of Neanderthal ancestry, and the admixture event has been dated genetically, using information from the breakdown of haplotypes by recombination, to 50-60 kya (Fu *et al.,* 2014). The roughly similar amount of this ancestry in all these populations, and the low diversity and uniformity of introgressed segments, suggests that humans experienced one major pulse of admixture from Neanderthals (Bergstrom *et al.*, 2020). This also indicates that the admixture occurred soon after modern humans expanded out of Africa, potentially in the Middle East, and provides more evidence for a single major expansion out of Africa. Small differences in the amount of Neanderthal ancestry between contemporary populations have been identified, with East Asians found to harbour around 8-20% more ancestry in comparison to European populations (Chen *et al.,* 2020). This could be explained by East Asians having additional Neanderthal admixture events, weaker purifying selection on introgressed variants due to East Asians having a smaller effective population size, and/or by Europeans having descent from a population that did not participate in the admixture event with Neanderthals. This hypothetical population, called Basal Eurasians, is proposed to have diverged from other non-African populations before their radiation (Lazaridis *et al.,* 2014; 2016). The existence of such a population raises questions on why they seem to lack Neanderthal ancestry: could have they resided in Africa while other populations met Neanderthals in the Middle East? When did they come into contact with ancestral populations contributing the remaining sources to Europeans? Ancient DNA from a Basal Eurasian sample is needed to illuminate their history.

Denisovans admixture history shows differences and similarities in comparison to Neanderthal. The most apparent contrast is its regional stratification: individuals with Australasian and Melanesian ancestry harbour the highest ancestry (2-5%), whereas East Asian and American populations show a very small, but detectable, amount of about 0.1% (Reich *et al.,* 2010; Reich *et al.,* 2011; Meyer *et al.,* 2012). Denisovan segments also show longer lengths in comparison to Neanderthal, suggesting a more recent admixture time (Sankararaman *et al.,* 2016). The sequenced Denisovan genome shows some divergence from the Denisovan segments found in Melanesians, and lineages found in East Asians and Melanesians come from different Denisovan sources, suggesting at least two different pulses of admixture (Browning *et al.,* 2018). Moreover, Melanesian populations themselves have been reported to harbour two independent Denisovan lineages, separate from the East Asian lineage (Jacobs *et al.,* 2019). A study has suggested an almost linear decrease in Neanderthal ancestry since the time of admixture, potentially due to continuous selection against introgressed segments (Fu *et al.,* 2016). However, recent analysis have shown that overall levels of Neandertal ancestry did not decrease significantly in the past 45ky; with strong negative selection on Neanderthal segments likely appearing in the first few hundred generation after the introgression event (Petr *et al.,* 2019). Both Neanderthal and Denisovan segments appear depleted around functional sequences such as genes, suggesting they were in general deleterious and removed from the population through purifying selection (Sankararaman *et al.,* 2014). This appears especially pronounced on the X-chromosome and in genes expressed in the testes, suggesting that Neanderthal ancestry reduced fertility in males when in a modern human genetic background (Sankararaman *et al.,* 2014). The introgression event may occasionally have introduced variants into modern humans that were positively selected, a process called adaptive introgression (Gittelman *et al.,* 2016). A striking example has been reported in the high-altitude living Tibetans, where a haplotype thought to have introgressed from Denisovans appears at high frequency (Huerta-Sanchez *et al.,* 2014). This haplotype overlaps *EPAS1*, a gene which encodes a hypoxia-inducible factor and is believed to have helped Tibetans to survive in the low-oxygen environment of the Tibetan plateau. Thus admixture with divergent archaic hominins has had benefits and disadvantages to modern humans, providing access to a gene pool with neutral, beneficial or detrimental variants.

1.6.3 Europe

Once out of Africa, modern humans quickly populated much of Eurasia and Australasia. The region with the best understood genetic history is Europe, due to the large amounts of studies

sampling modern and ancient DNA which have documented population turnovers and migrations. Europe is thought to have been settled by modern humans around 45 kya (Benazzi *et al.,* 2011; Higham *et al.,* 2011). These early populations appear to have not contributed to contemporary Europeans, as a discontinuity in ancestry appears before the Last Glacial Maximum (LGM), potentially as a consequence of climate change. Later populations, living from between around 37 to 14 kya, descend from a founder population which did contribute to present-day Europeans (Fu *et al.,* 2016). Major population transformations took place in the Holocene, after the LGM, when agriculture developed in the Middle East during the Neolithic ~10 kya. This 'Neolithic transition' began as hunter-gatherers became sedentary and transitioned to lifestyles involving agriculture and animal husbandry. Farmers from Anatolia expanded into nearby parts of Europe ~8 kya resulting in admixture with, and sometimes the replacement of, local hunter gatherers. This change in lifestyle supported larger populations, documented genetically by increases in effective population sizes during and after this period (Schiffels and Durbin, 2014). During the Bronze Age ~5 kya, another major event that changed the genetic landscape of the region is the movement of pastoralists from the Eastern Eurasian Pontic-Caspian steppe into Europe, replacing up to half of the ancestry in some regions (Allentoft *et al.,* 2015; Haak *et al.,* 2015). These herders, attributed to the Yamnaya culture, were themselves descended from various hunter-gatherers from the Caucasus and Russia (Jones *et al.,* 2015). The change in ancestry of the region is thought to be associated with technological innovations such as horseback riding assisting conquests, and may have spread Indo-European languages. Today, Europeans thus form a mixture of these three divergent ancestries, with relative differences in proportions in different regions. North Eastern Europeans have high steppe-related ancestry, which appears low in Sardinians who instead have high Neolithic farmer ancestry (Lazaridis *et al.,* 2014). These population transformations greatly reduced the genetic structure within prehistoric Europe, and modern-day European genetic variation shows strong correlation with geography (Novembre *et al.,* 2008).

1.6.4 Asia

The initial peopling of Asia is not well understood, but the analysis of early ancient genomes from the region have provided some insights. An individual from who lived in Western Siberia (Ust'-Ishim) 45 kya displays similar genetic affinity to modern-day East Asians, Western Eurasians and Aboriginal Australians, after accounting for archaic admixture (Fu *et al.,* 2014). This implies that the sample comes from a population that lived before, or around, the separation of these populations. Another important sample that comes from Southern Siberia

dating to 24 kya (Mal'ta), shows genetic affinity to both Western Eurasians and Native Americans, and less affinity to Siberians and East Asians (Raghavan *et al.,* 2014). A ~36 kya genome from European Russia (Kostenki14), appears closer genetically to Western Europeans than to East Asians, suggesting that these two populations have diverged by that period (Seguin-Orlando *et al.,* 2014). Two subsequent expansions into Asia transformed the genetic landscape of the region and mixed with and replaced the local Mal'ta-like hunter-gatherers. The first, was the movement of steppe pastoralists ~5 kya, around the same time they also expanded into Western Europe. The second was replacement of these steppe pastoralists in Central Asia ~3 kya by a population affiliated with the Sintasha culture who inhabited the Northern Eurasian steppe and subsequently admixed with East Asians (Allentoft *et al.,* 2015).

In the Middle East, agriculture appears to have been independently developed by different populations in the Fertile Crescent, as continuity between hunter-gatherers and early farmers has been found in both Iran and the Levant (Lazaridis *et al.,* 2016). This suggests that the movement of ideas and farming technology spread faster than the movement of people. These groups were strongly differentiated, but subsequently these farmer populations mixed with each other before expanding into different regions. Anatolian farmers moved westward into Europe, while Levantine farmers spread southwards into East Africa. Iranian farmers moved northwards and admixed with Eurasian steppe populations, and also moved eastwards into South Asia. Most modern-day south Asians can be modelled as a mixture between two historically divergent populations: Ancestral North Indians (ANI) and Ancestral South Indians (ASI) (Reich *et al.,* 2009). The latter is related to, although distantly, modern-day indigenous Andaman Islanders while the former is descended from a West Eurasian source: Iranian-ancestry and Steppe-ancestry. Iranian-related ancestry first entered the region over 4 kya, followed by steppe pastoralists during the Bronze Age. Within the past 3ky, populations with varying proportions of ANI and ASI mixed with each other to create the modern Indian cline observed today (Narasimhan *et al.,* 2019).

1.6.5 Oceania

Evidence from archaeology shows that Oceania was populated ~50 kya, indicating that modern humans quickly reached the region after expanding out of Africa. However, as with other regions, earlier occupations have been reported (~65 kya; Clarkson *et al.,* 2017). Archaeological and linguistic data previously suggested likely multiple founding events of the region; however, recent genetic studies of modern-day Aboriginal Australians and Papuans indicate a single

founding event (Malaspinas *et al.,* 2016). The two populations then diverged from each other ~35 kya, long before the rise in sea levels which separated New Guinea and Australia after the end of the last glacial period around 12 kya. Despite the relatively short geographical distance between them, this separation time is similar to when Europeans and East Asian split from each other. Within Australia, relatively old split times, 20-30 kya, are identified when comparing South Western and North Eastern populations (Malaspinas *et al.,* 2016). This surprising long-term isolation and lack of gene flow has been proposed to be due to changes in environment leading to the desertification of Australia, which today is the driest continent. The European colonization of Australia in the late 18th century introduced substantial European gene flow to Aboriginal Australian groups, which is not apparent in New Guinea, who instead show admixture from populations of Southeast Asian ancestry, which Australians lack. This ancestry appears mostly confined to coastal regions of New Guinea, as the highlands do not appear to harbour Southeast Asian ancestry. Within New Guinea, old separation times are observed between highlands and lowlands, ~15 kya (Bergstrom *et al.,* 2017). In addition, high levels of population structure are observed on the island, with F_{ST} pairwise comparisons over 5% (higher than British and Sri Lankan populations) being common, reflecting its cultural and linguistic diversity. This is despite the independent development of agriculture in New Guinea. While agricultural transitions homogenised populations in Africa, Europe and Asia, strong genetic structure still persists on the island today.

1.6.6 Americas

The Americas are thought to have been populated 15-20 kya (Jenkins *et al.,* 2012), as humans moved through the land bridge in Beringia that connected Eurasia and North America, although a recent study suggested a much earlier presence around 31 kya (Ardelean *et al.,* 2020). Whole genome sequences of Native Americans and Siberians have estimated that they have diverged up to ~23 kya (Raghavan *et al.,* 2015), indicating there may have been a period of isolation in Beringia for a few thousand years before moving into the Americas. Until around 13 kya, a large ice sheet covered most of North America and is thought to have made movements southwards across the land difficult (Goebel *et al.,* 2008). An ice-free corridor subsequently opened which could have facilitated migrations; however, archaeological sites south of the corridor have been uncovered dating before its opening, suggesting that modern humans reached such locations before the ice melted (Dillehay *et al.,* 2015). An alternative path that has been proposed is via a coastal route along the Pacific West of North America, avoiding the ice sheets (Pedersen *et al.*, 2016). The expansion quickly reached the south of South America, with southern Chile showing

signs of human occupation ~15 kya (Dillehay *et al.,* 2015). The ancestry of the founding population appears to be a mixture of one third related to Mal'ta hunter-gatherers and two thirds East Asian (Raghavan *et al.,* 2014). The former population also contributed to modern-day Europeans, resulting in Native Americans being closer genetically to Europeans than to East Asians. The modern-day Inuit populations living in the American Arctic originate from a separate and more recent migration event from Siberia ~3 kya. However, the first population to inhabit the American Arctic, affiliated with the Paleo-Eskimo culture, were a group that moved independently from the previous migrations ~5 kya, and survived for 4 thousand years before going extinct and being replaced by Inuits 700 years ago (Raghavan *et al.,* 2014). The bottleneck associated with the movement into the Americas is reflected genetically, with the populations showing the continental lowest diversity (Li *et al.,* 2008). The European colonization of the continent in the last few centuries resulted in significant admixture from African and Europeans sources, transforming the population genetic landscape and structure (Moreno-Estrada *et al.,* 2014; Homburger *et al.,* 2015).

1.7 The effects of culture and lifestyle on adaptation

The expansion of humans, within and outside Africa, resulted in encountering new environments that brought with them new selective pressures. Innovations associated with cultural and lifestyle changes such as plant and animal domestication also introduced environmental changes related to diet and pathogen exposure. The availability of whole genomes from different human populations allows the identification of regions that are potentially linked with adaptation, and coupled with functional studies that elucidate the function of these variants, can illuminate the effect of natural selection on human populations.

Out of Africa, humans had to adapt to an environment with less sunlight, especially at higher latitudes. Around the equator, it is thought that human populations with darker skin colour, due to higher melanin, have better protection from skin damage and folate degradation by ultraviolet radiation (Jablonski and Chaplin, 2017). However, ultraviolet exposure is important for the production of vitamin D, which has wider importance for diverse cellular processes linked with human health such as skeletal development. Populations at higher latitudes will have less exposure to ultraviolet radiation, and it has been proposed that lighter skin colour is advantageous for vitamin D production; alternatively, light skin colour may have been considered sexually attractive (Darwin, 1871). Variants within several genes implicated in

pigmentation have been shown to be under positive selection at higher latitudes including *SLC24A5* and *SLC45A2* (Lamason *et al.,* 2005; Wilde *et al.,* 2014); although the former has also been suggested to have increased to high frequency due to migration (Mathieson *et al.,* 2015). Transition in lifestyles also brought about changes in diet. Selection for lactase persistence appears as one of the strongest signals of recent positive selection in some populations (Mathieson *et al.,* 2015). Multiple different variants that increase the expression of *LCT* in adults have been identified in different farming populations in Europe, the Middle East and Africa (Enattah *et al.,* 2002; Tishkoff *et al.,* 2007; Enattah *et al.,* 2008). In addition, *AMY1* which produces the amylase protein found in saliva that breaks down starch into sugars, has been found at variable copy number in human populations (Perry *et al.,* 2007). Higher copies are associated with higher expression of the protein, and populations that consume higher amounts of starch, such as the Japanese and Hadza, have been found to have, on average, higher copies than groups with low-starch diets (Perry *et al.,* 2007). An important selective agent for new environments is pathogen exposure. A classic example is malaria infection. Heterozygotes for the p.Glu6Val variant in the β-globin encoding gene HBB have some protection from malaria, but homozygous individuals suffer from sickle cell anaemia (Kwiatkowski, 2005). Another variant that protects against malaria results in the Duffy-null phenotype, which prevents the expression of a red blood cell receptor required for Plasmodium vivax (Miller et al., 1976; Kwiatkowski, 2005).

Selection can act immediately on *de novo* mutations, or on variants that are already present and polymorphic in the population but only later become advantageous: standing variation (Barrett and Schluter, 2008). Additionally, as mentioned previously, introgression can also introduce new variants that can be selected, either straight away, or later when a selective pressure arises. The long term habitation of archaic populations in Eurasia likely led to adaptations to their environment. Thus through introgression, modern humans could rapidly adapt to their new environments by incorporating beneficial variants present in archaic hominins. This appears particularly important for immunity, as humans are likely to have encountered pathogens that they were not previously exposed to in Africa during their expansion (Abi-Rached *et al.,* 2011).

It is also notable that natural selection has responded to selective pressures that were introduced by human actions. The transition to a more sedentary lifestyle after the invention of agriculture both settled populations and allowed them to grow in size into the major cities we see today. Such a much larger and denser population increases the pathogen load, and the

domestication of animals exposes infectious agents that can cross into humans (Fan *et al.,* 2016). Thus cultural and genetic evolution are closely associated; as we engineer the environment, the changing conditions will create new selective pressures (Nielsen *et al.*, 2017). It is also apparent that many variants that show high stratification in human populations, possible due to local adaptation, are linked to visible phenotypic traits such as hair, skin and eye colour. However, these differences in pigmentation-associated genes are outliers to the average level of differentiation across the genome. Consequently, human populations are much genetically closer to each other than might be predicted based on just observable traits (Nielsen *et al.*, 2017).

It is also becoming apparent that single variants with a very strong effect on a trait, such lactase persistence or hair colour, are rare. Complex traits are affected by a large number of variants across the genome, each with a relatively small contribution (Pritchard *et al.,* 2010). Consequently, selection on a complex trait would require a shift in allele frequencies across multiple variants, but the magnitude of such change at each locus would be small (Pritchard *et al.,* 2010). This complicates the identification of such polygenic selection, and methods for their detection are an active area of research (Stern *et al.,* 2021). There have been reports of polygenic adaptation, most notably for height (Turchin *et al.,* 2012; Tucci *et al.,* 2018). However, recent studies have shown that the analysis of polygenic traits has been confounded by uncorrected population structure (Soheil *et al.,* 2019; Berg *et al.,* 2019). The identification of variants associated with a polygenic trait or a disease requires large sample sizes in order to have adequate statistical power to identify variants with weak effects. This led to studies that rely on meta-analyses, combining summary statistics from multiple studies of different populations (Turchin *et al.,* 2012). However, these populations may have subtle differences in allele frequencies, and introduce systemic biases when analysing hundreds of thousands of variants across the genome. As we are now in the biobank era, where countries have generated large datasets of their populations documenting hundreds of traits along with genotypes, the issues of population structure affecting meta-analyses are likely to reduce, but not completely go away (Zaidi and Mathieson; 2020). Such large datasets are important to understand complex traits and polygenic adaptation.

Chapter 2: Analysis of Global Human Structural Variation

This chapter has been published in two papers: Bergström *et al.,* 2020 and Almarri *et al.,* 2020a. A small portion of the work presented here was included in the former paper which mostly covered SNVs and indels, while most of the analysis in this chapter is presented in the latter study, which was restricted to structural variation. I performed all the analysis presented in this chapter, except for the fluorescent *in situ* hybridization analysis which was performed by the molecular cytogenetics team at the Sanger Institute. DNA library preparation and sequencing was performed by the Wellcome Sanger Institute sequencing facility.

2.1 Introduction

After the initial sequencing of the human genome, efforts began to identify and catalogue genetic variants found in human populations. One of the important early projects was the International HapMap Project (International HapMap Consortium; 2005), which studied initially 269 samples from 3 continental populations: Africans (Yoruba from Nigeria), Europeans (Utah residents of European Ancestry) and East Asians (Japanese from Tokyo + Chinese from Beijing). Subsequently, a major development in the population genomics era began with the pilot phase of the 1000 Genomes Project (1000GP), which analysed four different populations using low-coverage sequencing (1000 Genomes Project Consortium, 2010). The final Phase 3 release produced a global reference of variation by analysing 2,504 individuals from 26 populations (1000 Genomes Project Consortium, 2015). This resource identified ~88 million variants in total: 84.7 million SNVs, 3.6 million indels, and 60,000 structural variants. The goal of the 1000GP was to identify common genetic variation (minor allele frequency (MAF) > 1%); and they identified >99% of such variants with mostly low-coverage sequencing (~7x coverage). An additional advantage of the 1000GP is that both the samples as cell lines and the dataset are open access and individual genotypes, in addition to allele frequencies, are available with essentially no restrictions.

The continued decreasing costs of DNA sequencing has provided genetic data aggregated from hundreds of thousands of individuals, particularly sampled from medical genetics studies. This

has established resources such as the Haplotype Reference Consortium (McCarthy *et al.,* 2016) and Exome Aggregation Consortium (Lek *et al.,* 2016). The consent for these studies do not allow open access to genotypes which restricts their utility; however, they are still valuable to estimate allele frequencies and the creation of large reference panels for imputation. Another limitation of these resources is individuals of European ancestry comprise the majority of such samples. Recent studies have sampled individuals from more diverse populations and sequenced them to high-coverage. The Simons Genome Diversity Project (SGDP) analysed 300 samples from 142 populations (Mallick *et al.,* 2016), while the Estonian Biocentre Human Genome Diversity Panel (EGDP) sequenced 483 individuals from 148 populations (Pagani *et al.,* 2016). These studies were focused on population history, and have extended our understanding of human diversity, archaic admixture, and temporal changes in population size. Both projects, however, have limitations: the EGDP contains very few African samples, while the SGDP contains mostly 2 samples per population, hindering detailed analysis within each population.

This over-representation of European populations in genetic and genomic studies limits our understanding of genetic variation and population history; and has the potential to exacerbate healthcare inequalities as the field moves towards precision medicine (Siriguo *et al.,* 2019). The “Centre d’Etude du Polymorphism Humain” (CEPH) foundation and the Human Genome Diversity Project (HGDP) have collaborated to establish a cell-line collection from diverse human populations (Cann *et al.,* 2002). Throughout this thesis I refer to this dataset as the HGDP. This repository has been used extensively to study human diversity and evolution using different classes of genetic variation, including microsatellites (Rosenberg *et al.,* 2002) and SNV arrays (Li *et al.,* 2008; Jakobsson *et al.,* 2008). The collection has many advantages for the study of human population genetics. Extracted DNA samples and the data generated from them are available open-access with essentially no restrictions, and the resource contains a potentially unlimited amount of DNA. It contains samples from 54 diverse populations (Figure 2.1), with an average of 17 samples per population, although this ranges from 6 (San) to 51 (Palestinians). Previous projects such as 1000GP sampled metropolitan populations from major continental groups, with criteria set such as relevance to medical studies and being from non-vulnerable populations. The HGDP panel constitutes a much wider cross-section of human diversity, which are of linguistic, anthropological and historical relevance. It contains populations that practice different lifestyles: hunter-gathers, agriculturalists, and nomadic pastoralists. Within Africa, it contains the Khoe-San population, shown to be the most divergent living human

population; in addition to Mbuti and Biaka, rainforest forager populations from Central Africa. In the Middle East it contains desert-inhabiting Bedouin nomads and the Druze ethno-religious group. The panel comprises a large number of groups from East Asia, including the Yakut living in North-eastern Siberia and several ethnic groups from China, including the Tungusic-speaking Oroqen from Northern China and Turkic-speaking Uyghurs from Western China. Importantly, it contains samples from Oceania which are severely underrepresented in genetic studies, with two populations from Papua New Guinea and one from Bougainville. While in the Americas it contains Native American samples from groups that do not have recent European admixture, such as the Karitiana. In Europe it contains the isolated Orkney islanders, as well as the Basque population who do not speak an Indo-European language which is common in the continent. Although comprising a diverse set of populations; it should be noted that several regions are not sampled in the HGDP: India, Polynesia, Australia and Arabia, while others such as Europe and the Americas have limited representation.

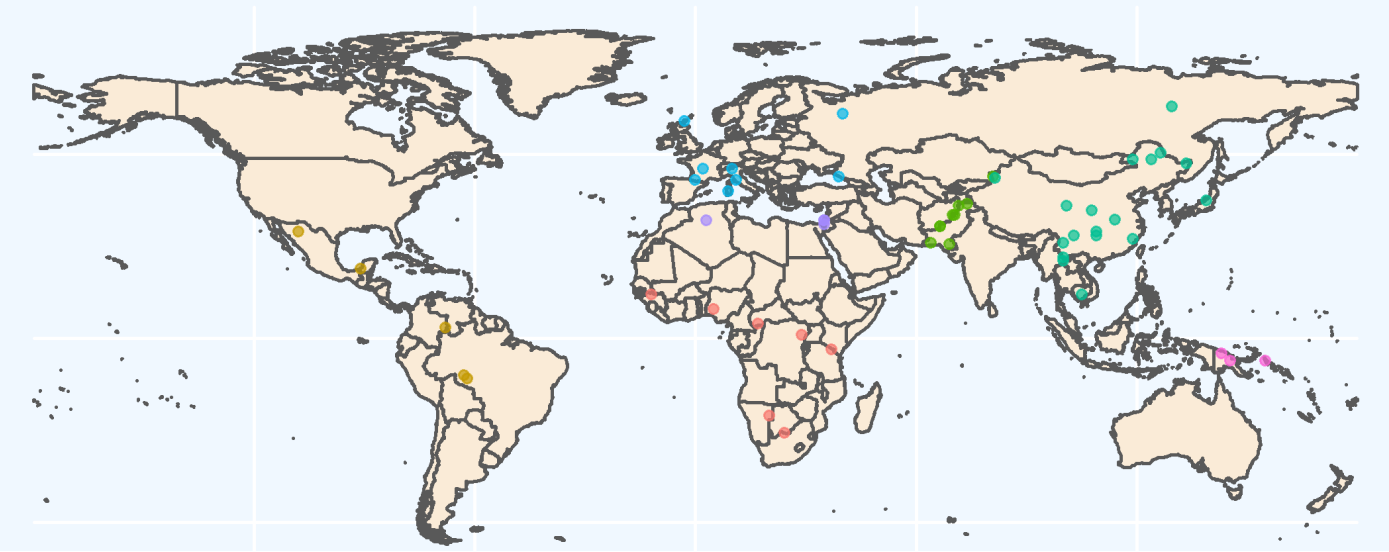

Figure 2.1: **The HGDP dataset**. Each point represents a population with colours illustrating regional groups.

The majority of whole-genome sequencing population studies have concentrated on SNVs, and have excluded structural variants (SVs). A typical human genome differs from the reference at around 4-5 million sites, with over 99.9% of variants being SNVs and short indels (1000 Genomes Project Consortium, 2015). SVs, although smaller in absolute number, contribute a greater diversity at the nucleotide level than any other class of variation as they affect more bases. Consequently, SVs are important in disease susceptibility and genome evolution. SVs

arise through different mutational processes associated with DNA replication, DNA recombination and DNA repair (Carvalho and Lupski, 2016). The organization of repetitive sequences in a region, including their orientation, distribution and density, is important in the formation and evolution of SVs. For example, repeats such as segmental duplications that flank a unique sequence can result in genomic rearrangements by nonallelic homologous recombination. Complex SVs can result by template-switching during DNA replication, linked with fork-stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication. Non-homologous end joining, where double-strand breaks in DNA are repaired without information from a homologous template, is another process generating SVs. SV formation is also associated with the simultaneous generation of SNVs near breakpoints, proposed due to the activity of error-prone DNA polymerase linked with DNA repair (Carvalho *et al.,* 2013). The allele frequency spectrum of SVs show that deletions within genes are significantly rarer than ones found in intergenic regions, a finding not apparent for duplications, suggesting that deletions are less tolerated (Sudmant *et al,.* 2015b). In addition, a strong negative correlation is found between the size of a deletion and its frequency, which appears much weaker for duplications, illustrating the effect of negative selection on CNVs (Sudmant *et al.,* 2015b).

Large-scale SVs which are detectable with traditional karyotyping, such as chromosomal aneuploidies, can result in well-characterized disorders such as Down and Turner syndromes. Moreover, sub-microscopic genomic rearrangements are also associated with multiple traits and disorders, including Angelman and Prader-Willi syndromes (Weischenfeldt *et al.*, 2013). High-throughput short-read DNA sequencing technology have allowed the identification and analysis of SVs at a much finer-scale resolution, however, in contrast to SNVs, SVs are much more challenging to detect (Sudmant *et al.,* 2015a). Multiple variables affect the sensitivity of detecting SVs, including the class, size, sequencing coverage and read length. The latest release of the 1000 Genomes Project estimated sensitivity rates ranging from as high as 80% for deletions to around 32% for inversions, with duplications in between at ~65% (Sudmant et al. 2015a). The size of SVs plays an important role in their discovery; around 80% of relatively small variants (50 bp - 1 kb) are thought to be undetected using short-read methods regardless of their frequency (Chaisson et al. 2015). At the other end of the spectrum, large common SVs that are segmental duplications, which account for ~5% of the human genome are unresolvable using short-reads (Sudmant *et al.,* 2015a). Undetected SVs have been suggested as an important source of disease-causing variation and a significant fraction of the “missing

heritability" in complex disorders (Manolio *et al.,* 2009), but are not usually included in most genome-wide association studies (GWAS), partly due to the challenges associated in their identification and genotyping. This is important as SVs are three times more likely to be associated with GWAS signal than SNVs, and large SVs (>20 kb) are up to 50 times more likely to affect gene expression (Sudmant *et al.,* 2015a; Chaisson *et al.,* 2015).

In this chapter I present the SV analysis of the HGDP dataset sequenced at high coverage using short-read data. The aim of this analysis is first to create a comprehensive catalogue of SV from these populations, and second to use the catalogue to understand human population history.

2.2 Samples, data and genome sequencing

DNA samples were provided by the HGDP-CEPH (Cann *et al.,* 2002). Of the 1063 in the complete panel, a total of 951 core samples (Rosenberg, 2006), 628 of which are male, from 54 populations were analysed. All samples were sequenced at high coverage using Illumina sequencing with either PCR or PCR-free libraries. Ten samples (PCR) were sequenced in a previous study for comparison with the Denisovan genome (Meyer *et al.,* 2012), all using PCR-based libraries (subsequently called "Meyer" samples). 142 samples were sequenced previously as part of the Simons Genome Diversity Project (SGDP), mostly using PCR-free methods (Mallick *et al.,* 2016). The remaining 808 samples were sequenced at the Wellcome Sanger Institute using both library preparation methods, and in some cases both on the same sample, resulting in 823 genome sequences ("Sanger" samples). Twelve SGDP and 2 Meyer samples were also independently sequenced at Sanger with PCR-free libraries. Collectively this generated 975 high-coverage genomes using short-read WGS from 952 samples, which underwent further QC as described in the next section. Each of the Sanger, SGDP and Meyer samples used sequencing technologies with different read lengths (150, 100, and 92 respectively). In addition, 26 samples from 13 populations (2 samples per population) were independently sequenced at high coverage using the pseudo long-read 10x Genomics chromium library preparation platform.

2.3 Sample Quality Control

The heterogeneous composition of samples within this project introduces challenges. Ideally, all samples would be sequenced using the same technology, using the same library preparation and read length. However this is not the case within this project, which may introduce batch effects. In addition, as samples are from lymphoblastoid cell lines, the repeated culturing steps may cause artefactual chromosomal rearrangements, such as large duplications and deletions. These are particularly of concern as they may be confused with genuine germ-line structural variants we are attempting to identify in this study. These artefacts may sometimes affect the viability of the cells, while others may induce proliferation. This illustrates that cell lines are a population of individual cells, and if a *de novo* variant arises within this population, it may proliferate to a maximum extent to become present in all cells, or it may be found in just a subset of them.

I first attempted to examine evidence for chromosomal artefacts by analysing coverage across each chromosome in each sample, by manually visualizing changes in coverage. If a chromosome contains a duplication (one copy addition) it should appear at ~1.5 times coverage, while a deletion (one copy less) should have ~0.5 times coverage. Many samples showed evidence of such abnormalities, which range from multiple whole chromosome duplications to smaller (~3Mb) deletions and duplications. Many more gains of chromosomes than losses were found, and most trisomies affected chromosomes 9 and 12. This has also been found in other cell-line bases studies (Redon *et al.,* 2006), suggesting that these chromosomes harbour certain sequences that increase proliferation once duplicated in cell cultures. For SNV analysis, duplications were shown not to have large effects on genotype calling, and such samples were included for downstream analysis (Bergstrom *et al.,* 2020); however structural variation analysis requires more stringent filtering. I excluded 41 samples from subsequent analysis; these showed chromosomal abnormalities across multiple chromosomes. An example is presented in Figure 2.2. In addition, 74 samples contained regions that show abnormalities, but were limited to mostly a single chromosome or part of a chromosome. To prevent losing important samples that do not have issues across most of the genome, but have limited and observable artefacts, I chose to include these samples in downstream analysis. I, however, masked these problematic regions, i.e. all variants within these regions were set to missing. Artefacts within the sex chromosomes were identified in particular, with mosaic loss of the Y chromosomes being common. A single XXY male was

identified, potentially a natural occurrence rather than a culturing artefact. These quality control steps left 919 samples, which are divided into 644 Sanger PCR-free, 147 Sanger PCR, 111 SGDP PCR-free, 9 SGDP PCR and 8 Meyer samples.

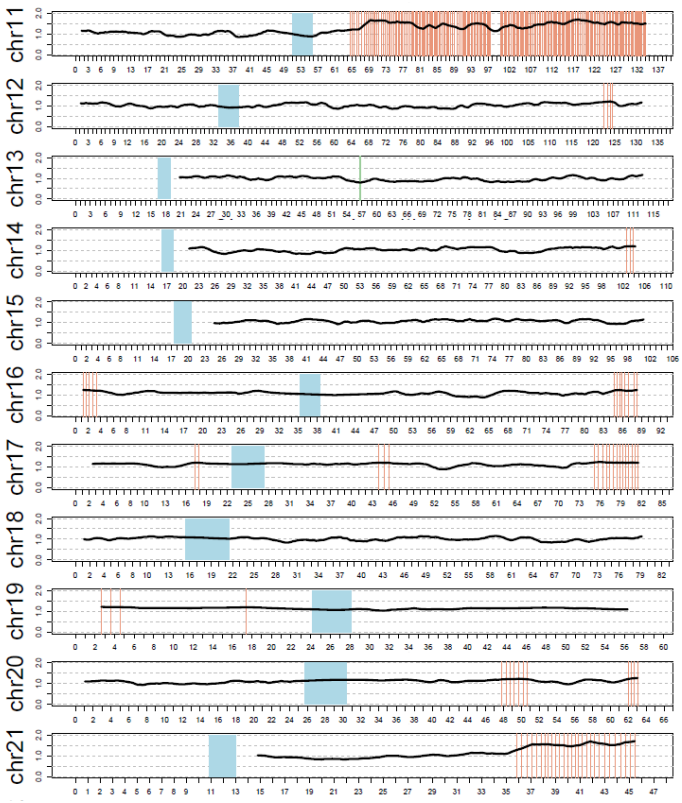

Figure 2.2: Putative cell-line artefacts. Coverage plot illustrating a sample (HGDP00452) that was excluded from the analysis because of likely cell-line artefacts. Coverage was calculated at ~300,000 single positions across the genome and a rolling mean was plotted normalized by the genome-wide median (Y-axis). HGDP00452 shows a large duplication in most of chromosome 11 in addition to smaller duplications in other chromosomes. Orange bars indicate coverage is >25% than chromosome average, green <25%. Blue represents centromeres. X-axis represents chromosome position (Mb).

2.4 Structural Variation Calling and Quality Control

A large number of structural variation callers have been developed and used by different projects. Two studies have recently evaluated different SV callers and provided best-practice recommendations (Cameron *et al.,* 2019; Kosugi *et al.,* 2019). I chose to use Manta (Chen *et al.,* 2016), which is an assembly-based caller, as it seemed to perform well in these studies and generates accurate breakpoints for SVs. Additionally, to complement Manta, I also used GenomeSTRiP (GS, Handsaker *et al.,* 2015), which uses coverage information to identify CNVs, and estimate the copy number of each variant. GS uses joint-calling across all samples

to identify variation in coverage caused by CNVs, and requires at least 30 samples, with higher numbers providing an increase in sensitivity. Manta, on the other hand, uses single sample calling. An advantage of GS is it can also be used to genotype copy number variants in ancient and archaic genomes. As the data in these genomes are mostly short (<50bp) and single-end reads, they cannot be used by many SV discovery algorithms that use information from split-read and read-pair discordance. A limitation of GS is it only identifies copy number variants over 1kb. Manta identifies insertions, deletions, tandem duplications, inversions, and interchromosomal translocations > 50bp. Manta can assemble insertions, but up to a maximum size of approximately twice read-pair fragment size, generally up to ~1kb.

2.4.1 GS quality control

To test for batch effects, I first ran a PCA on genotypes of the GS callset after separating different classes of CNVs (biallelic deletions, biallelic duplications and multiallelic variants). Plotting the first two PCs shows a noticeable batch effect between different libraries preparations and sequencing location. After setting variants with genotype quality (GQ) < 20 to missing, this batch effect becomes negligible (Figure 2.3). However, another noticeable batch effect is that PCR-based samples seem to have more variants than PCR-free. After setting variants GQ<20 to missing, these samples have higher rates of missingness than PCR-free, suggesting these additional low-quality calls are artefacts. Analysing the distribution of missing variants per library and sequencing location also shows a pattern, with the Meyer samples having high missingness > 2%, while the PCR-free libraries showed the lowest missingness (< 0.5%). I therefore excluded the Meyer samples (8 individuals) from subsequent analysis leaving 911 samples. I subsequently tested for excessive heterozygosity: variants that are heterozygous across all samples suggest they are artefacts, and this excluded a small number of samples but the biases in number of calls still remained. I then ran an excessive heterozygosity test for each library and sequencing location separately, and this appears to remove the remaining batch effects. However, after visualising variants I found another issue, that larger variants were split into smaller, sometimes non-overlapping segments. This is a previously reported and known behaviour of the GS algorithm that occurs when there are smaller variants within a sub-segment of a larger polymorphic variant. It also seems to occur if a low-quality variant resides within a larger variant. I subsequently merged high quality (CNQ > 12) calls that have the same diploid copy number and are within 50 kb of each other, in each sample separately, to more accurately estimate the total number of identified CNVs in our dataset. This identified a total of 39,634 autosomal variants, 1,102 variants on the X-chromosome and 289 variants on the Y-

chromosome. Stratifying this dataset by CNV class identifies 22,914 variants as biallelic deletions, 16,012 were duplications, and 2,099 were variants carrying both deletion and duplication alleles (Figure 2.3).

2.4.2 Manta quality control

Manta generated individual VCFs with structural variation calls for each sample, using default parameters (Methods). Only variants that pass all the quality thresholds of the algorithm were used for downstream analysis. I masked the cell-line artefacts identified as described in the previous section and subsequently merged all samples to create one merged VCF. The merged dataset contained a total of 160,958 variants. An issue for Manta callset is that it is not joint-called, and consequently differences in coverage, read lengths, insert sizes and library preparation may also introduce batch effects. The same variant found in one sample may be missed in another due to the differing variables mentioned. To address this issue, I discarded the original genotypes identified by Manta for each sample and then jointly re-genotyped the merged dataset across all samples using Graphtyper2 (Eggertsson *et al.,* 2019). This algorithm generates a graph structure at each variant site which represents the reference genome and the previously-identified structural variants. Reads are subsequently re-aligned to this graph and genotyped. I subsequently excluded variants with size over 10 Mb, as they may be potentially culturing artefacts, and set all variants with GQ < 20 to missing and excluded variants that show excessive heterozygosity as in the previous section. To evaluate whether these steps were successful in removing any batch effects, I ran a PCA for each class of variants (deletions, duplications, insertions and inversions). A batch effect still appeared, so to explore this further I studied the genotype confidence tags provided by the algorithm. I set genotypes with a ‘FAIL1’ tag to missing, and for duplications especially, I also set genotypes ‘FAIL2’ and ‘FAIL3’ to missing. After these adjustments, no batch effects were found across all classes, with the top PCs displaying continental clustering and subsequently population clustering (Figure 2.3). The final analysed Manta callset included 68,089 deletions, 25,084 insertions, 7,290 duplications, 1,895 inversions and 1,667 translocations (Figure 2.3).

2.4.3 Combining datasets, novelty and genotyping evaluation

The two datasets are likely to have some overlap in variants. I used a threshold of 50% reciprocal overlap to identify non-overlapping variants in both callsets and this identified 126,018 unique variants. I extracted overlapping regional-specific deletion calls from the two algorithms and compared variant allele frequencies as a test for genotyping accuracy. High correlation is

found between both callsets (r = 0.97; Figure 2.4), with the slight differences partly due to varying missingness. This suggests the genotyping of the dataset is high quality.

To evaluate the novelty of the identified variants, I compared the HGDP dataset with two published global-scale structural variation callsets:

1) The 1000 Genomes Phase 3 Structural Variation Dataset (Sudmant *et al.,* 2015a)
2) The copy number analysis of the Simons Genome Diversity Project (Sudmant *et al.,* 2015b).

As the SGDP callset is mapped using GRCh37, I lifted over the calls to GRCh38. I used a threshold of 30% reciprocal overlap between variants to classify them as the same call. Using a higher threshold increases the novelty rate for our dataset, and since the published studies used older algorithms and are based on shorter reads, this increases the size of the confidence interval around the breakpoints of the SV. Therefore I used a more relaxed threshold to be more conservative. In addition, during the comparison I chose to be conservative by evaluating whether a locus is structurally variable, instead of comparing each class of variant between the callsets (e.g. deletions vs deletions). This is due to a possible misclassification of variant class (e.g., insertion vs. duplication). This test showed that 78% of variants in our dataset are not found in either the 1000GP or SGDP callsets. I also evaluated the number of variants in the SGDP and 1000GP that are not found present in our HGDP dataset, which is 53% and 64% respectively. In this comparison I included all variants identified in the 1000GP, even classes that were not included in this analysis, such as large mobile element insertions. Of the 1000G variants that are not present in our dataset, a large percentage, 93%, appear to be rare (MAF < 1%). This is expected to an extent, as around half of variants in published WGS datasets appear as singletons. As the 1000GP is a low-coverage dataset, this may also increase the number of false positive variants. A major advantage of our callset is it includes the abundant, but understudied, class of relatively small variants (50bp – 100bp) which were not effectively characterized by the previous projects (Figure 2.3). At this size range, 91% of variants in our callset are not found in either published resource. Of these novel calls, an appreciable number are common and even high-frequency within regional groups and individual populations (Figure 2.5).

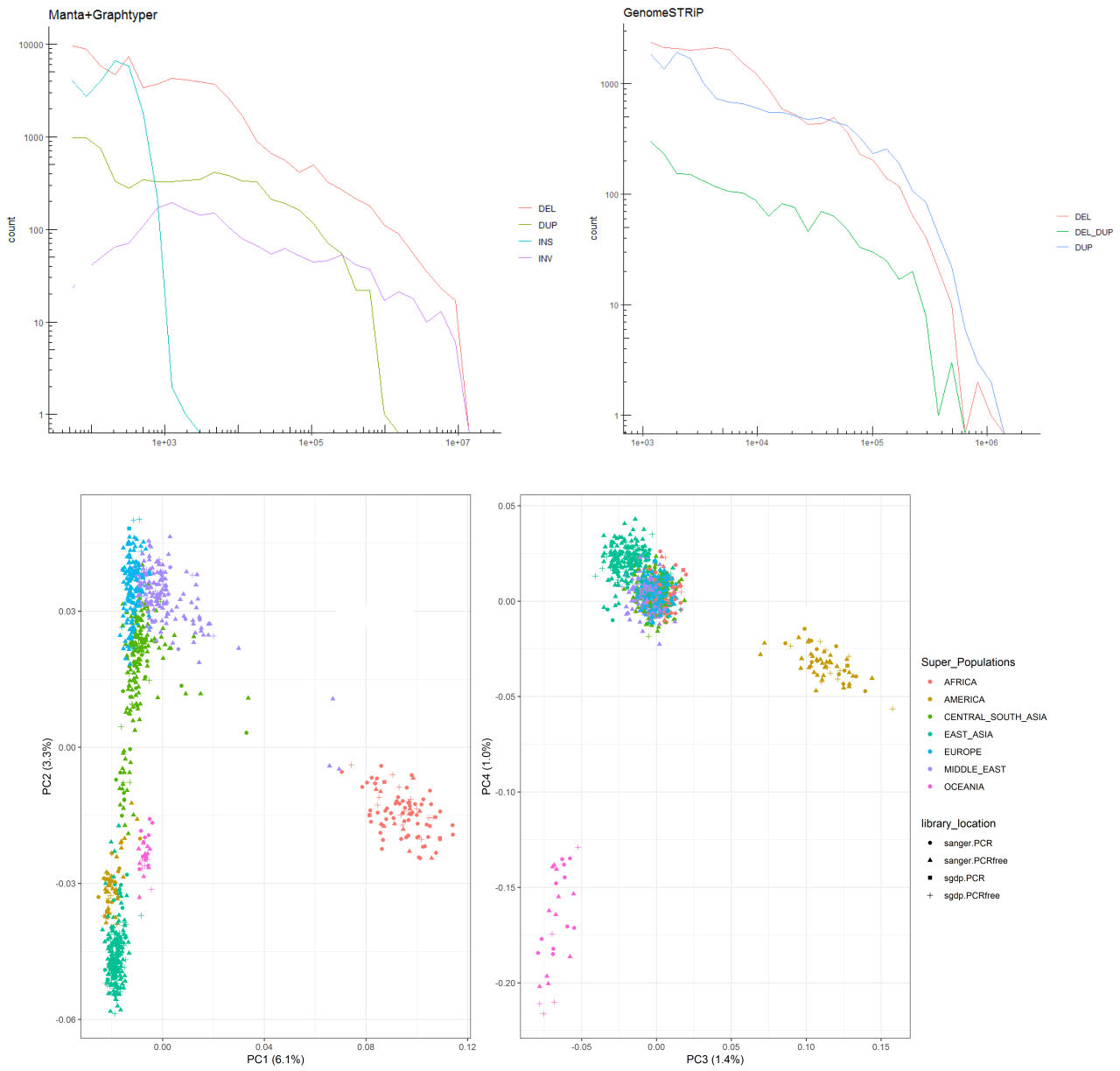

Figure 2.3: Variant Size distribution and Quality Control. Top: Size distribution of variants identified by Manta (Left) and GS (right). DEL: Deletion, DUP: Duplication: INS: Insertion, INV: Inversion, DEL_DUP: multiallelic. Bottom: PCA 1-4 of deletion genotypes called by GS stratified by library preparation method and sequencing location. After quality control, no batch effects are observed across all classes of SVs identified by GS and Manta after viewing 100 PCs.

To further test the quality of genotyping our dataset, I extracted common deletions (MAF > 5%) from African groups in the 1000GP that overlap variants in our callset (50% reciprocal overlap). I selected deletions as they had the highest sensitivity in the 1000GP dataset, and also because this dataset mostly consisted of deletions (82%, excluding mobile element insertions). Some variation in allele frequencies is expected as the HGDP and 1000GP include different African

populations, however, using common variants at a continental-level there should be some correlation. Indeed, a strong correlation is observed (r = 0.95; Figure 2.4). These results collectively suggest that the quality control steps performed resulted in a high-quality dataset free of batch effects that can be used in population-genetic analysis. It also demonstrates that our dataset contains a large amount of previously undocumented variation (Figure 2.5), likely due to the more diverse populations sampled, higher coverage and improved SV detection tools. This underscores the importance of studying underrepresented human populations.

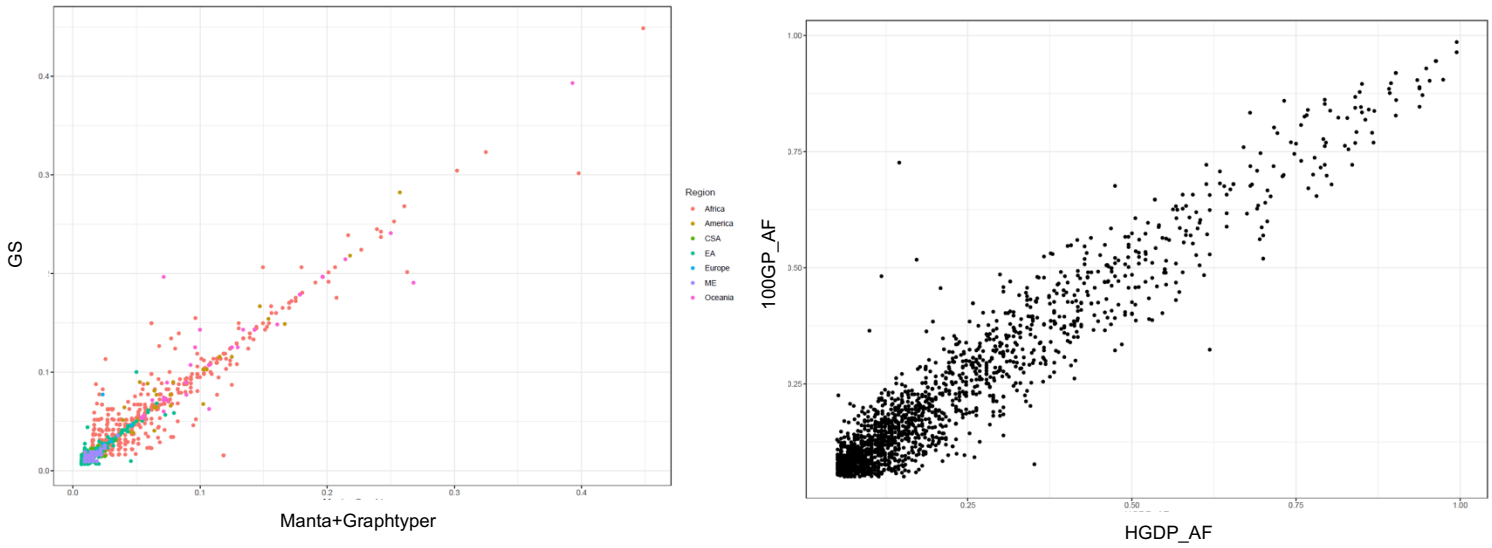

Figure 2.4: Testing the quality of genotyping. **Left**: Correlation of allele frequency of variants (Regional-specific variants, coloured by region) identified by both Manta+Graphtyper (X-axis) and GS (Y-axis) within the HGDP dataset. CSA: Central & South Asian, EA: East Asian, ME: Middle East. **Right**: Allele frequency (AF) correlations between deletions identified in the 1000GP (Y-axis) and the HGDP Manta+Graphtyper callset (X-axis) using common (> 5% MAF in 1KG) African-specific deletions.

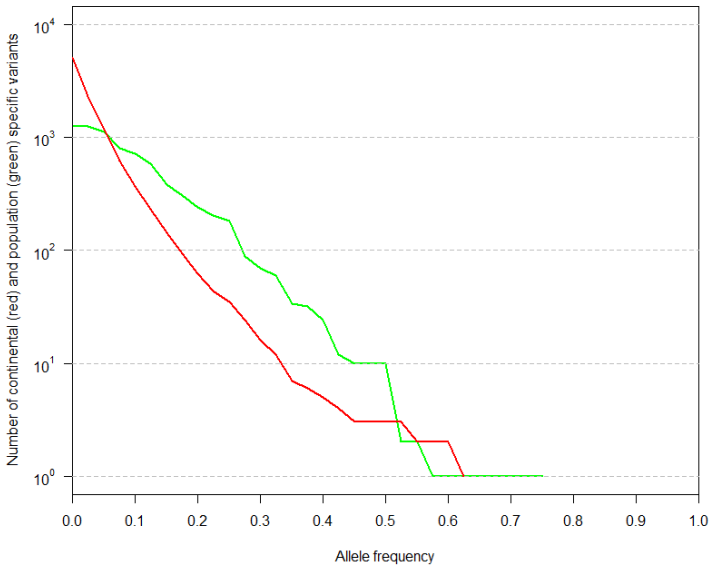

Figure 2.5: **Novel population-specific variants identified in this study.** Continental- (red) or Population- (green) specific variants (minor allele count > 2) in identified in the HGDP dataset but not present in the 1000GP or SGDP SV callsets binned by allele frequency. The same variant can be present in both distributions.

2.5 Population Structure

SNVs have been well-studied for the analysis of population structure, but SVs are less well understood. I separated each class of SVs and then ran a PCA on the genotypes of each class. For deletions, which are the most numerous class in the dataset, clear structure was found across a large number of PCs. As expected, the first PC separated African from non-African populations, while the second PC separates Europeans, Middle Easterners and South Asians from East Asians, Papuans and Americans. PC3 and PC4 distinguish Americans and Papuans, respectively. Subsequent PCs separate African populations, illustrating the diversity of the dataset. For biallelic duplications the structure was limited to the first 4 PCs, but surprisingly showed a different pattern than deletions, with PC2 separating Papuans from the rest of the dataset. I evaluated this further by analysing the PCA loadings of each variant, and found two variants with very high loading on PC2. When I excluded these variants from the PCA, the pattern returned to one similar to deletions. These two variants, on chr16p.12 and chr8p.21 are discussed in more detail in the archaic admixture section. Because GS only identifies a small number of biallelic duplications, a small number of variants that shows strong stratification will affect the patterns of variation observed in the PCA results. Insertions also showed a very similar pattern to deletions. To evaluate population structure further, I ran a 2-dimensional UMAP on all PCs that show structure (Figure 2.6). This provided much more

resolution; deletions show clear separation of continental groups, and even individual populations are clearly separated in many cases. The divergent African populations such as the San, Mbuti and Biaka form their own clusters away from the remaining African populations. Populations that are known to be admixed such as the Uygur and the Hazara cluster separately from the East Asian and Central & South Asian groups. Populations that have experienced relatively high amounts of genetic drift such as the Kalash, in addition to Oceanian and American populations, are also clearly differentiated. For populations that show less clear clustering and project into large continental clusters, I observe examples of finer structure, as samples from the same population generally projected closer to themselves than to other groups. Other classes of SVs (insertions, duplications, multiallelic variants) also show population structure, though not as well-defined as deletions (Figures 2.6). What is particularly notable is that Oceanian populations always remain well-differentiated in all classes. The observed differences in clustering between different classes of SVs likely partly reflect the varying mutational patterns generating each class of structural variant. Duplications, and multiallelic variants especially, are more commonly found in repetitive regions in the genome. This increases the probability of a recurrent mutation, with a variant mutating to a higher or lower copy number, and thus reducing the differentiation between populations. Additionally, the total number of variants in each class of SVs will also affect the pattern of structure (Figure 2.3), as higher numbers, such as in the case of deletions, will increase resolution. Another potential variable impacting the structure is the accuracy of genotyping, as low genotyping quality will also reduce the pattern of structure. However, this is unlikely to be a major factor in this analysis, as I previously showed that genotyping appears of high quality.

2.6 Population Stratification

I subsequently searched for highly stratified variants between populations, which can be potentially a result of selective pressures. As the HGDP dataset has a reasonably large number of samples per population, for each population pair I evaluated the relationship between average population differentiation, calculated using SNV-based F_{ST}, and the variant allele frequency difference. It has been proposed that variants which are outliers to this relationship are likely to be under selection (Coop *et al.,* 2009, Huerta-Sánchez *et al.,* 2014). The distributions of both deletions and biallelic duplications appear similar although duplications show lower stratification. Notable outliers are apparent (Figure 2.7): a deletion in *HBA2* is almost fixed in Lowland/Sepik Papuans (86%), while it is not found Papuan highlanders

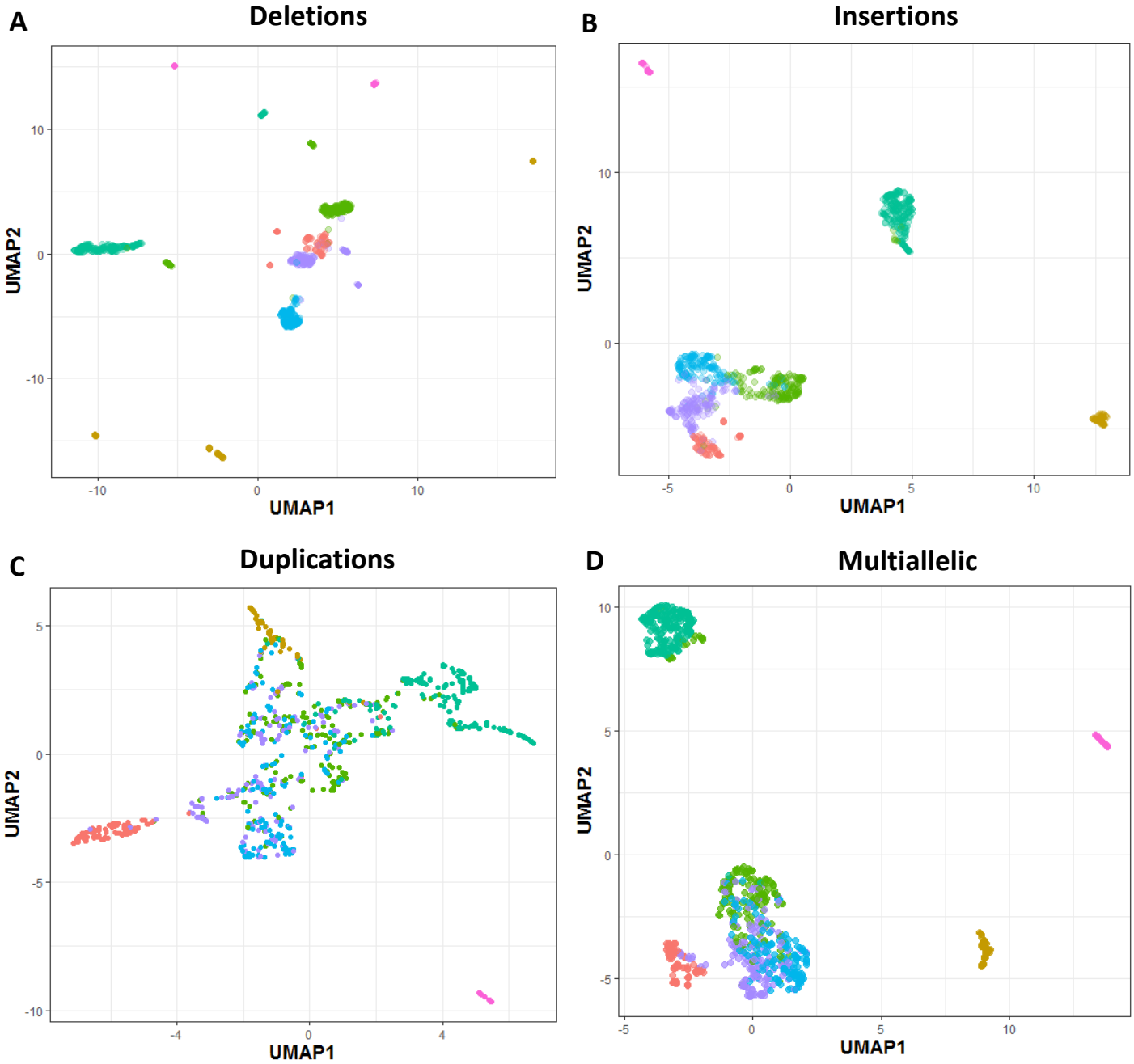

Figure 2.6: **SV population structure**. Each point represents a sample while colours illustrate regional labels as in Figure 2.1. **A:** UMAP of biallelic deletion genotypes. **B:** UMAP of insertions. **C:** UMAP of biallelic duplications. **D:** UMAP of multiallelic variants**.**

(p < 0.001, population stratification test using 1,000 permutations). *HBA2* encodes one of the alpha globin chains of haemoglobin, and it has been suggested that high frequencies of α-globin deletions are protective against malaria. Intriguingly, malaria is present in the lowlands of Papua New Guinea, but not in the highlands (Flint *et al.,* 1986). Another highly stratified variant is a deletion within *MYO5B,* which encodes a motor protein that is high frequency (88%) in the Lahu from China (Lahu-Hezhen, p = 0.001), while being rare in closely-related populations. This group shows a unique population history in comparison to its neighbours, as it also shows high numbers of private SNVs and also carries rare divergent Y chromosomes (Bergström *et al.,* 2020).

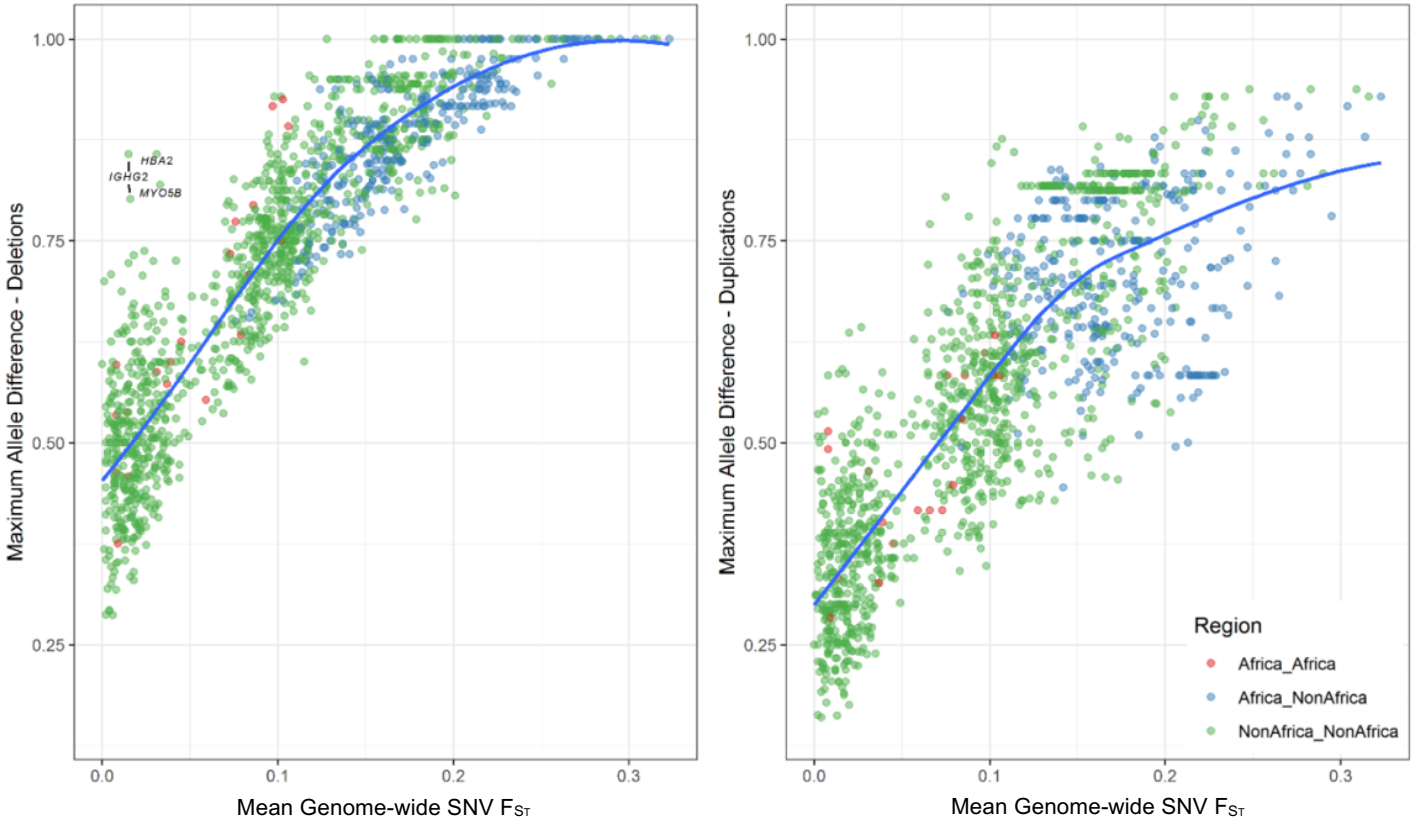

Figure 2.7. **Population Stratification**. Maximum allele frequency difference as a function of population differentiation for 1431 pairwise population comparisons. Blue curve represents loess fits. Deletions (**Left**) and Biallelic Duplications (**Right**) from the GS callset. Three outlying stratified variants are illustrated. *HBA2* deletion in Papuan lowlands, a deletion within *MYO5B* particularly common in Lahu, and a deletion downstream of *IGHG2* almost fixed in Dai (86% frequency).

2.6.1 Population-private variants

The dataset also allows us to look for variants that are unique to a population, which likely suggests that they are recent mutations (Figure 2.8). I searched for common population-private variants, and find some that both reach high frequencies and seem to have a functional effect (McLaren *et al.,* 2016). A 14kb deletion in the Karitiana population of South America is present at 40% frequency and removes the 5′ upstream region and up to the first exon of *MGAM*, potentially inactivating the gene. This encodes maltase-glucoamylase (MGAM), an enzyme that is highly expressed in the small intestine and is important in the digestion of plant starches (Nichols *et al.,* 2003). A recent ancient DNA study of ancient Andean individuals suggested that variants in this region show evidence of selection, and proposed it was due the transition to an agricultural lifestyle (Lindo *et al.,* 2018). In addition, in dogs this gene also shows signals of selection, which

Figure 2.8: Population-Specific Variation. Each point represents a variant private to a population (n > 2) with the x-axis reflecting its frequency. Colours represent regional labels and random noise is added to aid visualization. High-frequency variants discussed in the text are highlighted.

has also been suggested as an adaptation to a diet rich in starch during domestication (Axelsson *et al.,* 2013). If this deletion does limit the expression of the gene, it is puzzling why such a variant would reach high frequency in a population which includes starch in its diet. The population history of the Karitiana may provide some insights: they have suffered an extreme recent population crash which results in them having the highest levels of runs of homozygosity of any human population studied to date (Ceballos *et al.,* 2018; Bergstrom *et al.,* 2020). Using the population branch statistic (PBS), this variant shows suggestive but not strong evidence for a departure from neutrality (98.7% rank). The presence of homozygous individuals for this variant and its high frequency suggests that the population crash increased genetic drift to an extent that may have counteracted the advantage of the ability to digest starch. Nevertheless, even if the deletion does not have a major effect on the function of the protein, it could have also drifted to high frequency for the reasons above. In the HGDP dataset, there are distinct deletions present in this region, with one removing exons 34 to 38 (out of 48) present at 11% global frequency. In addition, a previous study reported an individual with a homozygous deletion within *MGAM*, but appeared to still have functioning protein activity (Eccleston *et al.,* 2012). The activity of maltase-glucoamylase in the small-intestine forms the final step in digesting linear regions of starch to glucose, and its activity is complemented by sucrase-isomaltase (SI). The exon structure of both enzymes is identical, however MGAM can hydrolyse maltose and starch, not sucrose, which can be hydrolysed by SI in addition to isomaltose (Nichols *et al.,* 2003). Further work is needed to understand the functional effect, if any, of the deletion and if its activity can be compensated by another enzyme.

Another private variant which reaches high frequency (54%) is a deletion that removes most of *SIGLEC5* in the Central African hunter-gatherer Mbuti. This gene is part of the Siglec gene family, which encodes cell-surface receptors expressed on immune cells. They bind to glycans that contain sialic acids expressed on host cells to differentiate ‘self’. Most Siglecs act to inhibit leukocyte activation after detecting host cells, including *SIGLEC5.* Adjacent to *SIGLEC5* is another member of the family, *SIGLEC14,* proposed to have evolved by gene conversion from *SIGLEC5* (Angata *et al.,* 2006). In contrast to most of the gene family, *SIGLEC14* binding results in leukocyte activation, and is thought to have originated by providing a selective advantage in combating pathogens that mimic host cells in expressing sialic acids (Akkaya and Barclay, 2013). These two genes are an example of paired receptors, one with an inhibitory function and the other activating, which are important in fine-tuning immune responses. The deletion we find in Mbuti likely removes the function of the inhibitory receptor, *SIGLEC5,* while

the activating receptor is maintained. Such an event has been proposed to result in immune hyperactivity and potentially autoimmune disease, and consequently is a surprising finding (Lubbers *et al.,* 2018). The deletion shows an extreme PBS (99.87% rank), indicative of positive selection. Additionally, it is the most extreme population-private variant in Mbuti. Another complication at this locus is the presence of another known and polymorphic deletion, which has the opposite effect, removing most of the activating receptor *SIGLEC14*. This deletion is found at 38% global frequency, but particularly higher in East Asians (63%). This common deletion results in the fusion of *SIGELC5* and *SIGLEC14*, creating a new gene which has the coding sequence of *SIGLEC5* but expressed by the promoter of *SIGLEC14* (Yamanaka *et al.,* 2009). I found only one individual within the dataset (HGDP00450), who has both deletions on separate haplotypes. Using only coverage information, the region appears ambiguous. However, fortunately this individual was also independently sequenced using linked-read technology. This resolved the phase at this locus and clearly shows two distinct deletions (Figure 2.9). The environment of the Mbuti population, who live in the Ituri Rainforest in the north-east of the Democratic Republic of the Congo, may have exposed them to a pathogen that created a selective pressure increasing the frequency of this private deletion. Without functional analysis of this variant, this is only speculation, but clearly this variant requires further study to elucidate its function and implications.

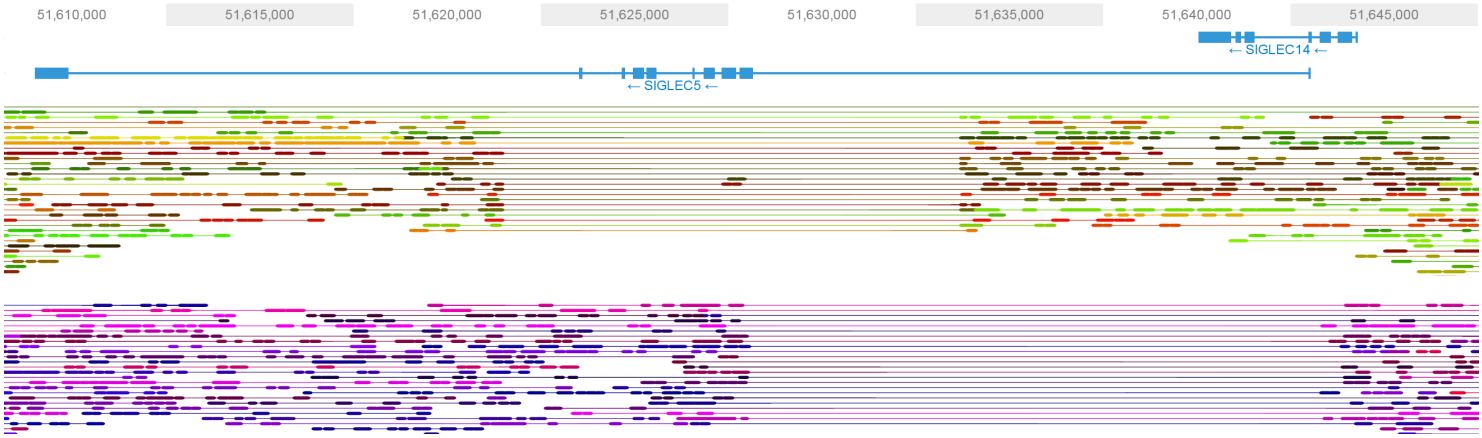

Figure 2.9: **Distinct deletions at the *SIGLEC5/SIGLEC14* locus in an Mbuti sample (HGDP00450) resolved using linked-reads**. One haplotype (top) carries the Mbuti-specific variant that deletes most exons in *SIGLEC5* and is present at high frequency (54%), while the second haplotype (bottom) carries a globally common deletion that deletes *SIGLEC14*, creating a fused gene.

A 2.7kb deletion on chromosome 15 within *OCA2* (also known as the *P* gene) appears private to the South African Bantu population at an unusual frequency (44%, Figure 2.8). However, this is a well-known deletion reported previously in African populations which causes Brown Oculocutaneous Albinism (Manga *et al.,* 2001). Many homozygotes for this deletion are found in the HGDP Bantu South Africans, suggesting that individuals who have donated samples to the resource had albinism. After contacting CEPH, which hosts the resource, about this information we were informed that the researcher who donated the samples to HGDP was Trefor Jenkins (now deceased). Since he has published many studies on albinism in African populations (e.g. Manga *et al.,* 2001), this suggests that the frequency of this deletion is a result of particular sample ascertainment instead of its general frequency in the South African Bantu population. Using a pair of samples homozygous for the deletion, I find that SNVs around the deletion (±150kb) have an excess proportion of identity-by-decent (IBD = 0.82) relative to the total calculated from all SNVs on a different chromosome (chromosome 1; IBD = 0.02). Moreover, IBD calculated on all SNVs on chromosome 15 (IBD = 0.49) also show higher values than chromosome 1. This suggests that this deletion has likely been inherited from a common ancestor, rather than arising independently in both samples.

2.7 Archaic Introgression

To understand the landscape of SV archaic introgression in modern populations, I genotyped variants identified in this dataset in two high-coverage Neanderthal (Altai and Vindija; Prüfer *et al.,* 2014, Prüfer *et al.,* 2017) and one Denisovan genome (Meyer *et al.,* 2012). I first compared the number of shared variants between the archaic genomes and modern-day populations, and found hundreds that are exclusive to archaic and African genomes. This suggests that these variants were part of the ancestral variation found in the ancestor of modern humans and archaic hominins, but were subsequently lost in non-African populations during the out-of-Africa bottleneck. I subsequently searched for variants that are present in both archaic and non-African populations, but not found in Africans. This filtering step will identify putatively introgressed variants, and similar strategies are used in SNV-based methods to identify introgressed haplotypes (Skov *et al.,* 2018). An important assumption of this approach is that African populations within the dataset do not have archaic sequences introduced by introgression. I focused on common regionally-stratified variants, and this identified ones with a wide range of sizes, the smallest 63 bp and largest 30 kb. Interestingly, all these variants reside within or near genes, suggesting they have likely functional implications (Table 2.1). The most

extreme putatively introgressed variant in frequency is composed of two regional-specific duplications found in Oceanians on chromosome 16p12.2, both present at 82% and appearing in perfect LD. This variant is shared with the Denisovan genome, but not with either Neanderthal genome. This result has also been reported in a study of smaller scale (Sudmant *et al.,* 2015b). Exploring the frequency of this duplication in our larger dataset within each Oceanian population, I found that it is present at a similar frequency in all three Oceanian populations (~82%). This result is intriguing, as the Bougainville Islander population, in contrast to the Papuan Highlanders, have significant East Asian admixture (~20%) which does not dilute the frequency of this variant. These duplications show a remarkably unusual allele distribution, with a PBS rank of 99.99% and are the most extreme regional-specific variant in the entire dataset (Figure 2.10 and Table 2.1). I also compared the distribution of regional-specific variants between SNVs and CNVs, and found that while they appear similar for most regional populations, the Oceanians appear as an exception: They have a higher excess of high-frequency private CNVs compared to what would be expected based on the number of private Oceanian SNVs (Figure 2.11). Most of these variants are also found in the Denisovan genome and appear introgressed (Figure 2.10). The unusual distribution of these variants (Figure 2.11; Table 2.1) indicates that positive selection may have increased their frequency in Oceanians.

Position	Size (bp)	Variant	EUR	CSA	EA	ME	AMR	OCE	Gene	PBS rank %	NEA	DEN
chr1:64992619-64992994	375	DEL	0	0	0	0	0	0.44	***JAK1***	*98.4*	REF	DEL
chr2:3684113-3690212	6099	DEL	0.02	0.003	0.05	0.03	0	0.26	***ALLC***	*90.3*	DEL Vin	REF
chr3:177287011-177292441	5430	DEL	0	0	0	0	0	0.39	*LINC00501*	97.7	REF	DEL
chr8:23124835-23130567	5732	DEL	0	0.02	0	0	0	0.36	*TNFRSF10D*	*96.8*	DEL	REF
chr8:23134649-23164796	30147	DUP	0	0	0	0	0	0.48	***TNFRSF10D***	*99*	DUP	DUP
chr11:60460681-60461880	1199	DEL	0	0	0.02	0	0.17	0	***MS4A1***	-	DEL	REF
chr12:101882163-101883377	1214	DEL	0.02	0.08	0.32	0.01	0.01	0.33	***DRAM1***	-	DEL	REF
chr12:104799951-104803150	3199	DUP	0.003	0.009	0	0.01	0	0.33	***SLC41A2***	*96.8*	DUP	REF
chr15:34920811-34925992	5181	DEL	0	0	0	0	0	0.63	***AQR***	*99.8*	REF	DEL
chr16p12.2	Complex	DUP	0	0	0	0	0	0.82	Multiple	*99.99*	REF	DUP
chr16:75059992-75060055	63	DEL	0	0	0	0	0	0.34	***ZNRF1***	*96.4*	DEL	DEL
chr17:3038851-3041981	3130	DEL	0	0	0	0	0	0.16	*RAP1GAP2*	*86.1*	DEL	DEL
chr19:42529806-42531042	1236	DEL	0	0	0	0	0	0.54	***CEACAM1***	*99.4*	DEL	DEL

Table 2.1: **Allele frequencies of regionally-stratified variants shared with high-coverage archaic genomes but not found in African populations**. Neanderthal refers to both published high-coverage genomes. If a variant lies within or intersects a gene it is highlighted in bold, otherwise the nearest gene is listed. The deletion within *ALLC* is only shared with the Vindija Neanderthal. The *TNFRSF10D* duplication common in Oceania is also present at low frequency (5%) in Africa. Africans do not have both deletion and duplication variants, which are in LD in Oceanians (r^2 = 0.48). The duplications at chr16p12.2 at high frequency in Oceania (82%) are part of a complex structural variant (Figure 2.12). PBS rank is presented for stratified variants common only in Oceania. EA - East Asia, ME - Middle East, AMR - America, CSA - Central South Asia, OCE - Oceania. NEA – Neanderthal, DEN – Denisova.

As the 16p12.2 duplication was identified through an increase in coverage using GS, the actual location of the variant in the genome is still unknown. This is because we are only mapping reads to the reference, but the duplication could be interspersed, even located on a separate chromosome. To further understand this variant, we characterized it in more detail using fluorescent *in situ* hybridization (Figure 2.12). This showed that the duplication consists of a region of the reference sequence that has duplicated and inserted in an inverted orientation into a gene-rich region ~7 Mb away in chr16p11.2. However, it does not seem to be simple duplication, as another sequence ~1Mb away from the original site is also present in the inserted site. This is consistent with GS calling two separate duplications in perfect LD.

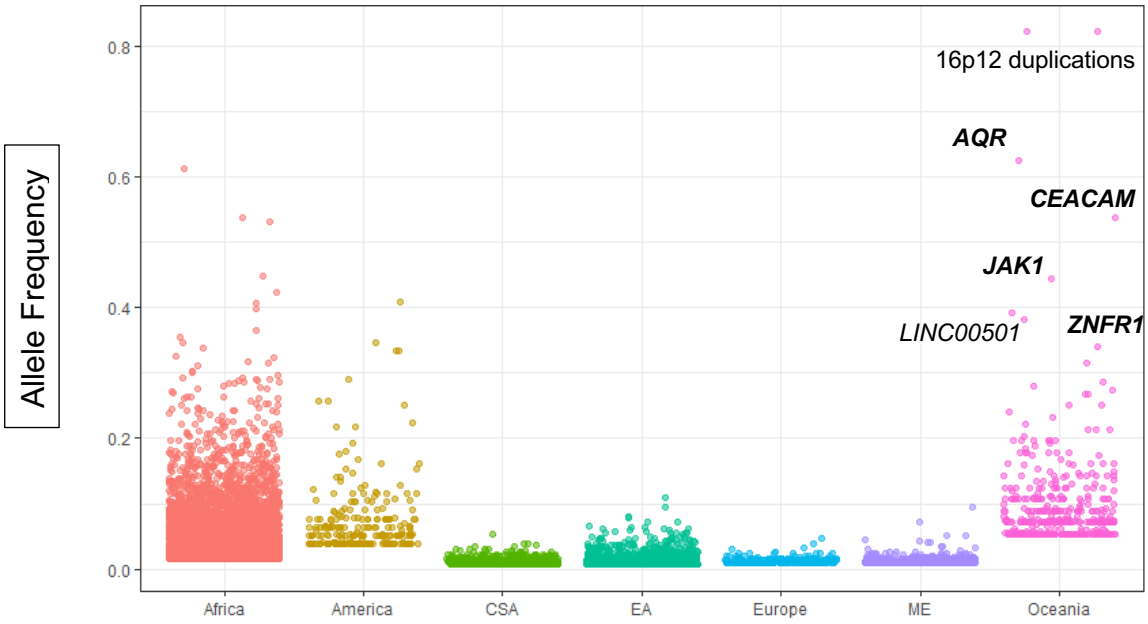

Figure 2.10: **Regional-Specific Variation**. Each point represents a variant private to a regional group (n > 2) with the y-axis illustrating its frequency. Random noise is added to aid visualization. The distribution reflects the ancestral diversity in Africa, the connectivity of Eurasia, the isolation & drift of the Americas and Oceania, and the separate Denisovan introgression event in Oceania. Oceania is notable for having private high-frequency variants that are all shared with the Denisovan genome and are within (**bold**) or near the illustrated genes, four of which are newly identified in this study (*AQR*, *CEACAM*, *JAK1, ZNFR1*). The Americas contain high frequency variants which are not shared with any archaic genomes, suggesting they arose and increased to high-frequency after they split from other populations. EA: East Asia, CSA: Central & South Asia, ME: Middle East.

I also evaluated the region using samples sequenced using linked-reads by analysing the shared barcodes in the region. Using all the information available, it appears that the duplication is a complex rearrangement involving a duplication-inverted-insertion, an inversion and a deletion. This locus is known to show complex recurrent structural changes as it lies in a repetitive region, and variants within this locus are associated with ~1% of autism cases (Weiss *et al.,* 2008). While we conclude that this variant increased in frequency after the archaic

admixture event, the target of selection remains unknown as the insertion lies in a region with a large number of genes. In addition, the selective pressure acting on this variant is unclear; however, its similar frequency across all three Oceanian populations appears in contrast to the variable frequency of the *HBA2* malaria-associated deletion in the region. This indicates that malaria infection is unlikely to be the selective pressure driving the increase in frequency of the 16p12.2 duplication.

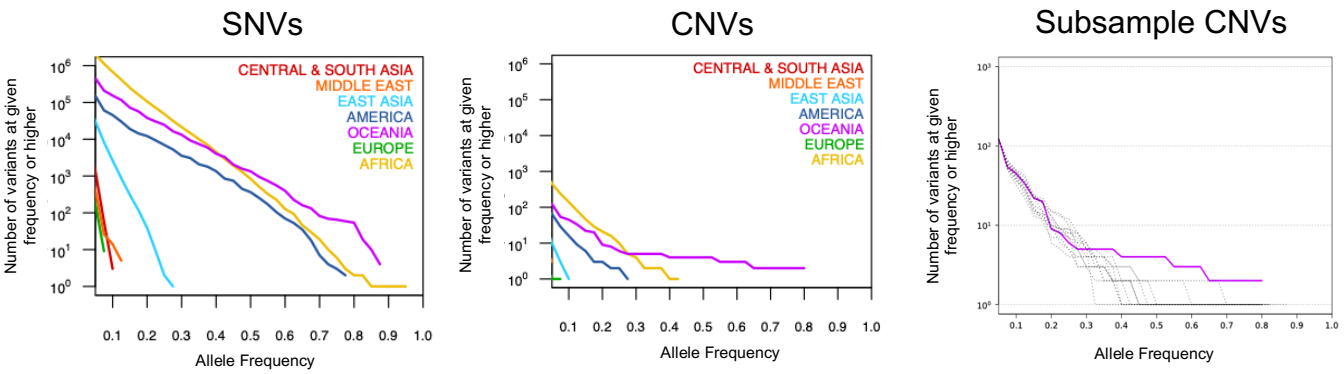

Figure 2.11: Comparison of regional-specific variants across SNVs and CNVs. Comparison of Regional-specific variants between SNVs (**Left**) and CNVs (**Centre**). To test whether the enrichment of high-frequency private Oceanian CNVs relative to SNVs could be due to sampling noise or represents positive selection, I randomly sampled private Oceanian SNVs (matching the number of private Oceanian CNVs) 1000 times (dashed lines) and compared the frequency distributions in these random samples to the observed CNV distribution (purple, **Right**). Only 12/1000 sample sets had a variant with an equal or higher frequency than the single most frequent CNV (at a frequency of 82.14%) and are shown. However, even among these 12 sets, none appear to match the distribution at the higher frequency range. Note the Y-axis scale in the right figure is different to visualize the random samples.

In chromosome 8p21.2 an intriguing deletion-duplication variant which is shared only with Neanderthals is located in a region containing two genes, *TNFRSF10C* and *TNFRSF10D*. These variants are common in Oceanians but rare globally (Table 2.1). The 5.7 kb deletion is located between *TNFRSF10C* and *TNFRSF10D*, while a 30kb duplication encompasses all of *TNFRSF10D*. The frequency distribution of the two variants appears complicated; the duplication is common in Oceania (48%), but also present at low frequency (5%) in Africa. However the deletion is not found in Africans, but is present in high frequency in Oceanians (36%) and is in moderate LD with the duplication (r^2 = 0.48). These two genes encode cytokine receptors which form members of the tumour necrosis factor receptor superfamily (TNFSF). These receptors are commonly found on leukocytes, and their activation is involved in diverse cellular processes, including inflammation and apoptosis (Johnstone *et al.,* 2008). In the same region two other genes with similar sequences, *TNFRSF10A and TNFRSF10B,* are found, also

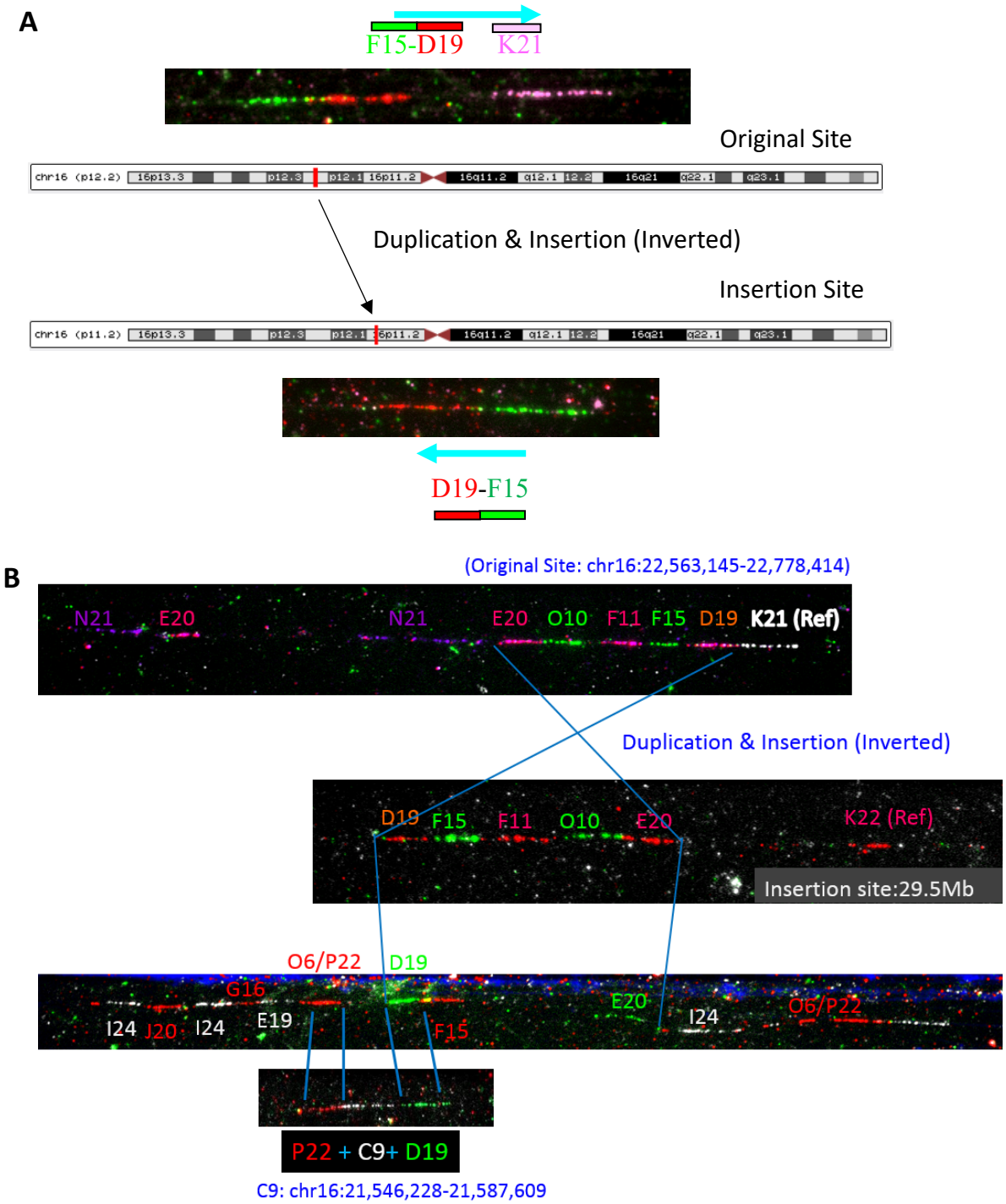

Figure 2.12: chr16p12 Papuan-specific expansion shared with Denisovan genome in more detail. **A:** Cartoon illustration of location of original (16p12.2) and inserted site 7Mb away (16p11.2). Names and colours of probes (e.g. K21) are indicated. **B:** Fiber-FISH illustrating the original site (top), the (inverted) insertion sites (centre) and the region surrounding the insertion site (bottom). Region flanking the insertion site (C9) is a sequence 1Mb away from the original site, consistent with GenomeSTRiP calling a second duplication at this site in perfect LD with the initial duplication. This suggests a complex event involving a duplication-inverted-insertion, an inversion and a deletion.

called Death receptor 4 and 5 respectively. These two contain receptors that mediate apoptosis upon binding TNF-related apoptosis-inducing ligands (TRAIL). However, *TNFRSF10C* and *TNFRSF10D*, do not contain receptors that induce apoptosis and are called Decoy receptors 1 and 2. It has been suggested that these receptors act as an antagonist to protect cells from apoptosis caused by TRAIL (Johnstone *et al.,* 2008). Since the four genes appear very similar in sequence, a duplication that encompasses all of *TNFRSF10D* could have potentially resulted in a new gene.

Multiple additional high-frequency variants are private to Oceanians and are shared with the Denisovan genome (Figure 2.10 and Table 2.1). This illustrates the long-term isolation of Oceanian populations, and the separate Denisovan introgression event (Browning *et al.,* 2018, Jacobs *et al.,* 2019). Many, but not all, show unusual PBS values (Table 2.1), indicative of positive selection. A very common deletion, at 63% frequency is found within *AQR*, an RNA helicase gene. RNA helicases play an important role in detecting viral RNAs and mediating antiviral immune response, and are a necessary host factor for viral replication (Ranji and Boris-Lawrie, 2010). In addition, AQR is known to be associated with regulating the integration of HIV1 DNA, and in recognizing and silencing of transposable elements (König *et al.,* 2008; Akay *et al.,* 2017). Two other Denisovan-shared and Oceanian-private deletions reaching high frequency are in *JAK1*, which encodes a kinase essential in cytokine signalling (44%), and in *CEACAM1* (also known as CD66a), a glycoprotein part of the immunoglobulin superfamily (54%) which modulates immune responses associated with inflammation and infection.

Outside Oceania, a deletion, shared only with Neanderthals, reaches ~26% frequency in the Surui and Pima of the Americas. This deletion removes an exon of *MS4A1*, which encodes the B cell differentiation antigen CD20. This gene has recently been a target of multiple developed monoclonal antibodies for B cell-associated leukemias, lymphomas, and autoimmune diseases (Kuijpers *et al.,* 2010, Marshall *et al.,* 2017), as it plays an important role in T cell-independent antibody responses. This suggests that this deletion could have medical implications, with therapies developed for one population potentially not effective in others.

2.8 Multiallelic variants and runaway duplications

CNVs are not limited to biallelic deletions and duplications, but also include multiple copies. I identified a dynamic range of copy number expansions, and some variants that were previously thought to be biallelic contained additional copies in our diverse dataset. I focused here on a

particular type of multiallelic variants, called ‘runaway duplications’ (Handsaker *et al.,* 2015). In these, the duplicated units are found in most populations at low copy numbers, but expand to much higher copies in a small number of populations, potentially due to a regionally-restricted selection pressures (Figure 2.13). It should be noted that the number of copies stated is for a diploid genome and expressed relative to the reference genome, which itself may have an atypical copy number. Experimental analysis have shown that GenomeSTRiP provides very accurate integer copy number estimation for multiallelic variants, even at high copy numbers (Handsaker *et al.,* 2015).

Some runaway duplications are only found within Africa, or in populations with recent African admixture (Figure 2.13A and 2.13F). The hunter-gatherer Biaka from the Central African Republic have a private expansion downstream of *TNFRSF1B*, where many individuals have 9 copies. All other populations within the dataset have only 2 copies. This gene is also a member of the TNRSF, as is the introgressed deletion-duplication variant previously discussed in the Oceanian populations. We also identified expansions in *HPR* (Figure 2.13A), which has been previously reported in some African populations (Handsaker *et al.,* 2015, Sudmant *et al.,* 2015b). This gene encodes a protein which is associated with defence against trypanosome infection (Smith *et al.,* 1995). Interestingly, populations that have the highest copy number are Central and West African groups, correlated with the geographic distribution of the infection (Franco *et al.,* 2014). However, we also identify the expansions in all Middle Eastern groups, but at a lower frequency. This is possibly due to recent admixture from African groups.

A striking expansion is found upstream of the olfactory receptor gene *OR7D2* which is restricted to samples with East Asian ancestry (Figure 2.13B). The expansion ranges from 2-18 copies, and through haplotype phasing I find that many samples contain the expansion on just one chromosome. This demonstrates that these expansions have mutated repeatedly on the same haplotype background. One particular Han Chinese sample had a high copy number, 18 copies. This individual appears to have nine copies on each chromosome, indicating that the same runaway haplotype is found twice in the same person. This could in the future lead to an even higher number of copies through non-allelic homologous recombination.

I then focused on medically-relevant expansions. One is found encompassing all of *HCAR2* (Figure 2.13C). Although the duplication is found in many continental groups, it is especially common in the Kalash population of Pakistan, with almost a third of the population carrying an

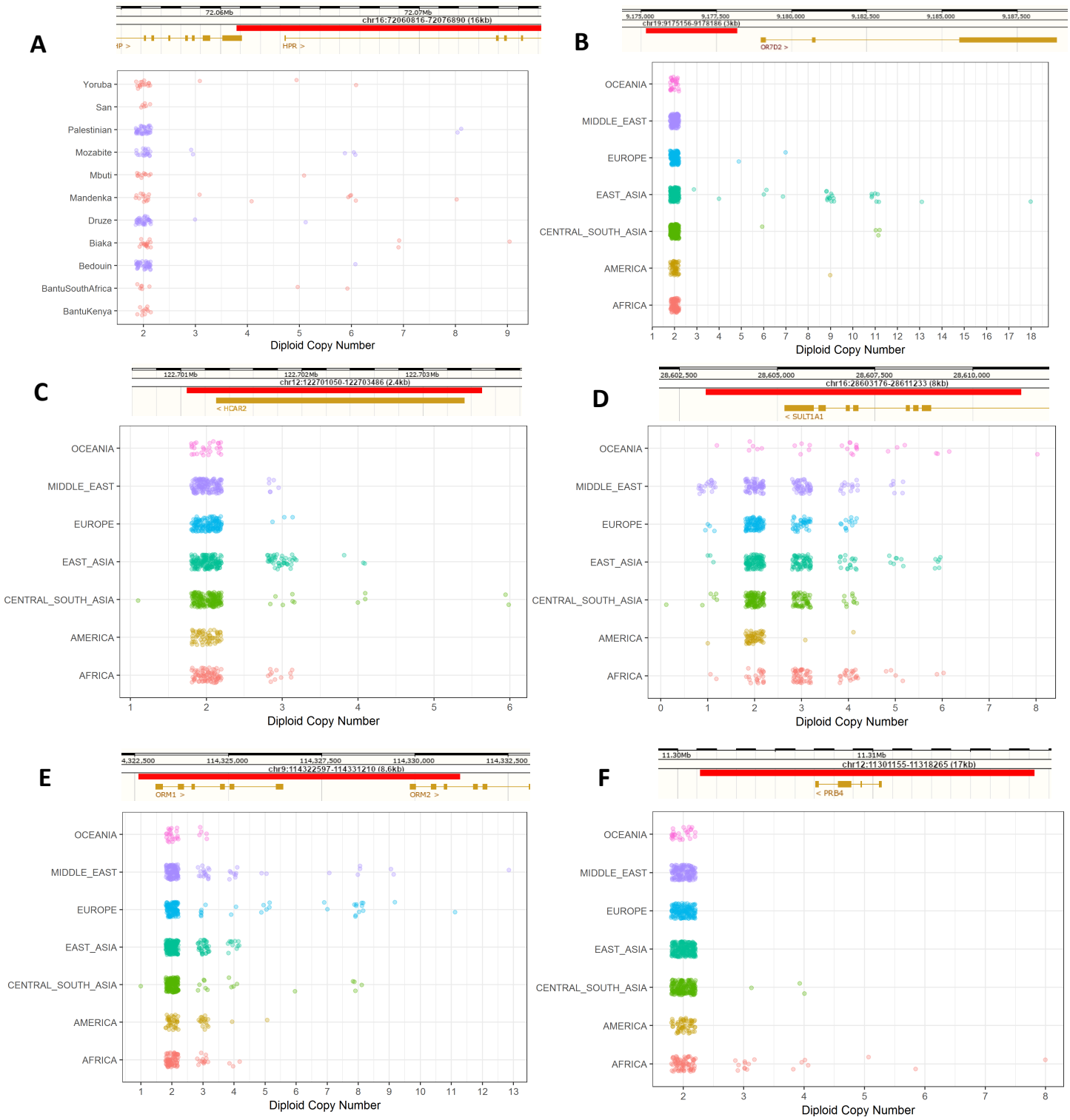

Figure 2.13: Copy Number Expansions and Runaway Duplications. Red bar illustrates the location of the expansion. Dots represent diploid integer copy number with random jitter added on the x-axis to aid visualisation. **A:** Expansion in *HPR* in Africans and Middle Easterners. **B:** Expansions upstream of *OR7D2* that are mostly restricted to East Asia. The observed expansions in Central & South Asian samples are all in Hazara samples, an admixed population carrying East Asian ancestry. **C:** Expansions within *HCAR2* which are particularly common in the Kalash population. **D:** Expansions in *SULT1A1* which are pronounced in Oceanians. **E:** Expansions in *ORM1/ORM2*. **F:** Expansions in *PRB4* which are restricted to Africa and Central & South Asian samples with significant African admixture (Makrani and Sindhi).

increase in copy number. *HCAR2* encodes HCA_2, a receptor expressed on immune cells and adipocytes. Once activated, it is involved in cellular processes that mediate anti-inflammatory effects, and as a result, has been proposed as a therapeutic target (Offermanns, 2017). Another expansion is found encompassing *SULT1A1* (Figure 2.13D), which encodes a sulfotransferase involved in metabolism of hormones and drugs (Hebbring *et al.,* 2008). All continental groups show variable copy numbers at this locus; however, the expansion is particularly pronounced in Oceanians (Figure 2.13D).

2.9 *De novo* Assemblies and Sequences Missing from the Reference

The results presented so far relied on short-read Illumina sequencing. To further explore insertions, which are difficult to identify and assemble using short-reads, we sequenced 26 samples from 13 populations (2 per population) using linked-read technology at ~50x coverage. We subsequently processed the sequences using the Supernova assembler to generate phased *de novo* assemblies (Weisenfeld *et al.,* 2017). One sample was subsequently excluded due to having substantially lower quality. By comparing these sequences to the reference genome (GRCh38), we identified 1,643 non-repetitive breakpoint-resolved insertions across all autosomes and the X-chromosome. In total, these sequences account for ~1.9 Mb not found in the reference (Figure 2.14). The sample that showed the highest number of insertions is from the San population, consistent with their known divergence from other populations. Notably, the number of insertions appears positively correlated with the quality of the assembly (r = 0.91, Contig N50 and number of identified insertions. Figure 2.14D) which suggests that many insertions remain to be identified.

I performed a PCA on the insertion genotypes and found that the populations show structure, with Central Africans and Oceanians appearing the most differentiated (Figure 2.14). This reflects the ancestral variation and deep divergences within Africa and the effect of long-term isolation, drift, and possibly private Denisovan introgression in Oceania. The majority of identified insertions are relatively small in size, with a median of around 500bp (Figure 2.14). However, large insertions of over 20kb are also found, but they are relatively rare. Ten of the insertions lie within or near exons, suggesting they have functional consequences. These genes are involved in different cellular processes, including regulation of glucose (*FGF21*), immunity (*NCF4*), and a potential tumour suppressor (*MCC*).

Many insertions appear uncommon, with 41% only appearing in one or two individuals. However, 290 insertions are found in over half of these diverse individuals, indicating that the reference potentially harbours rare deletions at these sites. To further explore this, I compared the inserted sequences with the chimpanzee, gorilla, and orangutan reference genomes (Gordon *et al.,* 2016, Kronenberg *et al.,* 2018). The majority of the identified insertions are also present in the great ape genomes, with 62% in chimpanzee, 59% in gorilla, and 35% in orangutan, values consistent with their evolutionary divergence from humans. In total, 68% of the sequences are also found in at least one great ape genome, and 33% in all three genomes. Notably, for insertions present in more than half of the individuals, 85% are also present in the chimpanzee reference. This value decreases to 18% for variants only found in two samples or less. The high number of common insertions also found in the chimpanzee genome indicates that instead of being them insertions, they are actually human-specific deletions that arose after the split from chimpanzees and are found in donors to the human reference.

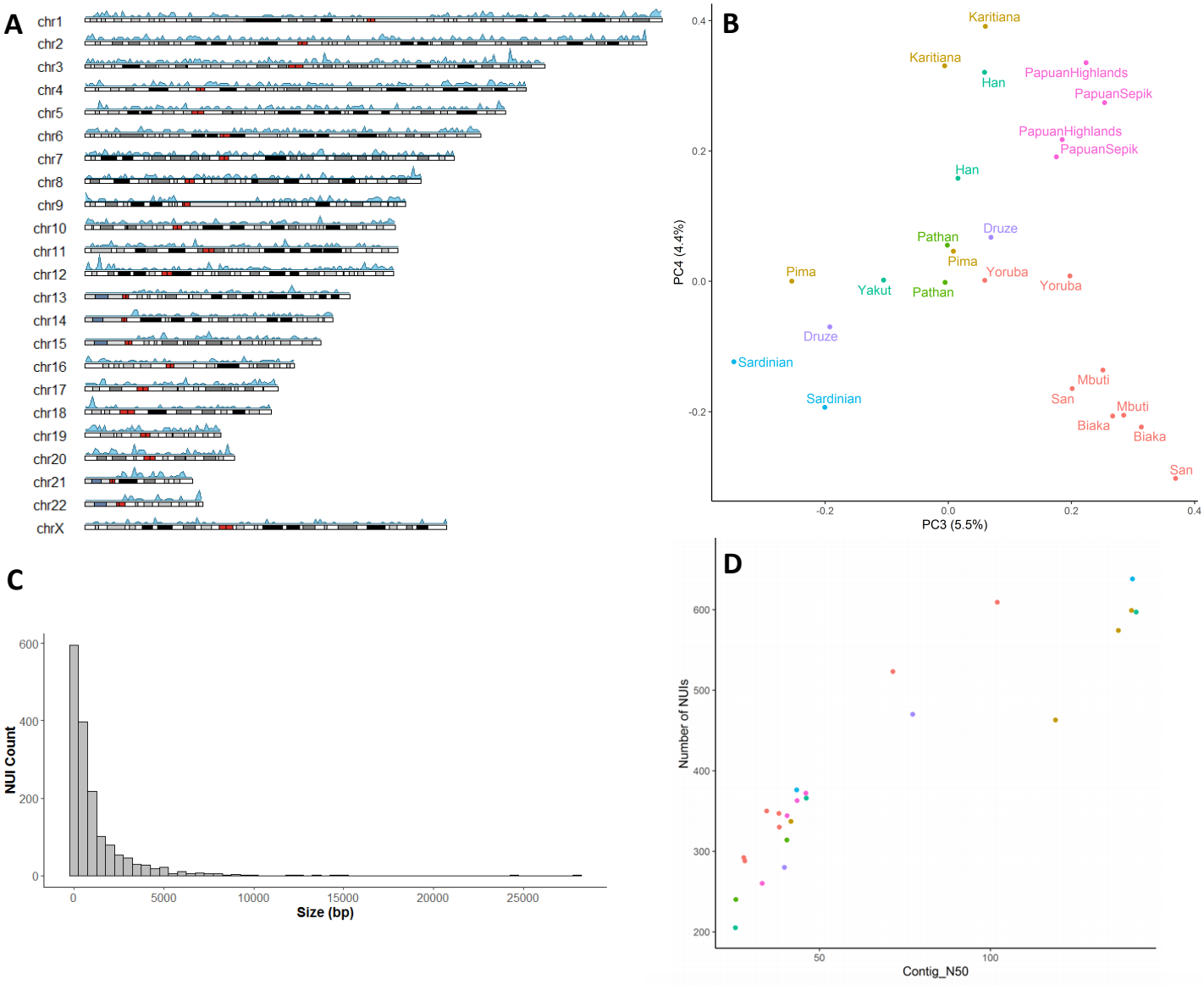

Figure 2.14: Non-Reference Unique Insertions (NUIs). **A:** Ideogram illustrating the density of identified NUI locations across different chromosomes using a window size of 1 Mb. Colours on chromosomes reflect chromosomal bands with red for centromeres. **B**: PCA of NUI genotypes showing population structure (PC3-4). Previous PCs potentially reflect variation in size and quality of the assemblies. **C**: Size distribution of NUIs using a bin size of 500bp. **D**: Positive correlation between Contig N50 and Number of identified NUIs (r = 0.91). Their relationship can be modelled by simple linear regression: y = 3x + 210 (the null hypothesis of the slope = 0 is rejected, *P* = 1.6E-10). Colours refer to the regional group of the samples

2.10 Discussion

In this study, I generated a comprehensive catalogue of SVs from one of the most diverse sets of human populations studied to date. I find that a substantial amount of genetic variation has not been documented in previous studies, with a considerable number of these SVs being common and high-frequency in regional groups and even in individual populations. This resource will be important for future medical studies, as the variants identified can be included in GWAS. SVs are important in disease susceptibility, and recent studies are identifying associations of SVs with very high effect sizes in many traits (Beyter *et al.,* 2020). As the scientific community moves to address the disparity in genomic studies by including underrepresented populations, our dataset becomes important as it catalogues variation found in diverse populations. Because of sample size, our dataset is restricted to mostly common variants. Another limitation is it still lacks adequate representation from many regions globally, such as Africa, the Americas, Oceania and Arabia. The identification of common and medically-relevant, regionally-private variants argues for further efforts in sequencing diverse genomes without data restrictions from under-represented groups.

Our resource allowed us to investigate the selective histories of SVs globally. The finding that even genetically closely-related populations can have variants with large allele frequencies differences illustrates the effect of geographically-localised selective pressures. This can happen even in geographically-nearby populations, such as in the malaria-associated *HBA2* deletion in Papuan Highlanders compared with Sepik/Lowlanders. In addition, the relatively large number of samples in each population allowed the identification of regional-specific and even population-specific variants. The latter finding in particular demonstrates that recent *de novo* mutations increased to appreciable frequencies after populations split from each other. It should be noted that the particular geographical sampling of the HGDP populations and the number of samples in each group will influence the number of private variants in this dataset. Nevertheless, we find a surprising number of high-frequency population-private variants that seem to have functional relevance. Some appear consistent with positive selection as noted above, while others may have increased in frequency due to genetic drift.

Our results demonstrate that admixture with archaic hominins has contributed potentially functional SVs to contemporary human populations. It is striking that many of the variants that have signatures of positive selection are involved in immune processes. This suggests that they may be associated with adaptation to newly-encountered pathogens after modern humans

expanded into new environments outside of Africa. Around the same time this work was posted as a pre-print and was under review, another study which focused on CNV archaic introgression was published (Hsieh *et al.,* 2019). Encouragingly, many of the putatively introgressed variants were independently replicated in the two studies. The two Oceanian variants I discussed in detail, on chr16p12 and chr8p21, were, in particular, characterized by Hsieh *et al.,* using long-read DNA and RNA sequencing. The duplication at chr16p12 was found to create a new member of the *NPIPB* gene family which has multiple amino acid substitutions indicative of positive selection. The chr8p21 variant, as speculated in this chapter, was found to generate a new *TNFRSF10D* gene which also appears to have amino acid substitutions. These results add more evidence that these variants have been under selection after archaic introgression. However, we still do not understand the selective pressures driving many of these introgressed variants, and other non-introgressed variants that seem to be under selection, and in most cases the target of selection is not understood. Experimental and association studies are needed to understand the function of these variants. It is also notable, perhaps even surprising, that despite the whole-genome sequencing of hundreds of thousands of human samples, we are still identifying new genes that are common in some populations today. This highlights the historical disparity of human genomics, and the need to study underrepresented populations to understand human history and adaptation.

By using linked-read sequencing, we generated one of the most diverse sets of phased *de novo* assemblies. Using these assemblies, we identified non-repetitive sequences that are absent from the human reference. These insertions also need to be included in future medical studies, as has been shown by a recent analysis in Icelanders that identified over 100 unique insertions in LD with a GWAS marker, and demonstrated that one insertion was associated with myocardial infarction (Kehr *et al.,* 2017). An important result of our analysis is that these insertions show population structure, demonstrating that each population harbours unique insertions missing from the reference. It is likely that even within each population, individuals also have rare and private insertions. Our findings, coupled with other recent studies that have identified megabases of such sequences, illustrates the limitation of a single human reference (Wong *et al.,* 2018, Sherman *et al.,* 2019), and the need of high-quality reference genomes from diverse human populations. It is encouraging that such efforts have already begun with the establishment of the Human Pan-genome Reference Consortium (Porubsky *et al.,* 2020).

Another limitation of our study is the use of mostly short-reads, which restricts the identification of complex SVs. Recent studies have uncovered substantially higher numbers of variants per

individual using multi-platform or long-read technologies (Audano *et al.,* 2019, Chaisson *et al.,* 2019). Although these studies were limited to very small sample sizes, with continuing decreasing costs of such methods it will become economically feasible to study thousands of populations, and such studies have begun (Beyter *et al.,* 2020). However, even with the use of, and further developments in, long-read sequencing, computational methods that can integrate the large number of SVs that will be discovered with many new diverse reference genomes need to be developed (Garrison *et al.,* 2018). Only then will the full spectrum of human structural variation can be understood.

2.11 Methods

This section provides a summary of the methods used in this chapter, more details are provided in Almarri *et al.,* 2020a and Bergstrom *et al.*, 2020.

Sample Sequencing and Quality Control

Samples were sequenced to an average coverage of 36x, minimum 25x, using Hiseq X or Hiseq 2500 Illumina instruments. All raw reads were subsequently processed using the Wellcome Sanger Institute automated sequencing pipeline and mapped to GRCh38. Coverage for each sample was visualized at ~300,000 positions across the genome and the rolling mean was plotted normalized by the genome-wide median.

Variant Calling and Quality Control

Structural variants were identified by GenomeSTRiP v2.00 (Handsaker *et al.,* 2015) and Manta v1.6 (Chen *et al.,* 2016) using default parameters. Individual sample VCFs generated by Manta were merged using svimmer using default conditions and re-genotyped across all samples using Graphtyper-v2.0 (Eggertsson *et al.,* 2019). In all callsets, variants with genotype quality < 20 were set to missing. Overlap between variants within the dataset and in comparison with other global datasets was performed using bedmap v2.4.35 (Neph *et al.,* 2012). LiftOver between genome builds was run using the online UCSC LiftOver function, and variants that failed liftover were not included in the novelty estimate. Bcftools v1.9 was used to filter and manipulate VCFs and add annotations such as excessive heterozygosity. Sequencing reads from the archaic genomes were extracted and realigned to GRCh38 using bwa aln v0.7.12 (Li & Durbin 2009) using the options -l 16500 -n 0.01 -o 2. Picard v2.6.0 was used to mark duplicates. Each archaic genome along with 30 Sanger PCR genomes were joint-called separately using GenomeSTRiP after supplying a site VCF of CNVs identified in this study. We restricted downstream analysis to archaic variant calls with CNQ ≥ 13 and manually confirmed putatively introgressed variants using IGV (Thorvaldsdóttir *et al.,* 2013).

Population Genetic Analysis

PCA was run using plink2 v2.00a2LM and v2.00a3LM including variants with MAF > 1%, missingness < 1% and pruned for LD using the option –indep-pairwise 50 5 0.2 (Chang *et al.,* 2015). UMAP was run in R-3.6.0 using the uwot package v0.1.3 using option ‘spca’, min_dist = 0.001, and n_neighbors = 16. The Variant Effect Predictor was used to identify the functional

effects of SVs (McLaren *et al.,* 2016). For the selection analysis, we calculated PBS distributions for each class of SVs and SNVs, and found all distributions to be very similar, but SNVs to be slightly more conservative, and consequently used it as a conservative null distribution. Variants were filtered for MAF > 1% and excluding > 10% missingness. A threshold of 99% of the PBS distribution was used as for evidence of departure from neutrality (i.e. top 1%). PBS was calculated using the following populations: (Oceanians; Sardinians, Han), (Karitiana; Surui,Han) and (Mbuti; Biaka, Han). The maximal variant allele frequency difference was calculated for each population pair (1431 pairwise comparisons) and compared to the average SNV differentiation (SNV F_{ST}). SNV F_{ST} was calculated for each population pair using EIGENSTRAT (Price *et al.,* 2006), on all biallelic SNPs within the accessibility mask generated in Bergström et al. (2020). Structural variant allele frequency and missingness was calculated for each population separately setting variants excluding variants with missingness > 25%. pVst from the vcflib package was used to test for significance of allele frequency differentiation using 1000 permutations. For multiallelic SNVs, we restricted the analysis to variants with CNQ ≥ 13. As this score is phred-scaled, CNQ ~13 represents ~95% confidence of diploid copy number.

Non-reference Unique Insertions

De novo assembly on linked-reads was run using Supernova v2.1.1 (Weisenfeld *et al.,* 2017). Phased BAMs and VCFs were generated using the Long Ranger v2.12 pipeline (Marks *et al.,* 2019). Non-reference unique (non-repetitive) insertions (NUIs) were identified using the NUI pipeline (Wong *et al.,* 2018), which compared each Supernova assembly to the GRCh38.p12 reference. Great ape reference genomes, chimpanzee (panTro6), gorilla (gorGor5) and Orangutan (ponAbe3), were downloaded from UCSC (Gordon *et al.,* 2016, Kronenberg *et al.,* 2018). Blastn was run to align NUI sequences to the great ape genomes, including 50 bp flanking NUI sequences, using the options –task megablast and –dust no. NUI Sequences were considered to be present in the great ape genomes if they aligned with ≥ 95% identity and 95% query coverage.

Fluorescent *in situ* Hybridization (FISH)

Melanesian lymphoblastoid cell lines (GM10543 and GM10540) were purchased from Coriell Institute for Medical Research. All FISH analysis was performed by the molecular cytogenetics team at the Sanger Institute.

Chapter 3: The Genomic History of the Middle East

This chapter has been published as a preprint (Almarri *et al.,* 2020b). I performed all the analysis presented in this chapter, except for some ancient DNA analysis (qpAdm and qpGraph) which was performed in collaboration with Dr. Marc Haber, and the Y-chromosome phylogeny which was performed in collaboration with Dr. Pille Hallast. DNA library preparation and sequencing was performed by the Wellcome Sanger Institute sequencing facility.

3.1 Introduction

The Middle East is particularly understudied by large-scale human genome sequencing projects. Geographically situated between Africa, Europe and South Asia, it forms an important region to understand human history, migrations and evolution. It is where modern humans first expanded out of Africa, where hunter-gatherers first settled and transitioned into farmers, where the first writing systems developed and where the first major known civilizations emerged. The underrepresentation of Middle Easterners in genetic and genomic studies has rendered much of the demographic history and prehistoric population movements of the region unknown. In addition, the relationships of Middle Easterners among themselves and to other global populations is unclear. The region contains some of the earliest evidence of modern humans outside Africa, with fossils dated to ~180 kya and ~85 kya identified in the Levant and North West Arabia respectively (Hershkovitz *et al.,* 2018; Groucutt *et al.,* 2018). In addition, tool kits and footprints suggesting their presence have been identified in Arabia dating to ~120 kya (Armitage *et al.,* 2011; Stewart *et al.,* 2020). Most of the limited number of studies in the region have focused on the Levant, despite Arabia, with an area of 3.2 million km^2, being larger than the Levant, Turkey and Egypt combined. The region has experienced large climatic fluctuations documented over the past 10 ky (Petraglia *et al.,* 2020). Although most of Arabia is a hyper-arid desert today, this was not always the case, as there were several late Pleistocene and Holocene humid periods resulting in a ‘green Arabia’, with the onset of the current desert climate starting around 6 kya (Petraglia *et al.,* 2020). The toggling from humid to arid periods has been proposed to result in population movements adapting to the climate. The Neolithic transition within Arabia may have developed independently within the region, or resulted from

an expansion of Levantine Neolithic farmers southwards (Drechsler, 2009; Crassard *et al.,* 2013; Hilbert *et al.,* 2015).

More recently, the genetic landscape of the region was affected by the Trans-Saharan slave trade from the 8th to 19th centuries, when millions of African individuals were captured, enslaved and traded. It has been historically suggested that individuals captured before the 16th century were Afro-Asiatic or Nilotic speakers from a region that today encompasses modern-day Ethiopia (Mirzai *et al.,* 2009). While from the 18th century, when the Omani Empire was in control of much of the African east coast, individuals enslaved are thought to be Bantu speakers from South Eastern Africa (Mirzai *et al.,* 2009). Most studies on modern-day populations in the Middle East have used mitochondrial, Y-chromosome or array-based analyses, and they have identified wide-spread, but variable, African admixture in the region (Abu-Amero *et al.,* 2008; Abu-Amero *et al.,* 2009; Haber *et al.,* 2013; Hellenthal *et al.,* 2014). Within the HDGP dataset, which includes Bedouins, Palestinians and Druze from the Levant, all populations show detectable African ancestry, highest in the Bedouins and lowest in the Druze. Cultural factors also appear to affect this pattern: in the Lebanese, studies have shown that Christian groups appear to have very limited African admixture, in contrast to much higher proportions in Muslim groups (Haber *et al.,* 2013). In Southern Arabia, in Yemenis in particular, while almost all populations harbour African ancestry, a subpopulation appear to lack African admixture (Haber *et al.,* 2019). It is important to account for African admixture within population-genetic analyses, as even small amounts can bias some tests.

Another social factor affecting the genomic landscape of the region is the widespread practice of consanguinity, among the highest in the world (Bittles and Black; 2010). A signature resulting from such practices is large blocks of homozygosity, or runs of homozygosity (ROHs). An exome-based study evaluating this in Middle Eastern and Northern African groups found that these populations can have extremely long ROHs (Scott *et al.,* 2016). This contributes to the relatively higher number of certain diseases in the region, as a rare and pathogenic variant acting in a recessive mode of inheritance can be present twice in the same individual due to consanguinity. Another issue is that consanguinity coupled with long-term genetic isolation, possibly due to cultural reasons such as religious practices (e.g. the Druze), increases genetic drift which also needs to be taken account in analysis.

In this chapter, I address the gap in the genomic research of the region by generating and analysing a high-coverage experimentally-phased open-access dataset of eight populations from the Middle East: the Arabian Peninsula, the Levant and Iraq. I note that there is no ideal name for this specific group of populations, and use “Middle East” because it is widely used and understood. I use the term ‘Arabian’ to refer to samples from the Arabian Peninsula (Emirati, Saudi and Yemeni), Levantine for Syrians and Jordanians, and Iraqi-Arabs and Iraqi-Kurds for samples from Iraq. As well as creating a catalogue of genetic variation from this understudied region that will assist future medical studies, I investigated the population structure, demographic and selective histories, and admixture events with modern and archaic humans. In particular, I use the dataset to address the following questions:

1) What is the population structure in the present-day Middle East? And how is this related to ancient populations who lived in the region?
2) Do modern Middle Easterners have traces of archeologically-documented earlier expansions out of Africa older than 60 kya ?
3) What is the demographic history of Middle Eastern populations? When did they separate and diversify from each other and from other global populations ?
4) What is the landscape of archaic introgression in the region and how does it compare to other populations?
5) Has the Neolithic revolution impacted all populations in the Middle East equally?
6) Has the spread of Semitic languages in the region left any genetic traces?
7) How did humans adapt to historical droughts and the desertification of the region?

3.2 Ethical Approval and Sample Collection

I applied and received approval from two different Research Ethical Committees to perform this study. First, I applied to the Dubai Scientific Research Ethics Committee which (DSREC-SR-02/2018_01). After receiving approval, I subsequently also applied and was approved by the Wellcome Sanger Institute Human Materials and Data Management Committee (HMDMC 18/026). All sample donors were interviewed and the aims of this research project explained, and provided informed consent. Saliva samples were collected using Oragene DNA kits (OG-600) from individuals from eight Middle Eastern populations (Figure 3.1). All populations included in this study speak Arabic, a Semitic language of the Afro-Asiatic language family, with the exception of the Iraqi Kurdish population who speak an Iranian language belonging to the

Indo-European family, Kurdish. DNA extraction, sample and variant quality control are detailed in the Methods section (3.10).

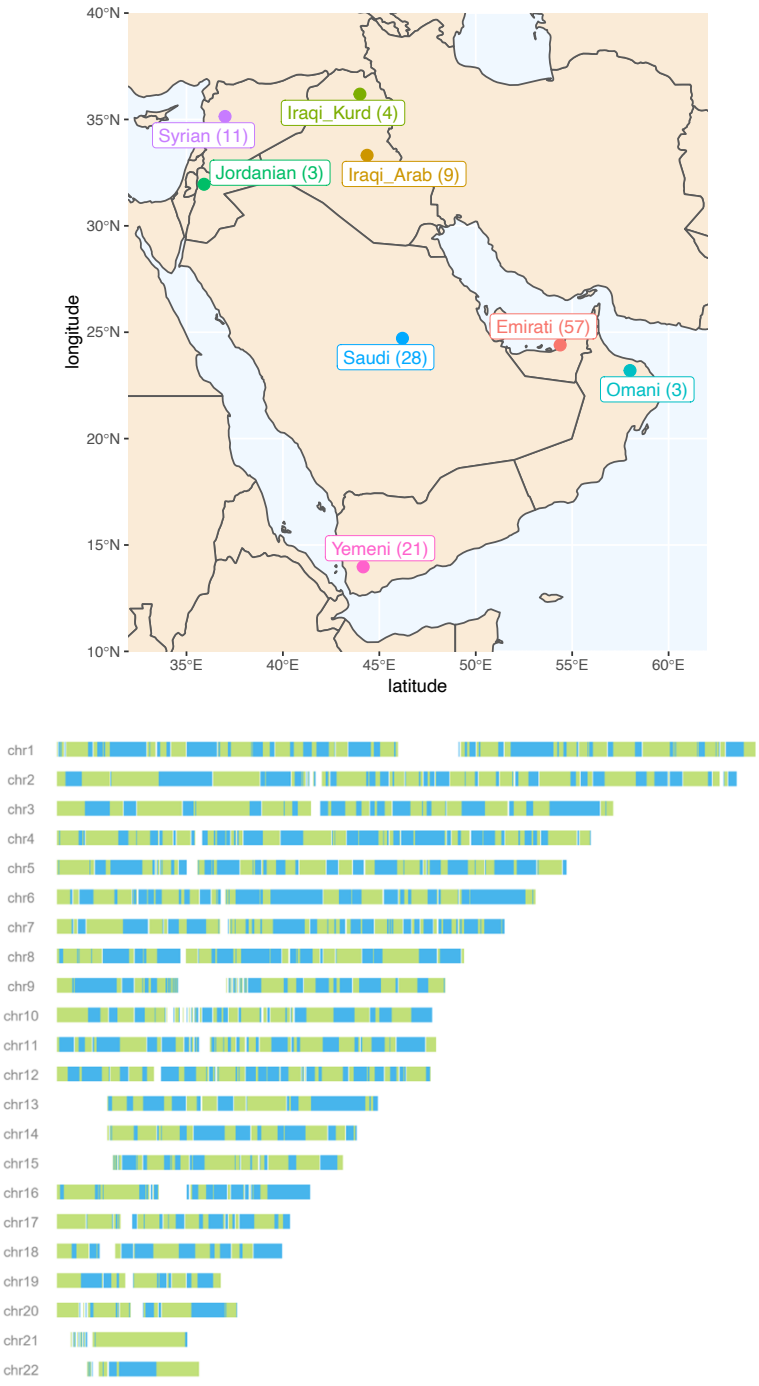

Figure 3.1: Overview of the dataset. **Top**: Map of sampled populations. Numbers in brackets refer to number of individuals samples from each population **Bottom:** Physically-phased haplotype blocks illustrated using alternating blue and green colours (sample APPG7555924). N_50 phase block: 5.2Mb. Longest phase block: 31.9 Mb

3.3 Comparison with the HGDP

I first compared the Middle Eastern dataset with the HGDP SNV callset (Bergstrom *et al.,* 2020), and found a total of 4.9 million autosomal SNVs in our dataset that are not present in the HGDP. As expected, most of the novel variants are rare (93%, MAF < 1%); however, I find ~370,000 that are common (> 1%). I subsequently evaluated whether these new variants lie within or outside the strict accessibility mask, and interestingly find that most of the novel common variants are outside the mask (66%, ~246,000). This demonstrates that although the HGDP dataset contains Middle Eastern samples from the Levant, it still does not capture all common genetic variation found in the region, highlighting the importance of studying underrepresented populations. In addition, it illustrates that, perhaps unsurprisingly, a large amount of undiscovered variation resides in regions that are not accessible to standard short-reads.

3.4 Population structure and admixture using single-variant methods.

I explored the structure and diversity of our dataset using both single-variant and haplotype-based methods. I first ran an unsupervised model-based clustering using ADMIXTURE (Alexander *et al.,* 2009); after combining our dataset with the HDGP and restricting to 1.3 million SNVs ascertained as polymorphic in archaic hominins, as they provide more accurate results from drift-based statistics (Bergstrom *et al.,* 2020). I chose K values ranging from 3 to 12, and find that K = 9 provides the lowest cross validation error. I show K = 5 and K = 9, as the former separates the whole dataset into 5 global clusters: Africans, West Eurasians, East Asians, Oceanians and Americans (Figure 3.2). At K = 5, ADMIXTURE shows that Middle Eastern populations appear to share ancestry from the component common in Europeans. It also shows that putative African ancestry is ubiquitous across Arabia and the Levant, with all population displaying such ancestry with the exception of the Iraqi Kurds. Generally, the African component appears to form a cline that is the highest in the south in Yemen which decreases moving Northwards. Within each population, variable amounts of such ancestry are detected (excluding major outliers): Yemenis (4-16%), Saudis (2-19%), Emiratis (0-20%), Iraqis (2-5%), and Syrians (1-6%), Jordanians (0.05-5%) and Iraqi-Kurds (0%). In the Arabians, Saudis and Emiratis in particular, 3 samples show very high African ancestry (>30%), likely due to recent admixture.

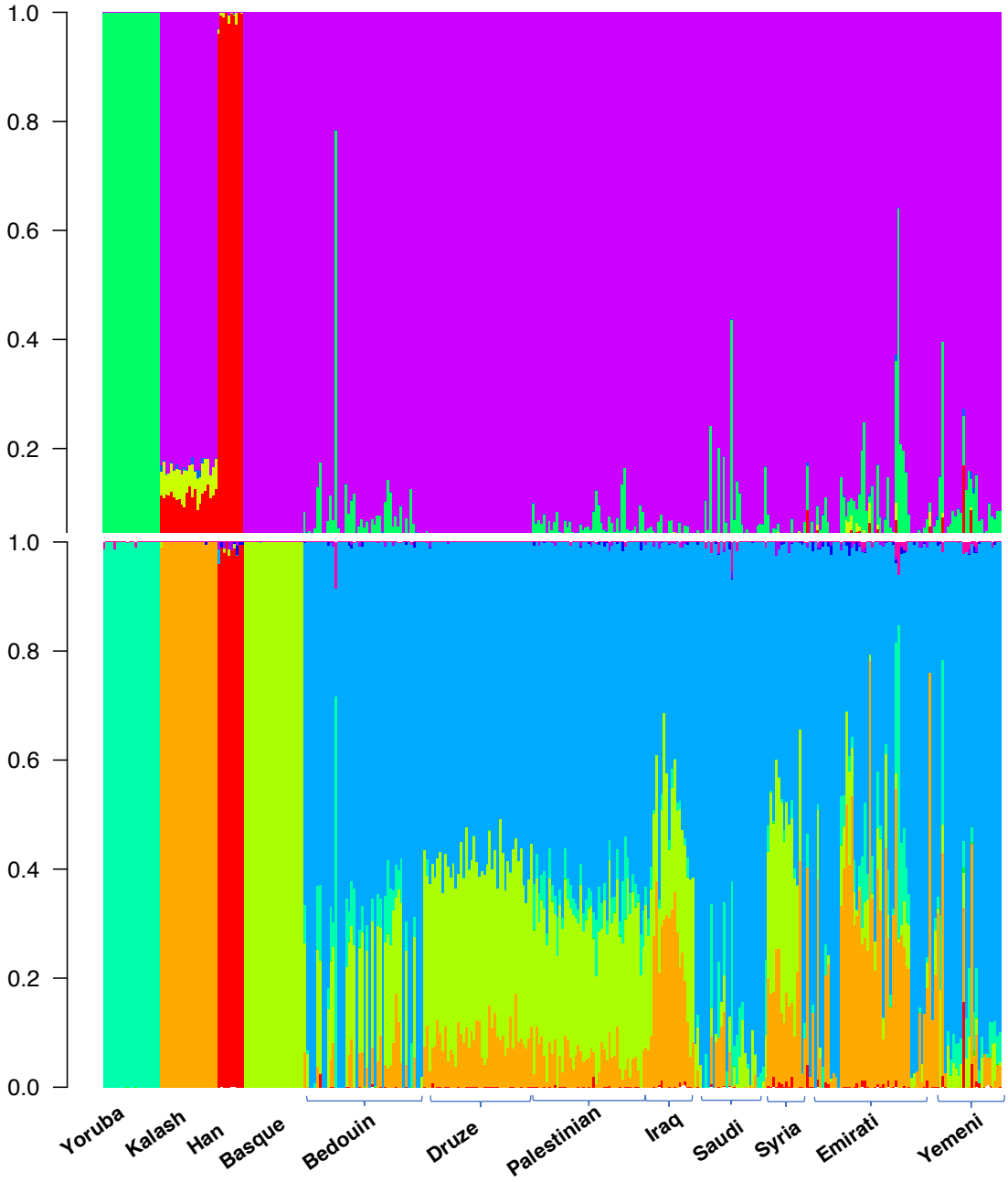

Figure 3.2: ADMIXTURE runs of our dataset with the HGDP samples. **Top:** Run setting K = 5. **Bottom:** Run setting K = 9, which has the lowest cross-validation rate. The Yoruba, Kalash, Han and Basque were included to help visualize different components.

At K = 9, a component appears that is the highest in Middle Easterners, in particular Arabians and Bedouins from the HGDP (>80%). Levantines (Druze and Palestinians) have a lower proportion of this component (40-70%) which is also present at an even lower proportion in

Southern Europeans (Sardinians, Bergamo Italians and Tuscans 4-20%). The Bedouin population from the HGDP appears to consist of two subpopulations, one with a high percentage of this Middle eastern component (many appearing ~100%), the other having percentages closer to Druze and Palestinians. A similar observation was reported using array data (Moorjani *et al.,* 2011), and we followed their approach by dividing the populations into BedouinA and B, with the latter being closer to Arabians. I also find the Emirati (UAE) population to show substructure, with a group similar to other Arabians, and other groups that appear to have ancestry related to Central & South Asians. Eastern Arabia has been previously shown to have groups with Iranian-related ancestry (Rodrigo-Flores *et al.,* 2016). At K = 9 much of the inferred African ancestry present in Middle Easterners appearing at K = 5 is reduced, with many samples also showing no such component. The next highest percentage component in Levantines (BedouinA, Palestinians, Druze and Syrians), ranging from ~20-40%, is maximal in Western Europeans (Basque and French, ~ 100%). Iraqis, both Arabs and Kurds, show a different pattern to Levantines: the European-like component is reduced (~10-30%) and the second highest component (~20-40%) is the one common in Central & South Asian populations (Kalash, ~100%).

It should be noted that the patterns displayed in ADMIXTURE are not necessarily caused by admixture, but also can result from shared ancestry. These values can also vary depending on the sample sizes of each population included in the analysis, however, it is still useful for initial exploration of the data. Another issue to note is populations that have experienced high rates of genetic drift, such ones caused by bottlenecks and long-term isolation, tend to form their own separate components, the Kalash being a notable example (Rosenberg *et al.,* 2002). If we were to make conclusions on population history solely, and naively, based on these ADMIXTURE results, it would appear that Arabians (Yemenis, Saudis and Emiratis) formed from an ancestral population, which contributed ancestry to Levantines and Iraqis, who themselves have mixtures from Europeans and Central & South Asians sources. However these conclusions would be premature and require further analysis. An issue that should be taken into consideration is whether Arabian populations, similar to the Kalash, experienced high rates of drift caused by bottlenecks which could create the patterns seen in the ADMIXTURE runs.

To measure the average differentiation between populations, I calculated pairwise F_{ST} using the Hudson estimator, as recommended by Bhatia *et al.,* 2013 (Figure 3.3). I surprisingly found high values for some pairs within the Middle East: within Iraq and the Levant, moderate

differentiation is observed, when excluding BedouinB. The Druze show the highest differentiation (F_{ST} = 0.6-0.9%) in these pairwise comparisons. When including BedouinB in the comparison, strikingly high values are observed (> 2%). It should be noted that these populations are sampled from a relatively small geographic range in the Levant, and the differentiation values are much higher than found in comparisons between some European populations separated by a greater geographic distance (e.g., HGDP French and Tuscan = 0.02%). It is clear that BedouinB is strongly differentiated from other surrounding populations, even from the other Bedouin group in the HGDP (BedouinA, 1.7%). When including Arabian populations in the comparisons (Saudis and Yemenis), BedouinB shows lower differentiation to these populations in comparison to other Levantines (~1.7%). Interestingly the Emirati population shows more differentiation to BedouinB (2%), and they also show relatively high differentiation to Saudis and Yemenis (0.8-1%). These F_{ST} values provide some context to understand the patterns seen in the ADMIXTURE runs, as it appears BedouinB and Arabian groups are highly drifted and differentiated from other populations. It also demonstrates that strong population structure exists in the Middle East, particularly between Arabia and the Levant/Iraq, and even within the two sub-regions.

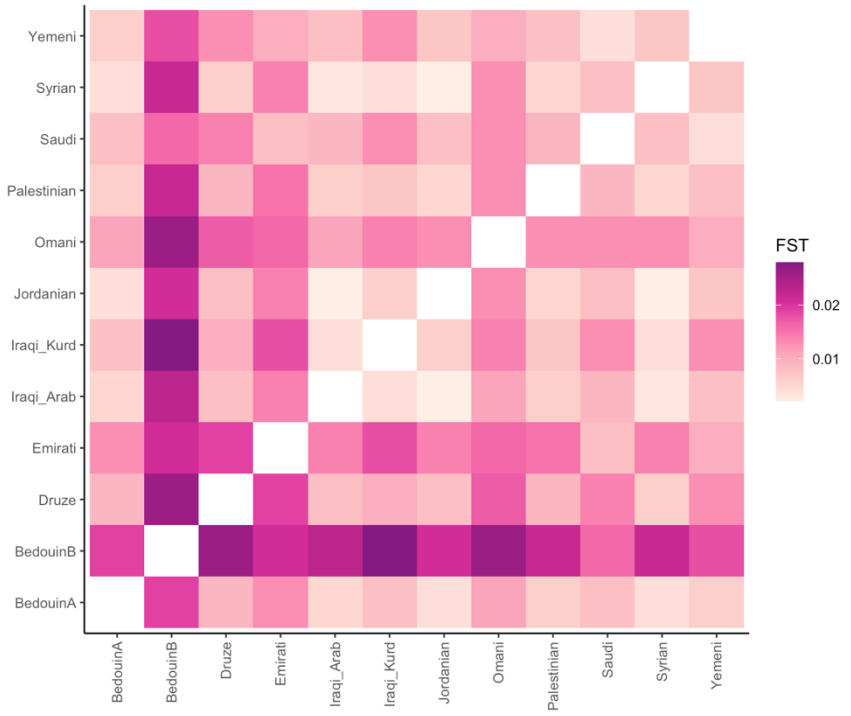

Figure 3.3: Pairwise F_{ST} estimates for Middle Eastern samples. F_{ST} calculated using 1.2M SNVs ascertained as polymorphic in archaic genomes as suggested by Bergstrom et al., 2020.

To run a formal test for the African admixture suggested in the ADMIXTURE results, I ran an f3 test on the Middle Eastern populations using f3(Middle Eastern; African, European/Central & South Asian) presented in Table 3.1. A significantly negatively result provides unequivocal evidence that the target population is admixed from the two source populations (or populations genetically related to the source populations), even if it occurred hundreds of generations ago (Patterson *et al.,* 2012). The test exploits the fact that allele frequencies of an admixed population, genome-wide, will be intermediate between the two source populations. Within Arabia, the Yemenis show significant negative f3 values ($Z < -3$) when including African and European populations, confirming the relatively high African admixture found in them in comparison to the other regional populations. For the Emiratis and Saudis however, surprisingly, no significantly negative values were identified. On the contrary, high positive values are found ($Z > 3$), especially for the Emirati population ($Z > 10$). Within the Levant, both Palestinians and BedouinA show significant admixture using African and European sources ($Z < -3$), which is not the case in Druze. The BedouinB stand out by showing very high positive f3 statistics ($Z > 20$). Although an admixture f3 test provides a robust test for admixture, non-negative values do not necessarily suggest admixture did not happen in the past. One issue that can affect this statistic is that if the target population has undergone strong drift since the admixture event, allele frequencies will fluctuate to an extent that they are no longer intermediate between the two admixing sources. This seems to be the case here, with the Emirati, Saudi and BedouinB, although showing small percentages consistent with African ancestry in ADMIXTURE, have drifted to an extent where the f3 test cannot detect admixture. An alternative way to test for admixture is exploiting admixture-induced LD (Loh *et al.,* 2013; Pickrell *et al.,* 2014). This test can also estimate the time of admixture, although with the assumption of discrete pulses of admixture. A limitation is it cannot detect admixture that occurred many generations ago due to the breakdown of LD with time through recombination. I tested for admixture using this LD-based method using the same sources (African and European/Central & South Asian) and Middle Easterners as targets. I confirmed African admixture in all Middle Eastern populations, with the exception of the Iraqi Kurds. All populations show a pulse of admixture, assuming a generation time of 29 years, at 500-1000 years ago, overlapping the Trans-Saharan slave trade period.

Reference_1	Reference_2	Target	Z-score
Yoruba	French	BedouinA	-20.604
Yoruba	Basque	BedouinB	19.514
Yoruba	Basque	Druze	3.241
Mbuti	Basque	Emirati	12.295
Yoruba	Basque	Iraqi_Arab	-8.927
Yoruba	Basque	Iraqi_Kurd	-1.83
Yoruba	Basque	Jordanian	-2.717
Yoruba	Druze	Omani	-11.371
Yoruba	Basque	Palestinian	-14.188
Yoruba	Basque	Saudi	3.227
Yoruba	Basque	Syrian	-9.404
Yoruba	Basque	Yemeni	-6.12

Table 3.1:Testing for African admixture using f3 admixture test. For Reference_1 (African), Yoruba, Mandenka, Mbuti, Biaka and South Africa Bantu, were used. For Reference_2 (Eurasian), Basque, French, Balochi, Druze, Brahui, Sindhi and Makrani were used. Most negative Z-score is presented for each target population.

I also used f4 statistics, which can be run as a symmetry test for population relationships based on the number of shared derived alleles. The test f4(Mbuti, BedouinB, [Druze/Palestinian/BedouinA]; [Saudi/Emirati]) always gives significantly positive statistics (Z > 3); indicating that BedouinB shares more alleles with Arabian populations compared to other Levantine populations. Within Arabia, Yemenis do not show any extra affinity to either Saudis or Emiratis, f4(Mbuti; Yemen; Emirati, Saudi); Z = -0.6. While Emiratis and Saudis show higher affinity to each other relative to Yemenis, f4(Mbuti; Emirati; Yemeni, Saudi); Z > 3, however, as the Yemeni population has higher African-related ancestry, this will bias this statistic and render them more distant to the other Arabian populations.

3.5 Population structure and admixture using haplotype-based methods.

I subsequently used the Chromopainter/fineSTRUCTURE pipeline to investigate population structure in more detail. fineSTRUCTURE analyses the haplotype matrix of shared segments generated by Chromopainter to distinguish samples into statistically distinct populations, and is useful in our analysis in two ways. First, it allows investigating population structure at a finer-scale and identifying sub-clusters from the same group not apparent by single-variant methods, and second, to identify representative samples that show minimal or no evidence of recent

admixture that could be used in subsequent more sensitive demographic history analyses. I first ran the pipeline using only samples within our dataset, and it generally supported the ADMIXTURE results; however, previously undetected sub-clusters of populations were identified. The tree shows that self-labelled populations generally cluster with each other (Figure 3.4). A distinction between the Levant and Arabia is identified, and the Levantine samples seem to cluster into two main subclusters. fineSTRUCTURE was generally not able to distinguish the self-labelled Levantine groups, in contrast to Arabian populations, suggesting potentially older structure and/or higher drift in Arabia.

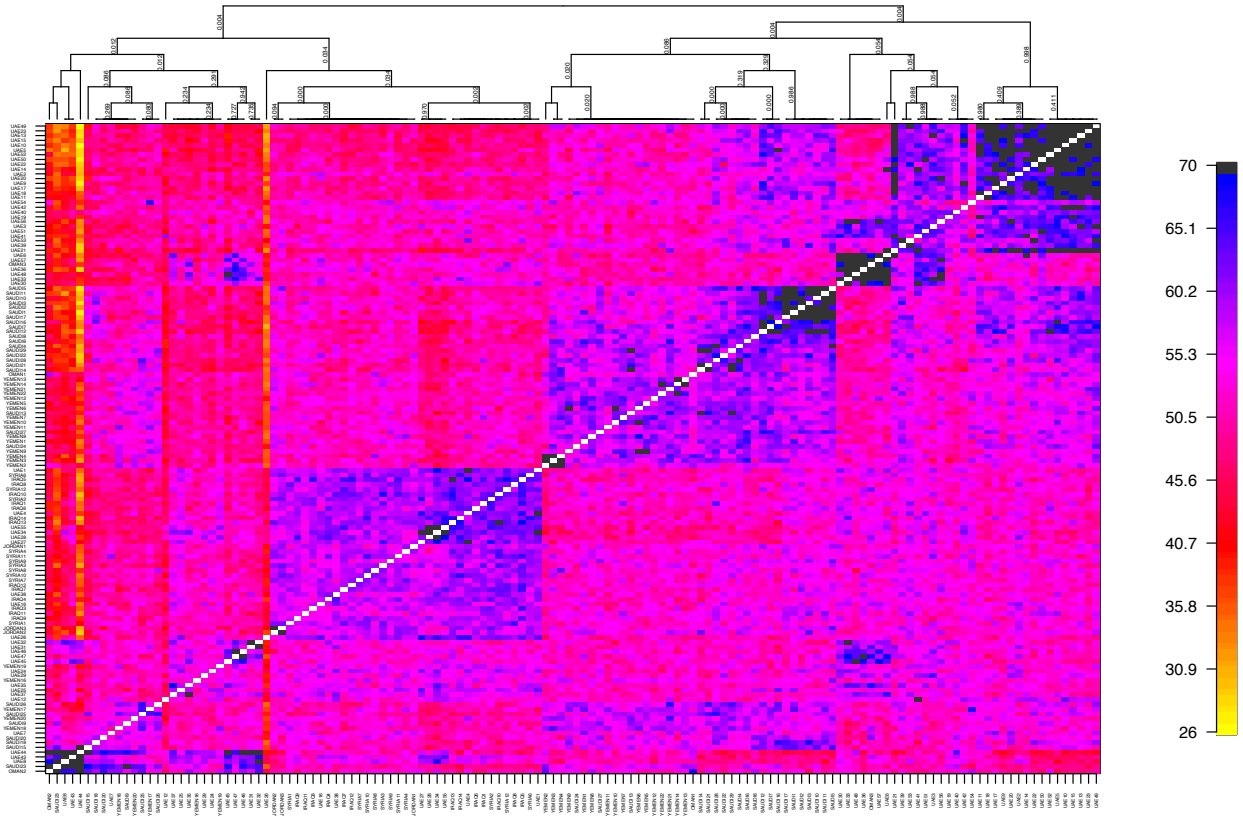

Figure 3.4: Coancestry matrix generated by the Chromopainter/fineSTRUCTURE pipeline illustrating shared haplotypic segments based on total chunklength (in cM) using 1.4 million SNVs. Numbers on tree edges represent the posterior assignment probabilities, edges with no numbers have a posterior assignment probability = 1. To increase visibility of haplotype matrix, the highest value has been capped at 70 cM.

Another result is that samples that appear to have relatively high African ancestry cluster together, even if they were from different populations. The Saudi and Emirati population showed multiple sub-clusters, while the Yemeni population showed two main clusters, with samples that have relatively lower African ancestry forming a separate cluster. The finding that Yemeni samples show little substructure is interesting, and possibly suggests recent structure, or high migration. A larger dataset sampling different regions would be required to further explore this pattern. Some clusters of Emirati samples appear closer to Levantines than to other Arabians, which is concordant with the ADMIXTURE results and potentially reflects samples with Iranian-like ancestry. At this step I separated the samples in each population into a 'core' subpopulation representative of the population and a non.core subpopulation, based on the fineSTRUCTURE and ADMIXTURE results. For the Emirati population I created 3 subpopulations, Emirati.core, Emirati.2 and Emirati.3. Emirati.core is the population that clusters with the other Arabian populations.

I subsequently ran Chromopainter/fineSTRUCTURE on a merged dataset composed of our samples with other Middle Eastern groups (Figure 3.5), in this case not just the HGDP populations, but including other groups such as Lebanese, Assyrians and different Iranian populations from published studies (Lazaridis *et al.,* 2014; Lazaridis *et al.,* 2016). A limitation is that these published studies are array-based; however, Chromopainter still performs well using a set of densely-typed markers. The fineSTRUCTURE results again show abundant structure across the Middle East, generally concordant with geography, with self-labelled groups, or related groups, generally clustering together. Groups known to be isolated such as the Druze and Assyrians form clearly-defined clusters. As with the previous analysis, populations from the Levant and Iraq (Lebanese, Syrians, Jordanians, Druze, BedouinA and Iraqi-Arabs) clustered together; however, Iraqi-Kurds clustered with Central Iranian populations, correlating with the linguistic affinity between the two populations. The Levantine BedouinB from the HGDP clustered with Arabian groups (Saudis, Emiratis, Yemenis and Omanis) in agreement with our single variant analysis.

Although single variant LD-decay methods can be used to test for and date admixture, they are unable to distinguish multiple admixing sources as only a pair of source populations are tested at a time. A set of methods that use the haplotype-sharing matrix produced by Chromopainter can exploit this information to test for and date admixture, and in addition provide proportions of multiple different ancestry sources. To perform this analysis I ran the Chromopainter pipeline with a larger number of individuals from diverse global populations,

particularly including groups that potentially admixed into Middle Easterners, such as African, European and Central & South Asian populations. The fineSTRUCTURE tree was then curated to exclude sample outliers and re-label populations that are indistinguishable to result in a total of 54 groups. To investigate potential sources of admixture, I subsequently ran SOURCEFINDv2 (Chacón-Duque *et al.,* 2018), a haplotype-based method that represents a population as a mixture of surrogates and has been shown to provide high accuracy. I divided the 54 curated populations into donor and recipient groups, with all populations from our dataset set as recipients. I performed this analysis multiple times, including setting one of our populations (Saudis or Emiratis) as a donor instead of a recipient to have a representative source from Arabia. This analysis confirmed that the Emirati cluster appearing closer to Levantines and more distant from other Arabians had modern Iranian-ancestry. Specifically, the closest source to these samples was inferred to be coastal Iranians (Iranian.Bandari). The third Emirati cluster had South Asian- and African-related ancestry. I find that the African ancestry present in all populations is from a source closest to Bantu-speakers from Kenya, with the exception of the Saudi population who also show ancestry from Nilo-Saharan-speaking Ethiopians. To date the admixture events I subsequently ran fastGLOBETROTTER, including only as surrogates populations that contributed >1% ancestry in the previous SOURCEFIND step (Figure 3.5). The only population where we don't find evidence for any recent admixture is the Iraqi-Kurdish population. Generally, the admixture dates presented by fastGLOBETROTTER are similar to those produced by ALDER/MALDER.

3.6 Modern population structure in the context of ancient populations

To further understand the population history of the region, especially the formation of modern-day populations, I also analyzed our dataset in the context of published ancient human populations from the region (Figure 3.6). A PCA including regional ancient groups shows that modern Middle Eastern samples form a cline positioned among ancient Levantine hunter-gatherers (Natufians) and Neolithic Levantines (Levant_N), Bronze Age Europeans, ancient Iranians (Iran Neolithic and Chalcolithic) and Neolithic Anatolians (Anatolia_N). This suggests that the modern populations can be potentially modelled using these ancient groups.

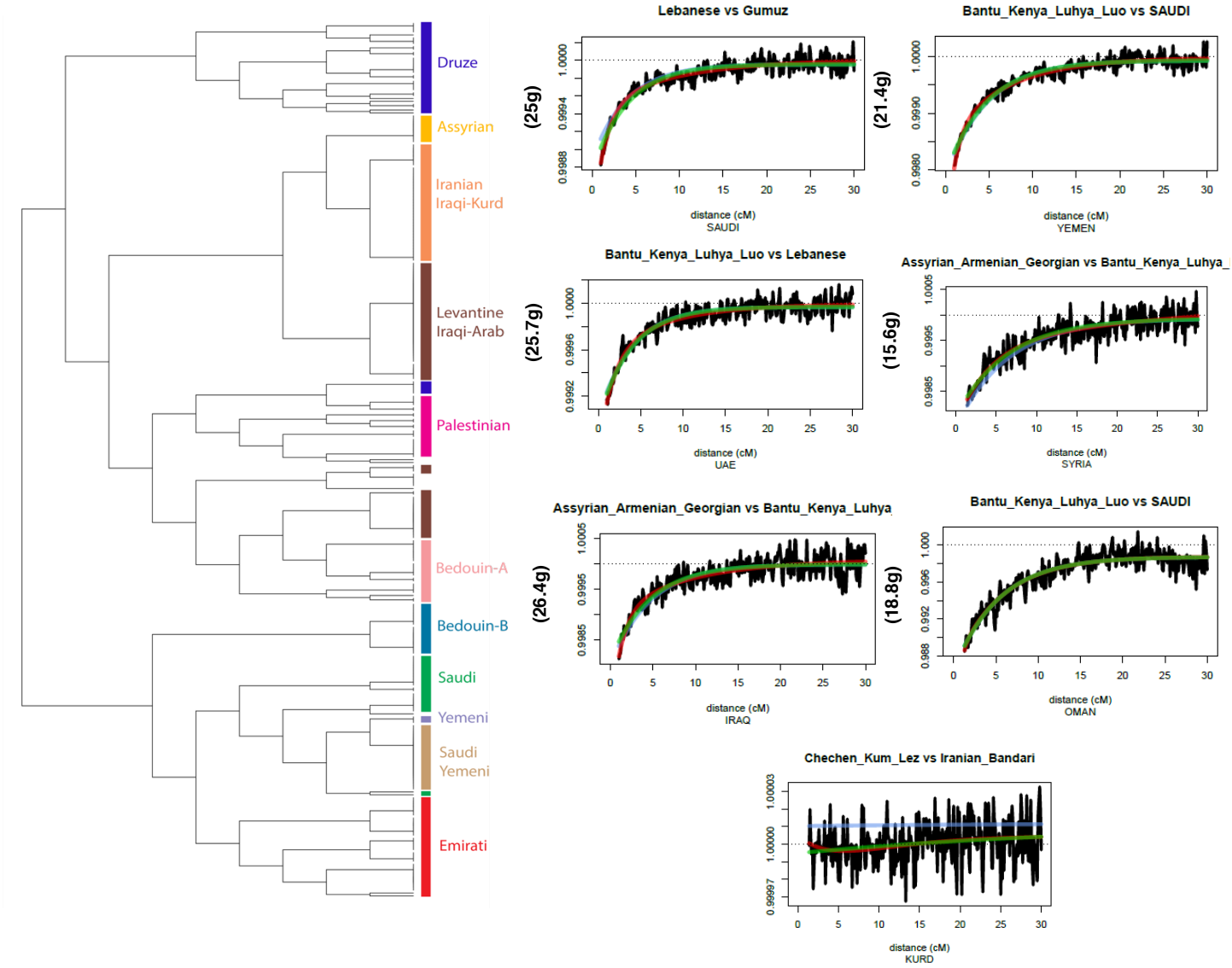

Figure 3.5: Population structure and admixture using haplotype-based methods. Left: fineSTRUCTURE tree of the modern-day Middle East population with population clusters highlighted. **Right**: Co-ancestry curves showing relative probability of jointly copying two chunks from donors at varying genetic distances. The curves fit an exponential decay (1-date green line, 2-date red line). The positive slope implies that these donors (illustrated at the top of each figure) represent admixing sources to the target (at the bottom of each figure). The estimated time of admixture is presented on the left of each figure, in generations (g).

The positioning of populations on this cline can also provide some insights into their ancestry: at the top of the cline, Arabian populations and BedouinB are located near Natufians and Levant_N. At the bottom of the cline lie the Iraqi-Arabs, Iraqi-Kurds and Assyrians, positioned

closer to ancient Iranians and Bronze Age Armenians. Positioned in the Middle of the cline are modern-day Levantines, closer to Bronze Age Europeans, and differentiated into two groups: the first composed of Palestinians, Jordanians and BedouinA, higher up the cline, while the second contains the Druze, Lebanese and Syrians lower on the cline.

We subsequently ran a temporally-aware model-based clustering on the same dataset above accounting for the time of death of the ancient samples (Joseph *et al.,* 2019, Figure 3.6 of this thesis). This approach may offer a solution to the high drift of modern populations which may have affected the standard ADMIXTURE analysis. Indeed, the results of this method provide more insights than the standard run and show that modern-day Middle Eastern populations can be modelled as deriving ancestry from Anatolian_N, Natufian/Levant_N, Iran_N and Steppe sources, similar to the PCA analysis. Differences between populations are apparent: the Arabian.core populations have very little Anatolia_N ancestry which is abundant in modern-day Levantines. This is an intriguing result, as the Levant_N population is known to share around a third of its ancestry with Anatolia_N, in comparison to the preceding Natufian hunter-gatherer groups (Lazaridis *et al.,* 2016). Consequently, a hypothesized Neolithic expansion from the Levant southwards into Arabia should have also introduced Anatolia_N ancestry in these populations, but that does not appear to be the case. The observed large difference in Anatolian_N ancestry could also be magnified due to subsequent post-Bronze Age events which introduced Eastern Hunter Gatherer (EHG) ancestry to the Levant (Haber *et al.,* 2020). Also observable in our results is that Iraqi-Arabs and Iraqi-Kurds have noticeably higher proportions of Iran_N ancestry than Levantines and other regional populations. The Iraqi-Kurds in particular have higher steppe-related ancestry than the other populations examined. This appears to be in agreement with the language spoken by the population, as in contrast to the Semitic-speaking regional populations, Iraqi-Kurds speak an Indo-European language proposed to have been introduced into the region by the movement and admixture of pastoralists from the Eurasian steppe (Haak *et al.,* 2015).

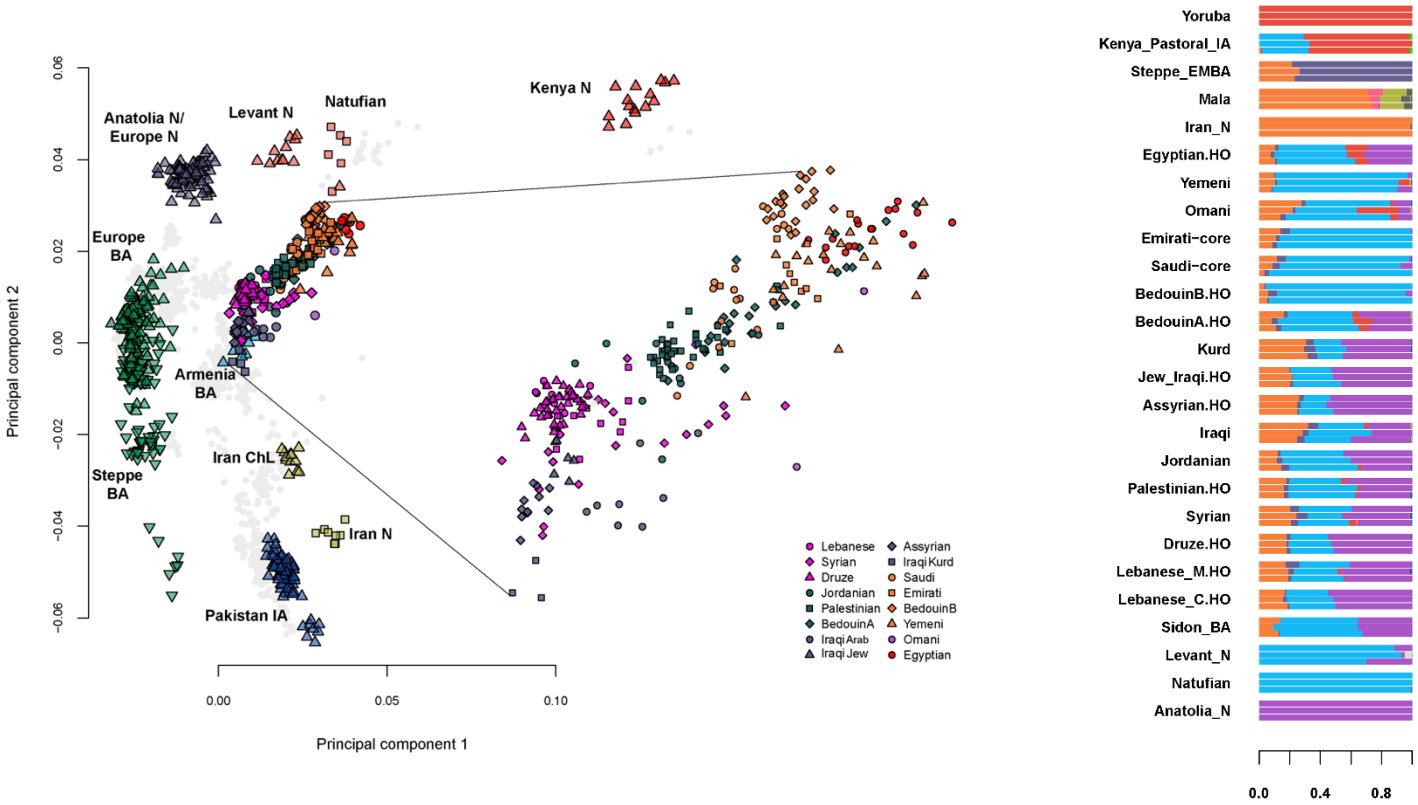

Figure 3.6: Population structure of modern samples in the context of ancient groups. Left: PCA of ancient and modern populations**.** Principal components were calculated from present-day populations and ancient samples were subsequently projected (all modern non-Middle Easterners shown as grey points). The Middle Eastern cluster is also magnified. **Right:** Temporally-aware model-based clustering using Dystruct (Joseph *et al.,* 2019) based on ~80,000 transversions and 9 time points. K=13 is presented, when the Natufian and Anatolia_N components split.

Informed by the previous results, we extended the admixture analysis by explicitly modelling modern populations as a mixture of a set of source populations using qpAdm (Patterson *et al.,* 2012; Haak *et al.,* 2015). We find that the majority of contemporary Middle Eastern populations can be successfully modelled as deriving their ancestry from four ancient populations (Table 3.2): Levant_N, Iran_N, EHG, and a ~4,500 year old East African who lacks recent Eurasian admixture (Mota; Gallego Llorente *et al.,* 2015). The results illustrate a contrast between the Levant/Iraq and Arabia: Levantines have higher EHG ancestry (~12-14%), almost double amount inferred in Arabians, which appears even higher in Iraqi Kurds (16%). Conversely, Arabians have higher Levant_N ancestry (~50%), which is much lower in Iraqis, especially Iraqi Kurds (23%). Higher African ancestry is also found in Arabian populations, although when

limiting to the 'core' populations it appears similar to Levantines (~3%). When substituting Levant_N with Natufian in the model and find that only Arabians can be successfully modelled, illustrating they can derive their local ancestry from Natufian-like hunter-gatherers without any input from Levant_N. Interestingly, none of the contemporary Levantines can be successfully modelled using Natufians instead of Levant_N.

		Ancestry proportions									Ancestry proportions								
Test	P value for rank=3	Levant_N	SE	Iran_N	SE	EHG	SE	Mota	SE	P value for rank=3	Natufian	SE	Iran_N	SE	EHG	SE	Mota	SE
Assyrian.HO	2.78E-03	0.32	0.02	0.61	0.02	0.10	0.01	-0.02	0.01	8.00E-06	0.38	0.02	0.56	0.02	0.09	0.01	-0.03	0.01
BedouinA.HO	**4.53E-01**	0.42	0.02	0.39	0.02	0.09	0.01	0.09	0.01	5.70E-04	0.48	0.02	0.36	0.02	0.09	0.01	0.08	0.01
BedouinB.HO	**7.94E-01**	0.54	0.02	0.35	0.02	0.06	0.02	0.05	0.01	2.47E-02	0.56	0.02	0.32	0.02	0.07	0.01	0.04	0.01
Druze.HO	1.86E-02	0.39	0.02	0.49	0.02	0.13	0.01	0.00	0.01	5.00E-06	0.45	0.02	0.44	0.02	0.12	0.01	-0.01	0.01
Egyptian.HO	**3.10E-01**	0.45	0.02	0.32	0.02	0.08	0.01	0.15	0.01	1.18E-03	0.50	0.02	0.30	0.02	0.08	0.01	0.12	0.01
Iraqi_Arab	**1.59E-01**	0.31	0.02	0.54	0.02	0.12	0.01	0.03	0.01	4.15E-02	0.38	0.02	0.49	0.02	0.11	0.01	0.02	0.01
Jew_Iraqi.HO	1.21E-02	0.35	0.02	0.55	0.02	0.11	0.02	-0.01	0.01	1.35E-03	0.41	0.02	0.51	0.02	0.10	0.01	-0.02	0.01
Jordanian.HO	**3.68E-01**	0.38	0.02	0.43	0.02	0.13	0.01	0.06	0.01	4.09E-03	0.46	0.02	0.38	0.02	0.11	0.01	0.05	0.01
Jordanian	**1.04E-01**	0.43	0.03	0.44	0.03	0.11	0.02	0.02	0.01	5.41E-03	0.48	0.02	0.40	0.03	0.11	0.02	0.01	0.01
Iraqi_Kurd	**6.97E-02**	0.23	0.02	0.62	0.02	0.16	0.02	-0.01	0.01	2.01E-02	0.31	0.02	0.56	0.03	0.15	0.02	-0.01	0.01
Lebanese_Christian.HO	2.01E-02	0.42	0.02	0.46	0.02	0.13	0.01	-0.01	0.01	9.00E-06	0.49	0.02	0.41	0.02	0.12	0.01	-0.02	0.01
Lebanese_Muslim.HO	**1.23E-01**	0.39	0.02	0.48	0.02	0.11	0.01	0.02	0.01	2.95E-04	0.45	0.02	0.44	0.02	0.11	0.01	0.01	0.01
Omani	**2.95E-01**	0.41	0.03	0.41	0.03	0.09	0.02	0.10	0.01	4.20E-02	0.46	0.02	0.37	0.03	0.09	0.02	0.09	0.01
Palestinian.HO	**5.48E-02**	0.40	0.02	0.43	0.02	0.11	0.01	0.06	0.01	2.02E-04	0.47	0.02	0.39	0.02	0.10	0.01	0.04	0.01
Saudi.core	**1.09E-01**	0.49	0.02	0.42	0.02	0.06	0.01	0.03	0.01	**9.08E-02**	0.52	0.02	0.39	0.02	0.07	0.01	0.02	0.01
Saudi	**3.07E-01**	0.50	0.02	0.32	0.02	0.05	0.01	0.14	0.01	1.25E-02	0.50	0.02	0.31	0.02	0.07	0.01	0.12	0.01
Saudi.HO	**2.83E-01**	0.50	0.02	0.40	0.02	0.07	0.02	0.04	0.01	**1.42E-01**	0.51	0.02	0.38	0.02	0.09	0.01	0.03	0.01
Syrian	**1.81E-01**	0.34	0.02	0.50	0.02	0.14	0.01	0.02	0.01	1.12E-02	0.40	0.02	0.46	0.02	0.13	0.01	0.01	0.01
Syrian.HO	**6.62E-02**	0.38	0.02	0.45	0.02	0.12	0.01	0.05	0.01	9.50E-05	0.44	0.02	0.41	0.02	0.11	0.01	0.04	0.01
Emirati.core	2.02E-02	0.49	0.02	0.43	0.02	0.06	0.01	0.03	0.01	**2.30E-01**	0.53	0.02	0.39	0.02	0.07	0.01	0.02	0.01
Emirati	2.00E-06	0.29	0.02	0.52	0.02	0.09	0.01	0.10	0.00	1.11E-03	0.34	0.02	0.48	0.02	0.09	0.01	0.09	0.00
Yemeni	3.69E-02	0.52	0.02	0.35	0.02	0.04	0.01	0.09	0.01	3.09E-02	0.55	0.02	0.32	0.02	0.06	0.01	0.08	0.01
Yemeni.HO	3.77E-02	0.38	0.02	0.40	0.02	0.06	0.01	0.16	0.01	**9.76E-02**	0.42	0.02	0.37	0.02	0.07	0.01	0.14	0.01

Table 3.2: Modelling present-day Middle Easterners as deriving their ancestry from four ancient populations. Using qpAdm we set seven outgroup: Ust'-Ishim, Kostenki14, WHG, CHG, Natufian (or Levant_N), Papuan, and Mbuti. SE = standard error. P value > 0.05 (bold) indicates the model is not rejected. ".HO" refers to samples from the Human Origins dataset. '.core' represents the curated samples, samples without '.core' represent the general population.

Another ancestry that is found in all Middle Eastern populations is from the Iranian Neolithic source. Previous ancient DNA studies have shown that this ancestry was not present in the region during the Neolithic period, but appears in the Bronze Age (Lazaridis *et al.,* 2016). This source was shown to replace around 50% of the ancestry in the region, and we confirm this in our analysis (Table 3.2). Such a large turnover of ancestry motivated us to explore the time of its spread. Using admixture-induced LD we tested whether or not this source has entered the region at a similar time in different populations. The results indicate that the admixture occurred following a generally North to South cline (Figure 3.7). The oldest admixture dates appear in the Levant (3,900-5,600 years ago (ya)), followed by Egypt (2,900-4,700 ya), East Africa (2,200-

3,300 ya) and Arabia (2,000-3,800 ya). Intriguingly, these admixture dates overlap with the dates suggested for the Bronze Age origin and spread of the Semitic languages estimated from linguistic data (Kitchen *et al.,* 2009; Figure 3.7 of this thesis). The Y-chromosome haplogroup J1 is present at the highest frequency in the Middle East globally, especially Arabia, and its origin has been hypothesized to be from the Zagros/Taurus mountain region, suggesting a population movement southwards (Chiaroni *et al.,* 2010; Lazaridis *et al.,* 2016). In addition, the published ancient Iran_N samples Y-chromosomes are mostly J1. As the Y-chromosome can provide independent, although from a single male-specific lineage, evidence towards understanding the history of the region, we created a Y-chromosome phylogeny of our samples with global populations (1333 samples). The phylogeny demonstrates that the majority of the J1 chromosomes in the Middle East coalesce around ~5.6 [95% CI, 4.8-6.5] kya, in agreement with a potential Bronze Age expansion (Figure 3.8). However, rarer earlier diverged lineages are also present coalescing ~17 kya. The Y chromosome haplogroup found in Natufians, E1b1b, is also common in our dataset and the majority of lineages coalesce ~8.3 [7-9.7] kya, though we also find a rare deeply divergent E1b1b Y-chromosome coalescing ~39 kya.

We subsequently attempted to create graph models of admixture history informed by the results above and by previous literature. We used two modern-day populations from the region to simplify the models, Lebanese Christians as a representative of the Levant and the Emirati.core of Arabia. Of the models we tested, one seems to have a good fit to the data (worst f-statistics Z < |3|). The best fit (worst f-statistic Z = -2.9, Figure 3.9), shows Lebanese forming from the Bronze Age Sidon (89%) population, with an additional contribution of Steppe-like ancestry (11%), consistent with the literature (Haber *et al.,* 2017). While Emirati.core, instead, form from a Natufian-like (25%) and Sidon-like population (75%). Other models explored, with Emiratis descending from Iran_N and Natufians, or from Levant_N and Iran_N show poorer fits (worst f-statistic Z = -3.7 and -7.1, respectively)

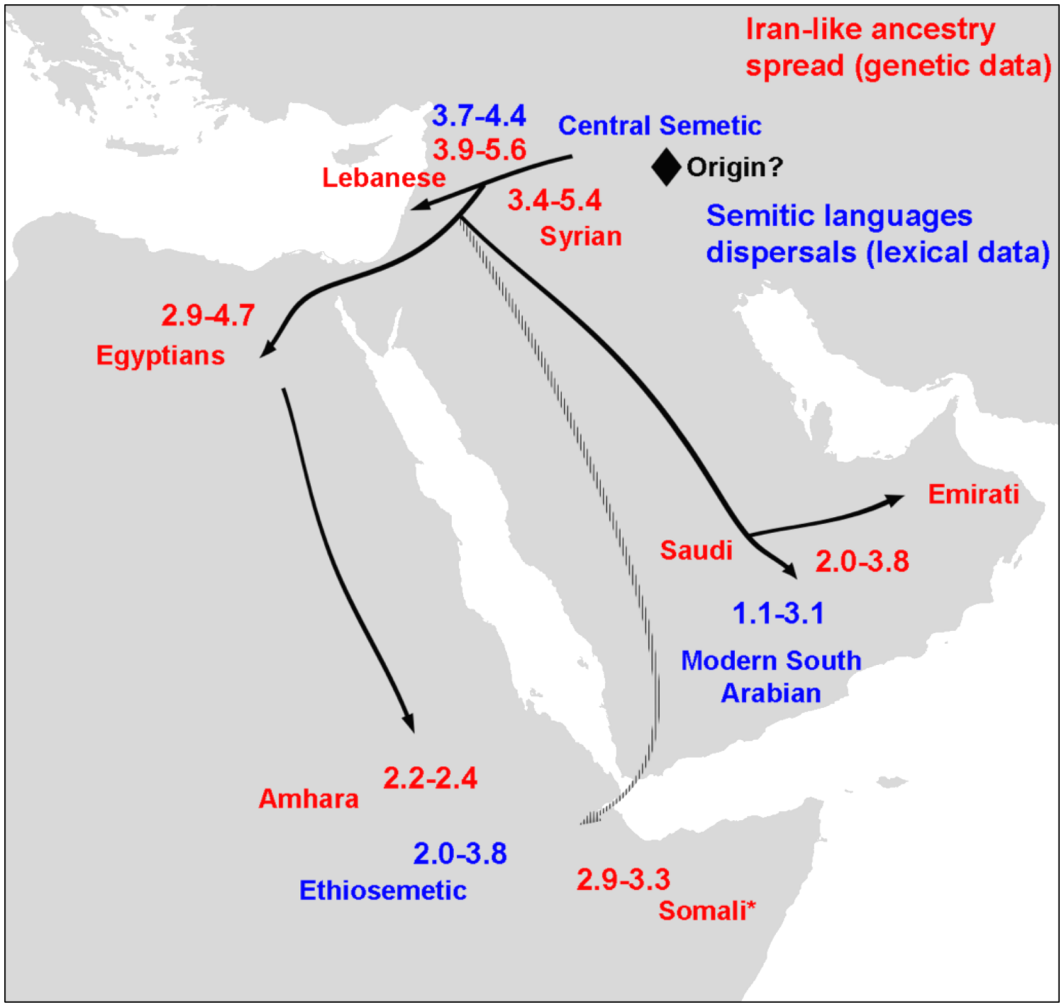

Figure 3.7: Correlation between the spread of Semitic languages and Iranian-like ancestry. Dates in thousands of years ago, (red) are based on our admixture analysis and Semitic languages dispersal dates were estimated by Kitchen et al. 2009 from lexical data (blue). Kitchen et al. estimated an Early Bronze Age origin for Semitic ~5.7 kya in the Levant. Admixture also appears in non-Semitic, Cushitic-speaking populations such as the Somalis. Kitchen et al. suggested that Semitic languages would have spread into East Africa with little gene flow, as Ethiosemitic-speaking populations share similar proportions of non-African ancestry and are genetically similar to Cushitic-speaking populations (Pagani et al., 2012). They proposed that the current distribution of Ethiosemitic languages reflect a language diffusion process through African populations, rather than gene flow.

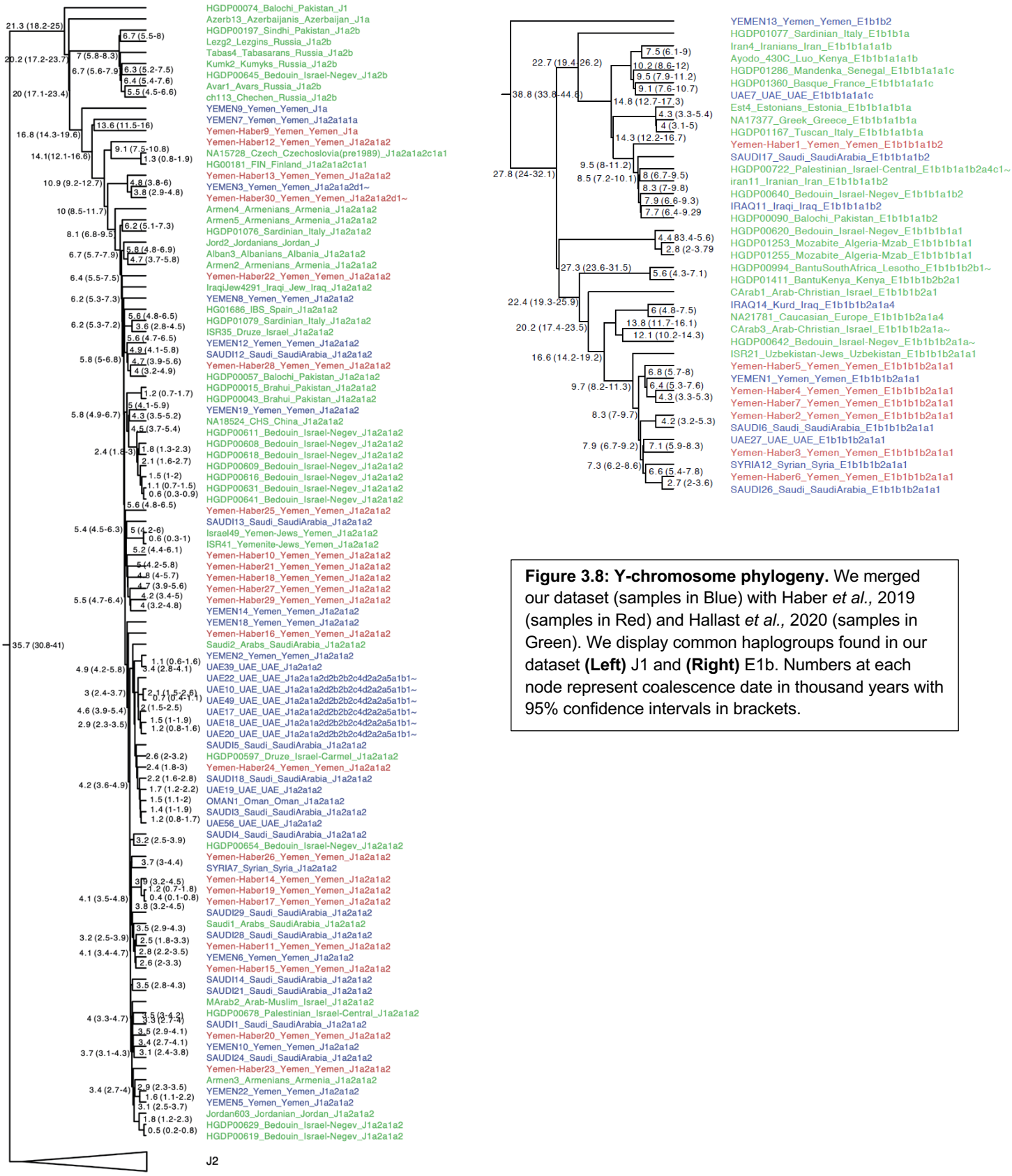

Figure 3.8: Y-chromosome phylogeny. We merged our dataset (samples in Blue) with Haber *et al.,* 2019 (samples in Red) and Hallast *et al.,* 2020 (samples in Green). We display common haplogroups found in our dataset **(Left)** J1 and **(Right)** E1b. Numbers at each node represent coalescence date in thousand years with 95% confidence intervals in brackets.

3.7 Effective population size and separation history

A major advantage of our dataset is that all samples are physically-phased using linked-read sequencing. This large number of accurately phased haplotypes can be exploited using coalescent-based methods to study population history from very old periods (>250 kya) to very recent periods (1 kya). The resolution at recent periods is of particular interest since they overlap historical and archeologically-documented events. Applying a new method, Relate (Speidel *et al.,* 2019), that generates genome-wide genealogies from the supplied haplotypes shows that Middle Easterners display a significant decrease in population size around the out-of-Africa event ~50-70 kya, typical of non-African populations (Figure 3.10). A recovery from this bottleneck follows a similar trajectory for all populations until around 15-20 kya, when Arabian and Levantine populations begin to diverge in size. All Arabian population maintain similar sizes, while Levantine and Iraqi groups continue to show a substantial population expansion. This contrast is intriguing as it begins after the end of the Last Glacial Maximum and is prominent during the Neolithic period, when agriculture developed in the region and resulted in settled societies supporting much larger populations. After this period, and around the beginning of the archeologically-documented aridification of Arabia ~6ya (Petraglia *et al.,* 2020), the Arabian populations appear to experience a bottleneck while Levantine groups continue their expansion. Around the 4.2ky aridification event, Levantine populations begin to plateau in size and subsequently decrease. Among the Arabian populations, the bottleneck in the Emirati group is particularly prominent, with an inferred effective population size of around ~5k, substantially smaller than Levantines during the same period (> 100k).

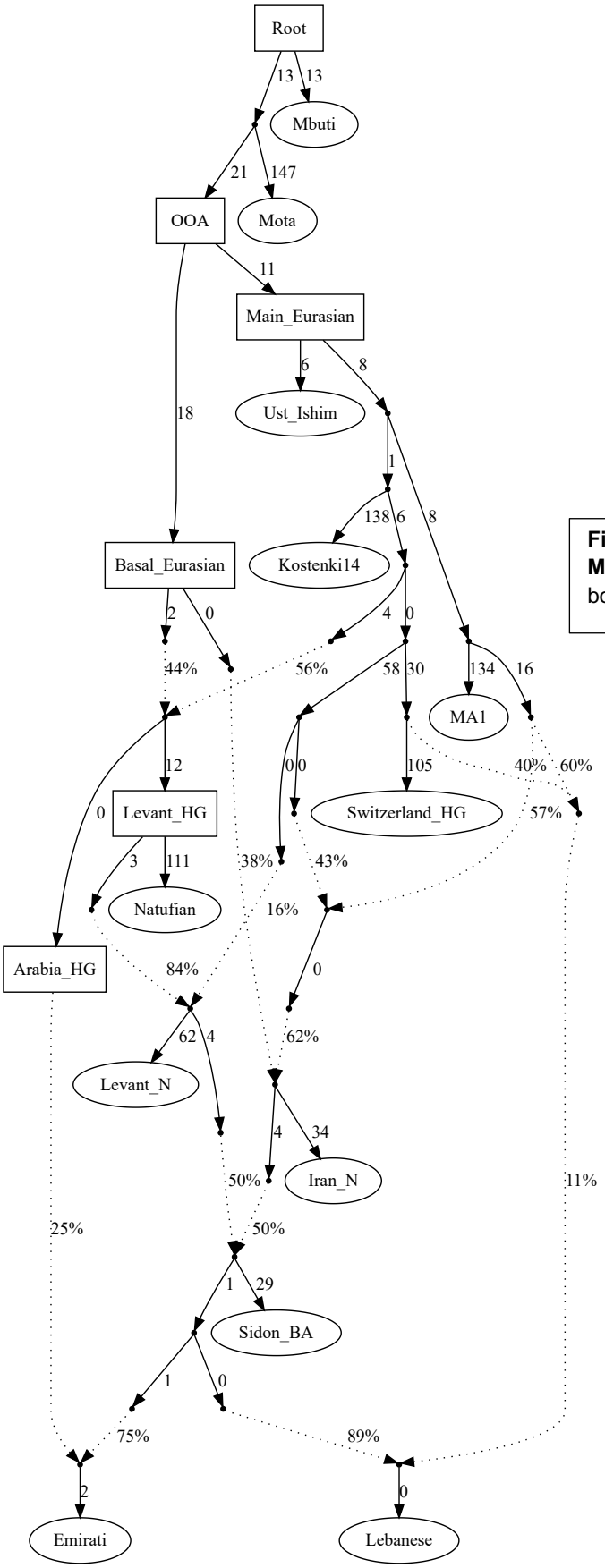

Figure 3.9: A possible model for population formation in the Middle East. Populations in ellipses are sampled, while those in boxes are hypothetical. Worst f-statistics: Z score = -2.9.

I was concerned about the potential effect of recent consanguinity on the estimated population size history, especially at the recent bottlenecks associated with climatic events. I calculated ROHs for each sample and found that some Arabian samples in particular have large ROHs, which is consistent with recent consanguinity. When I first ran the demographic history analysis I did not take into account such factors, and found that some of the curated samples chosen for this analysis had large ROHs. Using a threshold of 50 Mb of total ROHs of at least 1 Mb in size, I repeated the analysis three times. The first run I included some samples with >50Mb total ROH, the second all samples had < 50Mb, the third I only included one haplotype per sample. The latter test, even in highly consanguineous individuals, will remove the effect of recent consanguinity. Some differences are detected in the three runs (Figure 3.10): including samples with >50Mb ROHs leads to a more pronounced reduction in the past 4ky, while a modest recovery is seen in the last 1ky. Samples with <50 Mb ROHs show a similar history, but the recent bottleneck is smaller in magnitude while a stronger recovery is identified. The single haplotype analysis shows the recent recovery starts at 2 kya and is much stronger than the previous runs. The second bottleneck is observed in all three runs, suggesting it is not an artefact of recent consanguinity.

I subsequently analysed the separation history of populations within our dataset among themselves and in comparison to other global populations (Figure 3.11). Accurate phasing is crucial for this type of analysis, as demonstrated by a previous study that suggested that contemporary Papuans have traces of an earlier expansion out of Africa (Pagani *et al.,* 2016), while our HGDP SNV study, which repeated this analysis using physically-phased samples, did not replicate this finding and suggested that it was due to an artefact of statistical phasing (Bergstrom *et al.,* 2020). Physical-phasing is also crucial when attempting to explore population separation history at very recent times as rare variants become more important but are less accurately phased using statistical approaches, and are also unlikely to be present in reference panels. I first explored whether contemporary Middle Eastern populations have any detectable ancestry from an early expansion out of Africa by comparing the separation times against populations with physical-phasing from the HGDP (Bergstrom *et al.,* 2020). I used a relative cross-coalescent rate (rCCR) of 0.5 as an estimate of split time and find that all populations tested, Levantines, Arabians, Iraqis, Sardinians and Han Chinese, share the same separation history, and additionally the same gradual and complicated pattern of separation from Mbuti, around 120 kya. I subsequently compared the Middle Eastern populations with Sardinians and find that they all seem to separate ~20 kya, with Arabians exhibiting a slightly earlier divergence

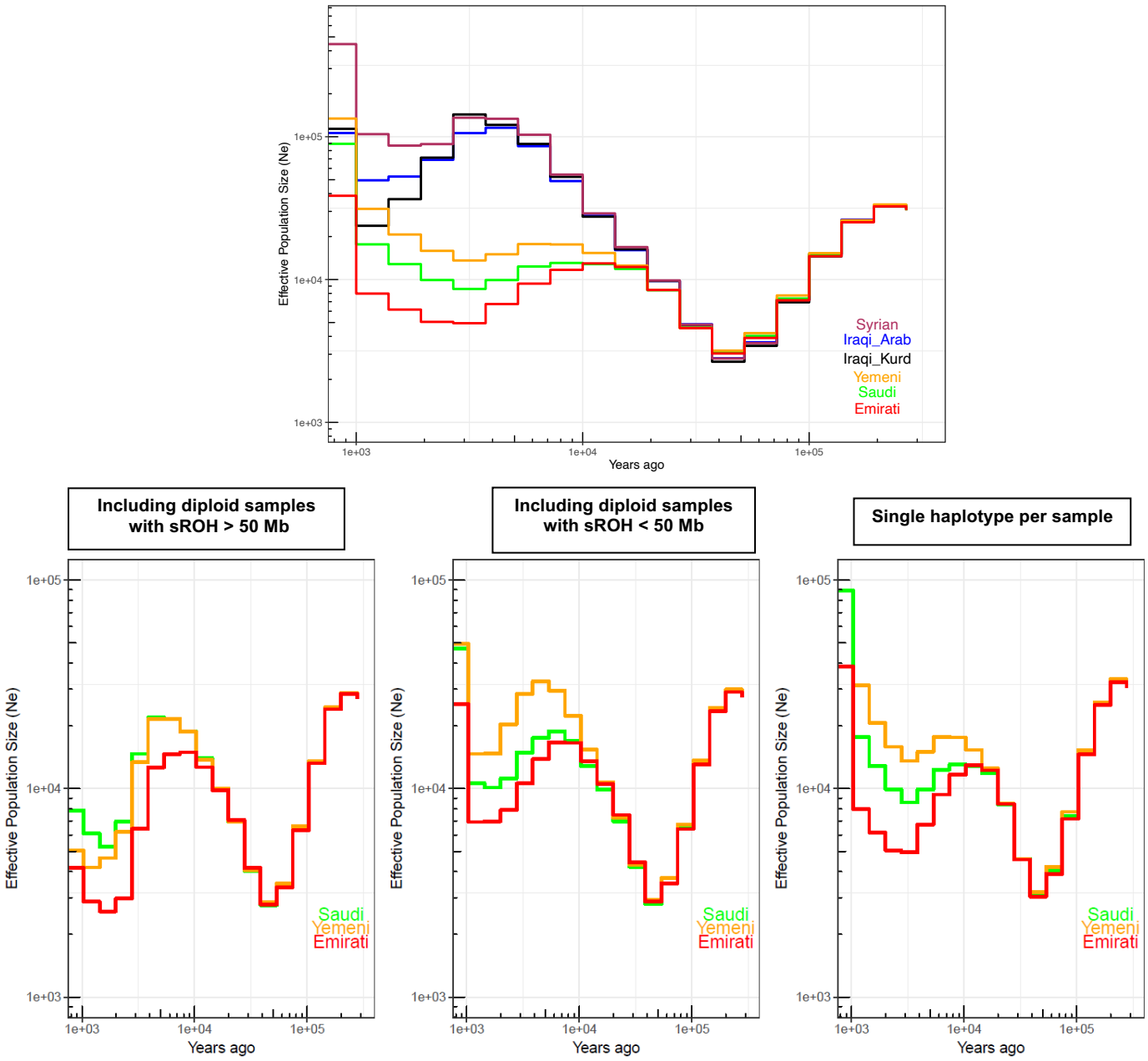

Figure 3.10: Effective population size histories for Middle Eastern populations. Top: Estimates using a single haplotype per sample. **Bottom:** Testing for the effect of recent consanguinity on population size histories in Arabian groups.

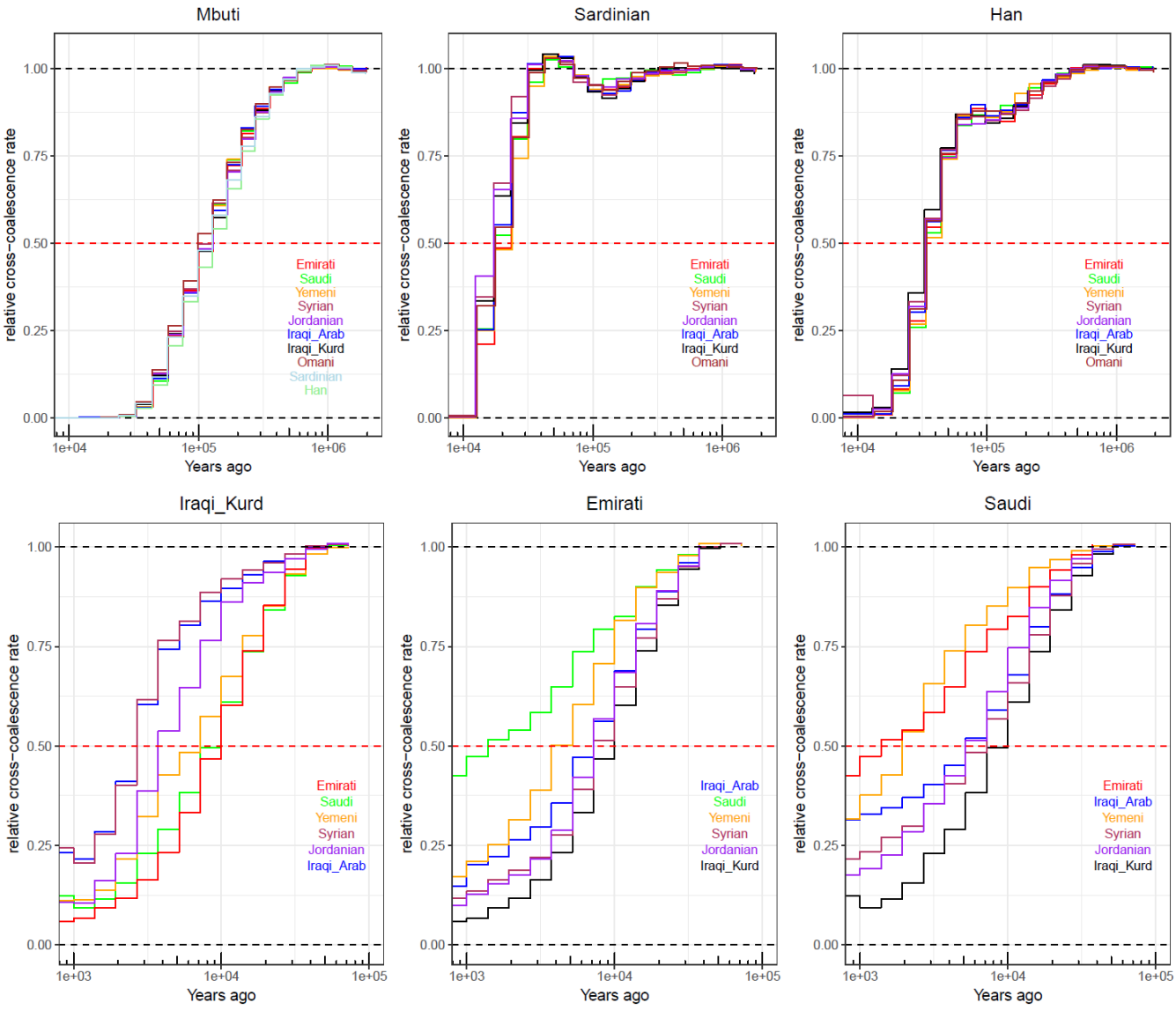

Figure 3.11: Separation history analysis. **Top:** Coalescent-based separation history between Mbuti, Sardinians and Han (indicated at the top of each panel) and each Middle Eastern group (identified within each panel). All Middle Eastern populations show a similar split time with each of these global populations. **Bottom:** Separation history within the Middle East (population indicated at the top of each panel, and within each panel). Note the different X-axis scales between the top and bottom sections.

time in contrast to Levantines. However in contrast to the gradual separation curves found in the comparison to Mbuti, Sardinians show an almost clean split from Middle Easterners. An important result in this analysis is all lineages within Arabia and the Levant, in addition to all lineages within Middle Easterners and Sardinians, coalesce within 40 kya (rCCR = 1). This indicates that contemporary populations do not have any detectable traces of earlier expansions out of Africa and descend from the same population that expanded out of the continent ~50-60 kya.

I subsequently compared the split times of the populations within the Middle East. The oldest separation times were between the Levant/Iraq and Arabia (Figure 3.11). The population appearing to diverge the first from Arabians was the Iraqi-Kurdish population. The Emiratis split from Iraqi-Kurds ~10 kya, and at a more recent time to Syrians, Jordanians and Iraqi Arabs (~7 kya). The Saudi separation curves from the same populations indicate a more recent divergence, 5-7 kya, while Yemenis appear as an intermediate between the Saudi and Emirati curves. These split times suggest that Arabian and Levantine populations separated before the Bronze Age, indicating any Bronze age expansion into Arabia from the north, if it indeed occurred, did not result in a complete replacement of local ancestry. Within Arabia, Yemenis appear to split from Emiratis ~3 kya, while Saudis appear to split more recently from both populations (<2 kya). Within the Levant and Iraq, all populations have separated from each other within the past 3-4ky. The separation curves of populations within the region appear gradual, suggesting possible ongoing gene flow after separation rather than clean splits. The separation curves also reflect the admixture histories of these populations.

3.8 Archaic introgression and deep ancestry in the Middle East

Previous studies have reported that Middle Eastern populations have lower Neanderthal ancestry in comparison to European and East Asian populations (Rodriguez-Flores *et al.,* 2016; Bergström *et al.,* 2020). However the interpretation of this is result is complicated by recent African admixture which will decrease Neanderthal ancestry. Moreover, many analyses require the use of an outgroup, in which African populations are commonly used. However, if the outgroup itself contains even small amounts of Neanderthal ancestry, for example due to back-to-Africa migrations introducing Eurasian ancestry, this will introduce bias into the analysis (Chen *et al.,* 2020). To explore the landscape of Neanderthal introgression on our dataset I relied on different analyses: first, I exploited the accurate phasing and compared the rCCR of

our samples with the high coverage Vindija Neanderthal genome (Prufer *et al*., 2017). The rCCR decreases to < 0.01 at round 200 kya for all populations we tested, including Africans (Figure 3.12A). However, in more recent time segments, a small increase in rCCR reaching 0.02-0.03 is only found in Eurasian groups, including Middle Easterners. The pattern reaches its peak ~50-60 kya, in agreement with the proposed time of introgression event.

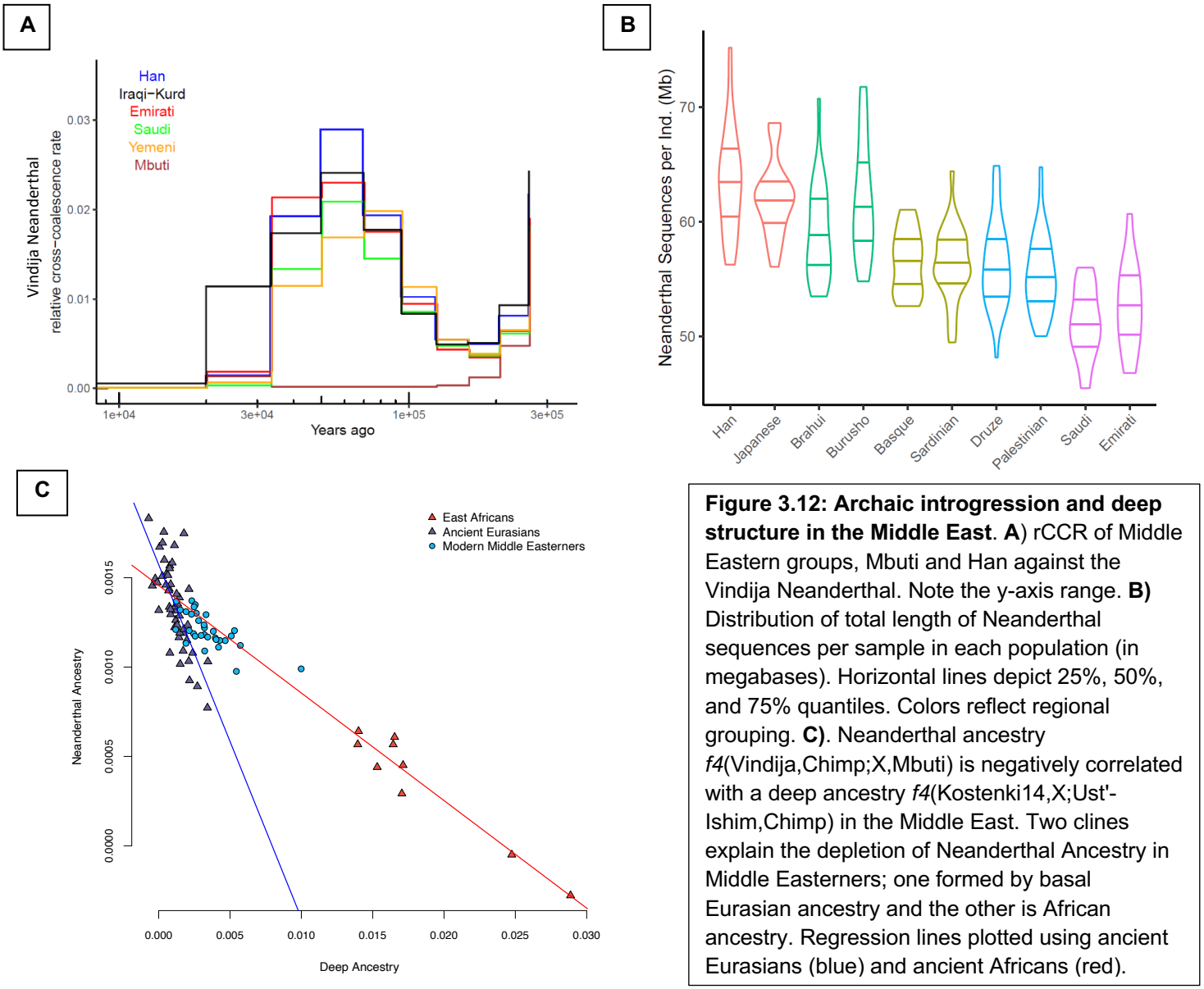

Figure 3.12: Archaic introgression and deep structure in the Middle East. **A**) rCCR of Middle Eastern groups, Mbuti and Han against the Vindija Neanderthal. Note the y-axis range. **B)** Distribution of total length of Neanderthal sequences per sample in each population (in megabases). Horizontal lines depict 25%, 50%, and 75% quantiles. Colors reflect regional grouping. **C)**. Neanderthal ancestry *f4*(Vindija,Chimp;X,Mbuti) is negatively correlated with a deep ancestry *f4*(Kostenki14,X;Ust'-Ishim,Chimp) in the Middle East. Two clines explain the depletion of Neanderthal Ancestry in Middle Easterners; one formed by basal Eurasian ancestry and the other is African ancestry. Regression lines plotted using ancient Eurasians (blue) and ancient Africans (red).

I subsequently used a recently-developed probabilistic identity-by-descent-based method (IBDmix), which directly compared a populations with a high-coverage Neanderthal genome to identify introgressed haplotypes (Chen *et al.,* 2020). Running this method on a merged dataset composed of our samples and the HGDP dataset identified segments totalling ~1.27 Gb suggested to be of Neanderthal origin. The amount of Neanderthal segments that were private to our dataset and not found non-Middle Eastern Eurasians was 25Mb, suggesting that most of the Neanderthal segments in the Middle East are shared outside of the region. I subsequently compared the total number of Neanderthal bases in each individual, and found on average, lower values in Arabian populations in comparison to other Eurasian groups, including Levantines. In this analysis I excluded the Yemeni population as they have high African ancestry and focused on the Emirati.core and Saudi.core populations which have ≤ 3% African admixture (Four different methods are in agreement with this estimated proportion: ADMIXTURE, qpAdm, qpGraph and fastGLOBETROTTER). The Sardinian and Druze populations have similar amounts of Neanderthal admixture, ~56.4 Mb per individual on average (Figure 3.12B). Arabia in contrast, Saudi.core and Emirati.core have 52.1 and 52.7 Mb of Neanderthal segments respectively. This value is ~8% lower than Sardinians and Druze and ~20% less than Han Chinese. The estimated African ancestry in the Arabian.core populations cannot explain the depletion of Neanderthal ancestry. A basal Eurasian population with low-to-no Neanderthal ancestry has been proposed to have existed and contributed different proportions to ancient and modern Eurasian populations, with the highest proportion in Neolithic Iranians and Natufians (~50%; Lazaridis *et al*., 2016). Since we have shown that Arabian populations have higher Natufian-related ancestry than other populations in the region, this indicates that this is the likely reason they have lower Neanderthal ancestry relative to other populations, including Levantines. To further explore this while controlling for African ancestry, we find a significantly negative correlation (Pearson's r = -0.81, $P = 2.7 \times 10^{-6}$) when plotting the statistics f4(Vindija,Chimp,X,Mbuti) and f4(Kostenski14,X,Ust-Ishim,Chimp). The former estimates the amount of allele sharing with the Vindija Neanderthal, the latter the amount of 'deep ancestry' relative to Ust-Ishim. In other words, populations with more Vindija-related drift, share less drift with Ust'-Ishim and thus derive some ancestry from a pre-Ust'Ishim ("Basal Eurasian") ancestry. Two clines are apparent in Figure 3.12C which thus explain the depletion of Neanderthal ancestry, one due to recent African-admixture, and the other to Basal Eurasian ancestry present in ancient Eurasians. Modern-day Middle Easterners seem to be affected by both clines since they have both ancestries.

To complement this analysis without direct comparison to the Neanderthal genome, I used another method, Sprime, which identifies divergent segments not present in an outgroup. Instead of including only African groups in the outgroup, I included all non-Middle Eastern HGDP populations with the aim of identifying divergent segments not present, or found at very low frequency, outside of the region. These diverged segments do not have to be of Neanderthal origin, as it could identify segments introduced from another hypothetical diverged hominin that admixed with the Middle Easterners. I compared the results of Sprime with published Vindija Neanderthal and Altai Denisovan high-coverage genomes, and find that most of the identified segments are Neanderthal, but a few match Denisova. The latter are perhaps also likely to be Neanderthal, but absent from the Vindija genome to due to polymorphism between Vindija and the admixing Neanderthals. This analysis identified two relatively large segments that are common in Arabia, but very rare elsewhere. The first is a 496kb segment located on chromosome 13 which is present at ~20% frequency in Saudis but found at 0.02% globally (1000GP). This segment overlaps *GPC5*, a gene expressed in brain tissues. The second is a 499kb segment on chromosome 4 present at ~20% frequency in the Emirati.core groups and less than 0.05% globally, which overlaps two genes: *CFAP299* which is expressed in the testes and plays a role in spermatogenesis, and *BMP3*, a cytokine which involved in cartilage and bone development. I searched for amino acid substitutions within these segments, but none were identified within canonical transcripts, with variants mostly in introns. As these segments show varied frequency within Arabian populations, which have diverged recently as shown in our demographic analysis, it is likely that their current appreciable frequency has increased recently due to genetic drift.

3.9 Selection in the Middle East

The current arid climate in Arabia and the long-term nomadic-like lifestyle of many of its populations may have exerted selective pressure for adaptations. The further understand this, I searched the genome-wide genealogies for lineages that carry mutations which have spread unusually quickly (Speidel *et al.,* 2019). I used $P < 5\times10^{-8}$ as a conservative genome-wide significance threshold of evidence of selection. In agreement with two previous candidate-based resequencing studies (Imtiaz *et al.,* 2007; Enattah *et al.,* 2008) I find evidence of positive selection in the *LCT/MCM6* locus (Figure 3.13A). It is known from these two studies that Arabian populations harbour two correlated variants (rs41380347 and rs55660827), distinct from the known European variant (rs4988235), which are associated with lactase persistence. I

estimated the strength of selection by measuring how fast the lineage spread (Stern *et al.,* 2019), and the Arabian *LCT/MCM6* variants show evidence of strong selection (s = 0.011, logLR = 13.27). Interestingly, this selection coefficient is similar to, but slightly weaker than, the variant identified in Europeans (s = 0.016-0.018; Mathieson and Mathieson, 2018; Stern *et al.,* 2019). Within our dataset, the haplotype appears at the highest frequency in the core Arabian groups: ~50% in Emiratis and Saudis, and at much lower frequency in the Levant and Iraq (4%). Notably, it is not present in any Eurasian and African population in the 1000GP, but has been reported at low frequency in some East African groups (Tishkoff et al., 2007). To explore the history of this variant further, I checked if it is present in any published ancient samples, including ancient Iranians and Levantine populations. As many studies use a capture-based approach that enriches for a set of variants (e.g. 1240K SNVs) in which this SNV is not present, I was limited to whole-genome-sequenced ancient datasets; however, none of the samples examined carry the variant. This appears consistent with a recent origin of the haplotype within the region, with a subsequent increased in frequency due to positive selection. This was confirmed after reconstructing the allele frequency trajectory of this variant using coalescent-based analysis (Figure 3.13A), as it shows a striking increase in frequency between around 6-9 kya and the present day. Notably, this overlaps the transition from a hunter-gatherer to a herder-gatherer lifestyle in Arabia, suggesting a change in lifestyle created a selective pressure for which this variant was a target. It has been proposed that the domestication of dromedary camel and the consumption of its milk are implicated with selection on this variant (Enattah *et al.,* 2008).

We find another significant locus showing strong selection at rs35241117 (Figure 3.13B, s = 0.007, logLR = 8.1, which lies downstream of the gene *TNKS* which encodes the enzyme tankyrase. Multiple microRNA genes are also found in the region, including MIR-124 and MIR-597. The variant is present at the highest global frequency in Saudis and Yemenis (~60%), and is associated with a many immunological, metabolic, and skeletal traits, including hypertension, glomerular filtration rate, diuretics and BMI (Canela-Xandri *et al.,* 2018; Watanabe *et al.,* 2019). This variant lies outside a haplotype that has been recently proposed to be under selection in Kuwaitis and Saudis (Eaaswarkhanth *et al.,* 2020). There is moderate LD between the variant and the haplotype (r^2 = 0.51). Although we do detect signals of selection on variants within this haplotype as well, after fine-mapping rs35241117 shows the highest evidence of selection. As the previous paper used a dataset genotyped on array data, it is likely that their signal is actually linked to rs35241117, so they underestimate the strength of selection by almost a half based on

our analysis. As many of the traits associated with this variant in GWAS appear to affect kidney-related functions, it is tempting to speculate that it is linked to adaptation to the hyper-arid climate of Arabia. However, further functional work is required to support this hypothesis.

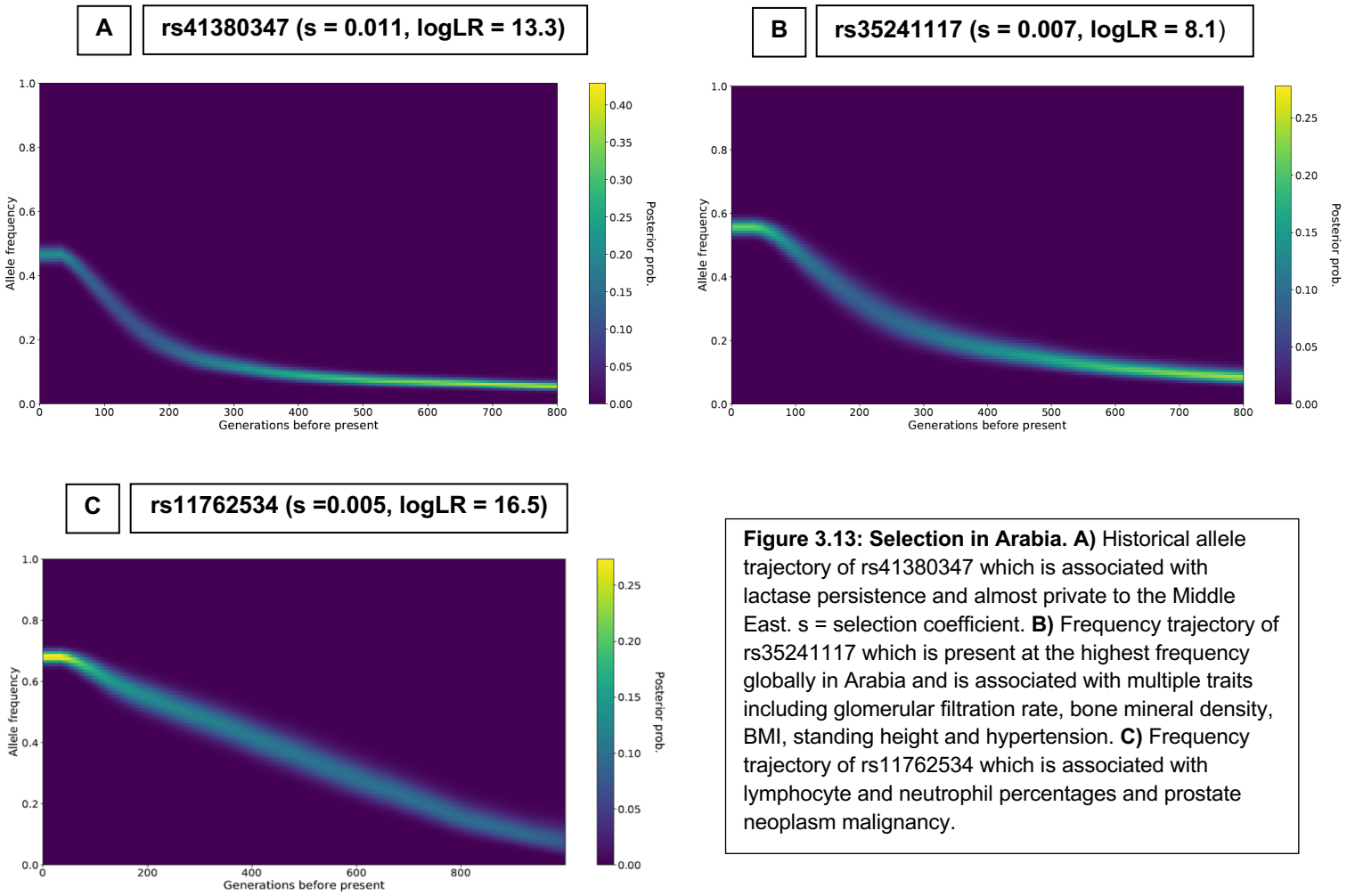

Figure 3.13: Selection in Arabia. A) Historical allele trajectory of rs41380347 which is associated with lactase persistence and almost private to the Middle East. s = selection coefficient. **B)** Frequency trajectory of rs35241117 which is present at the highest frequency globally in Arabia and is associated with multiple traits including glomerular filtration rate, bone mineral density, BMI, standing height and hypertension. **C)** Frequency trajectory of rs11762534 which is associated with lymphocyte and neutrophil percentages and prostate neoplasm malignancy.

Another variant, rs11762534, located within *LMTK2* shows evidence of moderate selection (s=0.005; logLR = 16.49; Figure 3.13C) and has been reported to be associated with malignant neoplasm of prostate and blood cell percentages (Canela-Xandri *et al.,* 2018; Watanabe *et al.,* 2019). In addition, this variant is an eQTL for many genes (The GTEx Consortium, 2020). *LMTK2* encodes a kinase that is implicated in multiple cellular processes including growth factor signalling and apoptosis, and appears to be necessary for spermatogenesis in mice (Kawa *et al.,* 2006; Cruz *et al.,* 2019). This variant shows high stratification globally: it is almost absent in Africans and East Asians, at low frequency in South Asians (< 10%), and is present at 45% in Europeans (1000GP). However it appears at higher frequency, 66%, in the Arabian populations

and notably the variant also shows an even higher frequency in BedouinB (81%), while appearing less common in Levantines (Druze and Palestinians, both ~55%).

Genome-wide genealogies are powerful in identifying variants under positive selection as they represent the data in a rich format, an ARG, in contrast to other selection tests that use summary statistics. However, a limitation of the method used here is it tests for selection only on derived alleles, and in addition it is not clear if it can identify post-admixture selection. To further explore this, I compared Arabians and Levantines and looked for strongly differentiated variants (Figure 3.14). The most extreme PBS value in Yemenis is rs2814778, which results in the Duffy-null phenotype and is almost confined to African populations (1000GP). In Yemenis, however, this variant appears at high frequency (74%), and appears to decrease higher up the peninsula and in the Levant (59% in Saudis, 6% in Iraqi-Arabs). I ran a local ancestry deconvolution algorithm to distinguish African from non-African segments in our dataset, and this locus shows the highest enrichment of African ancestry across the genome (Maples *et al.,* 2013), not only in Yemenis, but also in other Arabian populations. As the proportion of African ancestry in Yemenis and Saudis appears to be around 9% and 3% respectively, this over-representation of African ancestry at this locus is indicative of positive selection after African admixture. The derived allele is a well-known variant that has been suggested to be protective of *P. vivax* infection (Miller *et al*., 1976), historically present in Arabia, particularly in the South West (Yemen) and West of Arabia; moreover, the historical geographic range of *P. vivax* seems to correlate with the frequency of rs2814778 in the Middle East.

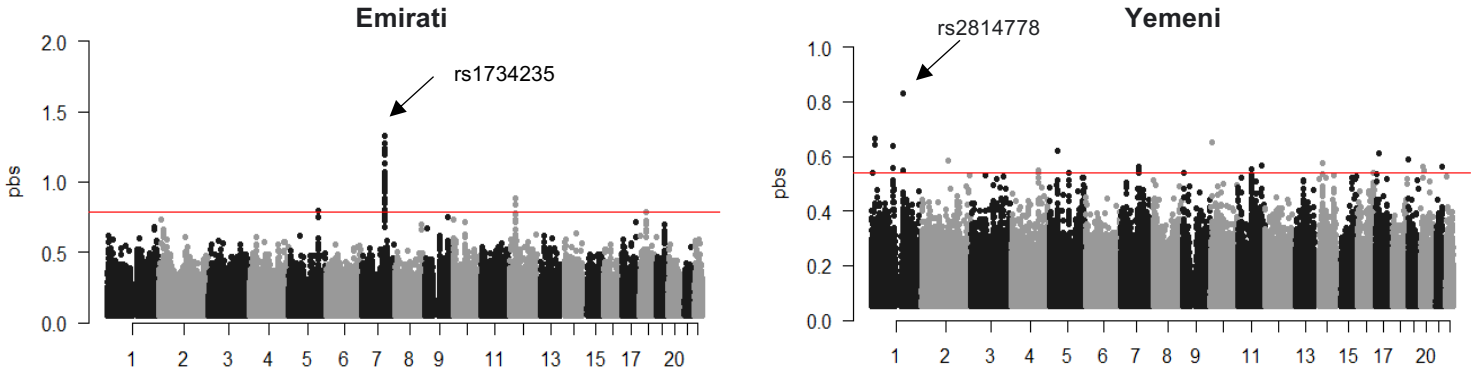

Figure 3.14: PBS comparing Arabians with Iraqi_Arabs and using Syrians as an outgroup. Red line illustrates 99.999% quantile. Note the different y-axis scales between the two figures. **Left**: In Emiratis (also in Saudis), we find a strong signal of differentiation at a 97kb haplotype on chromosome 7. Variants on this haplotype (rs1734235) almost reaches fixation (97% and 85%, in Emiratis and Saudis respectively) and are associated with increased expression of the lincRNA (AC003088.1). **Right**: rs2814778 is found at high frequencies in Yemenis and results in the Duffy-null genotype.

As genome-wide genealogies have the potential of identifying relatively weak selection, I subsequently searched for evidence of polygenic adaptation in Arabian populations. In contrast to selective sweeps, where strong selection on a variant results in a large change in allele frequency over a short period, polygenic selection is thought to occur through a more subtle allele frequency shift across a large number of variants. Using summary statistics calculated from Biobank-based GWAS, and genome-wide genealogies of Arabian populations generated by Relate, I tested for transient directional selection across 20 polygenic traits specifically over the past 2,000 years (Stern *et al.,* 2021). Most traits show no, or inconclusive, evidence of recent polygenic selection, including skin colour, height and BMI (Figure 3.15). However, I do find a few traits that show significant evidence of polygenic selection, with selection for higher years of education (EduYears) showing a consistent signal across all three Arabian groups (P = 0.0002 in Saudis). This result has also been found in the British populations (Stern *et al.,* 2021); however, this signal becomes much reduced when conditioning on other traits, which suggests that polygenic selection is not directly acting on the phenotype (EduYears), but indirectly through a correlated trait. Contrasting with findings in the British population (Stern *et al.,* 2021), I do not find any evidence of selection on traits such as hair colour, sunburn or tanning ability. Across the three Arabian populations, the direction of selection appears similar on most traits, potentially due to shared ancestry; however, it should be noted that the current varied environment across the region may result in different recent selective pressures on the populations. In the Emiratis in particular, I found a significant signal of selection on variants that increase type 2 diabetes (P = 0.004). This is an interesting result as the prevalence of type 2 diabetes in Emiratis is among the highest globally, and has been proposed, in part, as a consequence of a strong recent shift from a herder-gatherer and fishing-based mode of sustenance to a sedentary lifestyle (Malik *et al.,* 2005). In the same population we also find nominal evidence of polygenic selection on variants decreasing the levels of Apoliprotein B (P = 0.01) and increase the levels of low-density lipoproteins (P = 0.01); however, after correcting for multiple testing they appear only suggestive (P adj= 0.06 at 5% FDR).

3.10 Discussion

In this chapter I presented the analysis of a high-coverage open-access dataset from the Middle East, a region particularly understudied by global sequencing projects. This is the first human dataset where all samples are physically-phased using linked-read sequencing, which allowed the reconstruction of accurate haplotypes for analysis. I find that millions of variants are

identified within this dataset that are not catalogued in previous projects, with an appreciable number appearing common in the population. Notably, most of the common variants are located in regions that are inaccessible to short-read studies, highlighting the limitations of such technology, as a significant proportion of the human genome remains inaccessible for high confidence analysis (~25%; Bergstrom *et al.,* 2020).

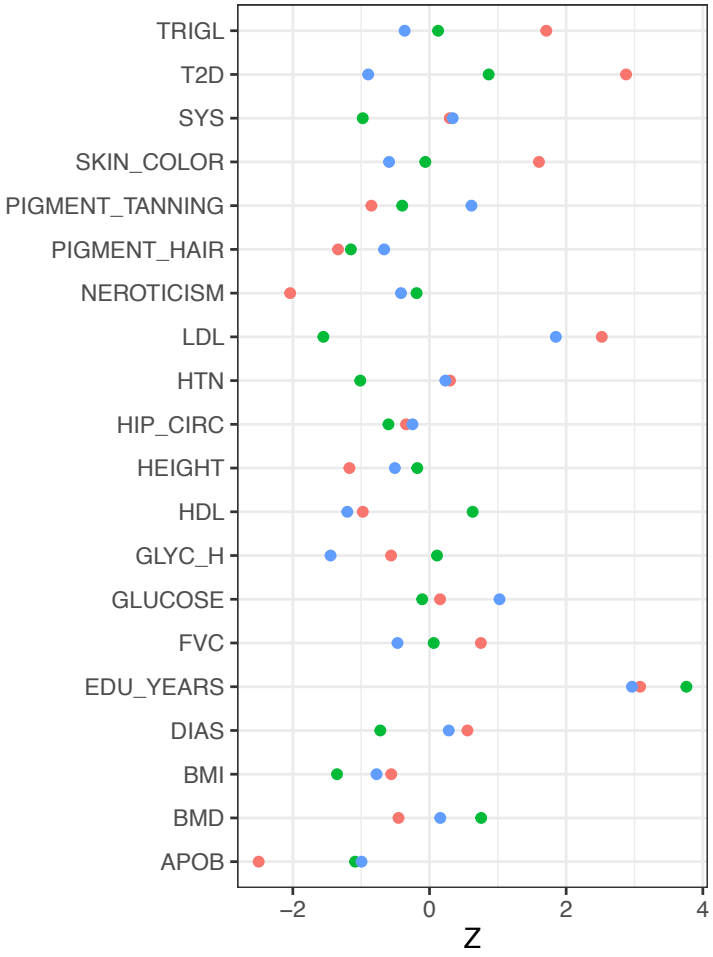

Figure 3.15: Polygenic selection in Arabia. Testing for recent polygenic selection, over the past 2000 years, on 20 traits within Arabian populations. Asterisks indicate the test is significant after correcting for multiple testing (FDR = 5%). TRIGL: Triglycerides; T2D: Type2 Diabetes; SYS: Systemic Blood Pressure; LDL: Low-density lipoproteins; HTN: Hypertension; HIP_CIRC: Hip circumference; HDL: High-density lipoproteins; GLYC_H: Glycosylated *haemoglobin;* FVC*:* Forced Vital Capacity*;* EDU_YEARS: Years of Education; DIAS: Diastolic blood pressure; BMI: Body Mass Index; BMD: Bone Mass Density; APOB: Apoliprotein B

I exploited the large number of experimentally-phased genomes to reconstruct the demographic history of Middle Eastern populations from relatively old periods (>250 kya), to up to the last thousand years. I found no evidence that contemporary populations have ancestry from archeologically-documented early expansions out of Africa (>60 kya) in the region, suggesting that, within the limits of resolution of the method (Schiffels and Durbin, 2014; Bergstrom *et al.,* 2020), these populations did not contribute genetically to modern groups. This result supports the growing consensus that present-day non-African populations descend from a single

expansion out-of-Africa, which was quickly followed by encountering and admixing with Neanderthals, probably in the Middle East, before populating the rest of the world (Mallick *et al.,* 2016; Bergstrom *et al.,* 2020).

I found that Middle Eastern groups have little Neanderthal DNA that is private to region, with the vast majority of introgressed segments shared with other Eurasians, implying that they share the same admixture event. Arabian populations appear to have a lower percentage of Neanderthal ancestry than Levantine, European and East Asian groups, due to them having relatively higher basal Eurasian ancestry than other non-African populations, in addition to more recent African ancestry. This hypothetical basal population did not contribute directly to modern-day Arabians, but through ancient populations who themselves had high basal Eurasian ancestry, Neolithic Iranians and Natufians. Differences between Arabia and the Levant are apparent: Arabians have higher Natufian ancestry, while Levantine groups have higher Anatolian and European hunter-gather-like ancestry. This suggests that many post-Bronze-Age events that changed the genetic landscape of the Levant did not reach Arabia. The contrast between the two regions also appears in population size history estimates, as they diverge around 15-20 kya, predating the Neolithic period. The Levantine expansion becomes especially prominent in the Neolithic period, suggesting that the transition to a sedentary agricultural lifestyle allowed the population to grow dramatically in numbers within a short time. This pattern is not found in Arabian populations who maintained similar sizes since their divergence. The population history analysis also shows that Arabian populations experience a bottleneck around the time of the aridification of Arabia 6-7 kya, while Levantines experience a distinct bottleneck overlapping the 4.2 kiloyear aridification event. This indicates that climatic events appear to have had strong effects on these populations, which is in agreement with archeologically data: the 4.2 kiloyear event has been suggested to be associated with the collapse of empires and kingdoms in the Middle East (Weiss *et al.,* 1993).

Our analysis also allows us to address ongoing debates in archaeology. It has been suggested that a complete population turnover occurred during the late Pleistocene/early Holocene in Arabia, and the peninsula was repopulated by Levantine Neolithic farmers moving southwards from the Fertile Crescent (Uerpmann *et al.,* 2010). Our analyses do not support such an event, a complete replacement. Moreover, our modelling indicates that Arabian populations can derive ancestry from Natufian-like local hunter-gatherers instead of Levantine Neolithic groups, in contrast to modern-day Levantines who cannot be modelled as such. The Levantine Neolithic

population shares around one third of their ancestry with Anatolian Neolithic groups; if such a Neolithic expansion occurred, we would expect to find Anatolian Neolithic ancestry in Arabian populations, which we find to be very low. Moreover, if Neolithic Levantines repopulated Arabia, we would expect to see similar population size histories during this period, which is not apparent. These results also address a second archaeological debate regarding prehistorical Arabian and Levantine connections which centres on whether, and how, animal domesticates moved between the Levant and Arabia, whether this represents population movements or cultural diffusion (Crassard *et al.,* 2013). Our results suggest the latter scenario, and if admixture did actually occur, it appears to be limited.

Another significant source of the ancestry found in contemporary Middle Eastern populations is the Iranian Neolithic, which replaced almost half of the local ancestry during the Bronze Age. In our admixture analysis we show that this ancestry was introduced into the Levant first, and then subsequently reached Egyptian, East African and Arabian populations. By comparisons with the published linguistic analysis, we find that the timing of this movement appears to be correlated with the origin and spread of Semitic languages during the Bronze Age (Kitchen *et al.,* 2009). While the broad formation of Middle Eastern populations appears to be somewhat understood, it is clear that more ancient DNA studies from the region are needed, especially from Arabia as no such study has been published at the time of writing this thesis. Although the hot, arid and sometimes humid climate in the region renders the preservation of DNA difficult, the success of several recent studies from tropical and hot climates raises some hope of the possibility of being able to analyse ancient remains from Arabia (Lipson *et al.,* 2020; McColl *et al*., 2018). In this study we suggest that local hunter-gatherers from Arabia, which have Natufian-like ancestry, could have directly contributed ancestry to modern-day populations; however, the epi-paleolithic period of Arabia is poorly understood. Samples are needed at this crucial period to test our findings, and to understand the population structure of the region at that time. It is known that the development of agriculture has led to population movements which have strongly reduced genetic structure in many regions, thus it is likely that the hunter-gathers of the Levant and Arabia were strongly differentiated.

Exploiting the accurate physically-phased haplotypes in generating genome-wide genealogies, I was able to reconstruct the historical trajectory of variants and identify ones that show evidence of positive selection. We replicated, refined and identified novel signals of selection within Arabian populations, although for most of them there doesn't appear to be an obvious reason

for why such variants were selected. Further functional studies are required to elucidate their function and pinpoint why, and how, they aided adaptation. The almost private lactase persistence allele within the region, which in just a few thousand years appears to have increased in frequency to reach 50%, correlating with lifestyle changes, illustrates the importance of including underrepresented groups in genomic analysis to illuminate genetic adaptations. In addition to single variants, I find evidence that polygenic selection seems to have also played a role in adaptations. It appears that variants that may have been beneficial in the past, potentially due to adaptation to the climate, are now associated with diseases such as T2D. However, it also should be noted that we find very few traits that show signals of polygenic adaptation. This could be potentially due to the long-term small effective population size of Arabian populations, which result in higher rates of genetic drift reducing the impact of selection. Moreover, as Middle Eastern groups are among the most understudied populations included in GWAS (Sirugo *et al.,* 2019), this creates challenges and limitations for the analysis of polygenic traits. Such long-term small effective population size, especially coupled with the recent practice of consanguinity, is also likely to increase the genetic burden of Arabian populations. This can be exploited for the study of Mendelian traits, such as in homozygosity mapping, and also to understand gene function as individuals are more likely to carry homozygous loss-of-function mutations and serve as natural ‘human knockouts’. In a positive development, many countries in the region have recently established national biobanks, a first step in hopefully reducing these disparities, and offer an opportunity to understand complex and disease traits in the wider Middle East.

3.11 Methods

This section provides a summary of the methods used in this chapter, more details are provided in Almarri *et al.* 2020b.

Sample Extraction and Sequencing

Saliva samples were collected using Oragene DNA kits (OG-600) and DNA was extracted using a high molecular weight method (Qiagen MagAttract HMW kit). Quality and size of the extracted fragments were assessed using a pulsed-field capillary electrophoresis (Femto Pulse system). For most samples the fragments appeared to be relatively large and of a useful size (>30kb). Based on these results, 137 samples were subsequently processed for library preparation at the Wellcome Sanger Institute sequencing facility. All libraries were prepared using 10X Genomics Chromium kits and each sample was then sequenced in a separate lane on a HiSeq X instrument. Raw FASTQ files generated from the sequencing instruments were processed using the Long Ranger pipeline (version 2.2.2, using GATK v3.7) and mapped to the GRCh38 reference supplied by 10x Genomics using barcode-aware alignment into phased BAM files and subsequently phased VCF files. The average sequencing coverage for all samples was 32x, median 31x. Around 98% of SNVs on average were physically-phased in each sample.

Sample and Variant Quality Control

Sample quality control was assessed using indexcov (Pederson *et al.,* 2017) which exploits coverage to identify chromosomal abnormalities and quality issues. No such large rearrangements were observed; however, a few samples showed issues with variable coverage, mostly affecting one sample, with a more limited issue in three other samples. These samples were sequenced in the same batch, suggesting a shared issue. After exploring this issue further, it appears related to relatively low input DNA used during library preparation, which was confirmed after contacting the sequencing facility. These samples still provide accurate genotypes and were included in most subsequent analysis, but were highlighted and excluded from sensitive demographic history analyses. Using the relative coverage of X to Y chromosome, 79 samples were inferred to be male and 58 female.

For variant quality control, the QUALITY tag of each variant was assessed within each sample VCF separately. The Long Ranger pipeline analyses of haplotype structure informed by the experimental phasing was used to tag variants that are potential false positives. Quality tags

produced include: PASS, 10X_PHASING_INCONSISTENT, 10X_QUALITY_FILTER, 10X_ALLELE_FRACTION_FILTER, 10X_HOMOPOLYMER_UNPHASED_INSERTION, 10X_RESCUED_MOLECULE_HIGH_DIVERSITY. Analysing each quality tag separately, non-PASS variants generally had low transition to transversion ratio (Ts/Tv) values, suggesting they contain false positives. This ranged from 0.55 in the 10X_QUALITY_FILTER variants to 1.16 in the 10X_RESCUED_MOLECULE_HIGH_DIVERSITY variants. The latter tag rescues variants in duplicated regions where the reference likely harbours deletions. I chose to be conservative by setting all non-PASS variants to missing and subsequently merged the samples into one multi-sample VCF file. I then set any variant with GQ < 20 and regions that show twice or more average sample coverage to missing, and also excluded variants that show excessive heterozygosity. The Ts/Tv after quality control was 1.97, and remained consistent throughout different allele frequency bins, suggesting that the variants are of high quality. This identified 23.1 million SNVs. I subsequently tested for possible relatedness using the --genome option in plink-v1.9 (Chang *et al.,* 2015) and excluded one Syrian sample from a pair that show evidence of being related (PI_HAT > 0.15).

Population Genetic Analysis

Our dataset was combined with published modern and ancient global populations extracted from the available curated dataset from the Reich Lab (https://reich.hms.harvard.edu/downloadable-genotypespresent-day-and-ancient-dna-data-compiled-published-papers). Variants were lifted over to GRCh38 using picard (v2.18.26, https://broadinstitute.github.io/picard/). To avoid bias from different phasing methods, as advised by the authors of the Chromopainter/FineSTRUCTURE software, we discarded for this specific test the physical phasing from our samples and phased the merged dataset with Eagle v2.4.1 (Loh *et al.,* 2016) using the 1000 Genomes Project phase 3 panel (1000 Genomes Project Consortium *et al.,* 2015) and subsequently run using the Chromopainter/FineSTRUCTURE pipeline v4.1.1. Based on the output, we divided the self-labelled populations into representative 'core' subpopulations that show limited-to-no recent admixture. To investigate potential sources of admixture, we used SOURCEFINDv2 (Chacón-Duque *et al.,* 2018) which was run using the default parameters. To test for and date admixture, fastGLOBETROTTER, was run using the parameters (prop.ind: 1, bootstrap.date.ind: 1, null.ind: 1, with all remaining parameters default) including only as surrogates populations that contributed >1% ancestry in the previous SOURCEFIND step. We also tested for admixture using MALDER (Loh *et al.,* 2013; Pickrell *et al.,* 2014) using default parameters using 590k variants using 6 references: (Luhya;Yoruba;Druze;Iranian;Indian Telugu;Punjabi).

For the demographic history analyses we leveraged the physical-phasing in our dataset to generate genome-wide genealogies using RELATE v1.1 (Speidel *et al.,* 2019). We limited the analyses to regions within the genome accessibility mask described in Bergstrom *et al.,* 2020 and set unphased variants to missing (i.e. excluded from analysis). We scaled the results using a mutation rate of 1.25e-8 and a generation time of 29 years. For the separation history analysis with global populations, we downloaded the Mbuti, Sardinian, and Han Chinese physically-phased samples from the HGDP (2 samples per population, Bergstrom *et al.,* 2020). Since the published data used an older version of Long Ranger, we recalled the samples using v.2.2.2 to be consistent with our dataset. We filtered the VCFs as described above for our dataset. We used MSMC2 v2.1.1 to infer split times between our populations and the HGDP samples using 8 haplotypes for each comparison (4 haplotypes from each population). MSMC2 was run using the --skipAmbiguous option, to calculate coalescent rates within and between populations, restricted to the genome accessibility mask described in Bergstrom *et al.,* 2020.

We used IBDMix (Chen *et al.,* 2020) to call Neanderthal segments in a merged dataset of our samples with the HGDP. We followed all the variant filtering steps performed in Chen *et al.,* 2020 and filtered the output using a minimum size threshold of 50kb and LOD score higher than 4. We also ran Sprime (Browning *et al.,* 2018) on a similar dataset. As Sprime requires non-missing genotypes, we removed variants that were >5% missing and imputed the remaining missing variants using Eaglev2.4.1 (Loh *et al.,* 2016). We set all non-Middle Eastern samples from the HGDP as outgroup (768 samples) and ran Sprime for each Arabian population. We filtered the output using a score threshold of 150,000. We applied MSMC2 using 4 haplotypes (2 diploid samples) to examine the separation history between our populations and the high coverage Vindija Neanderthal (Prufer *et al.,* 2017). Briefly, this analysis exploits the fact that the Vindija Neanderthal shows extremely low heterozygosity, which renders much of the genome homozygous and essentially phased. We used the --skipAmbiguous option to exclude sites with unknown phase and ran MSMC2 using the parameters previously described.

We used the Relate Selection Test in Relate v1.1 (Speidel *et al.,* 2019) to look for lineages that spread faster than competing lineages. We used BEAGLEv4 (Browning and Browning, 2016) to statistically phase the remaining unphased variants (~2%) using gtgl, setting the option usephase=true to take into account the physical-phasing already provided in the VCF. We only analysed variants that show selection at a “genome-wide threshold” of $P < 5e\text{-}8$. To further refine and understand the evolutionary history of the variant we used CLUES (Stern *et al.,*

2019). We fine-mapped variants using the likelihood ratio statistic produced by CLUES as suggested by Stern *et al.,* 2019, and focused on variants that show moderate to strong selection (s > 0.005). To test for polygenic selection, we used PALM (Stern *et al.,* 2021). We extracted GWAS summary statistics performed on the UK BioBank (Bycroft *et al.,* 2018; Gazal *et al.,* 2018; Hujoel *et al.,* 2020; http://www.nealelab.is/uk-biobank/). For each trait investigated, we split the genome into 1,700 approximately independent blocks (Berisa and Pickrell, 2016) and selected the variant with the lowest p-value within each block for analysis. Variants were filtered for minor allele frequency > 5%, Rsq > 0.5, INFO score > 0.8, and indels were excluded. To further explore the choice of significance threshold, and potential effects of uncorrected population structure, on the results, we repeated the analysis using more stringent significance thresholds (P < 1e-8 and P < 5e-9; which will drop blocks not passing the threshold) and found similar results.

For model-based clustering, ADMIXTUREv1.3 (Alexander *et al.,* 2009) was run on modern samples in an unsupervised mode from K=3 to K=15 using 1.3 million variants ascertained as polymorphic in archaic genomes (Bergstrom *et al.,* 2020). For tests including ancient samples, Dystruct (Joseph *et al.,* 2019) was run using default parameters on ~80,000 transversions. Samples were randomly subsetted to ≤ 10 individuals per modern population and ≤ 20 individuals per ancient population. We set nine time points binned as follows (in years ago): 14,500-10,000; 10,000-8000; 8000-6000; 6000-5200; 5200-5000; 5000-3000; 3000-1400;1400-200; and present-day.

We used the Phewas search option in the GWAS atlas (Watanabe *et al.,* 2019) and Gene Atlas (Canela-Xandri *et al.,* 2018) to look for trait associations with variants that show evidence of selection. We used the GTEx portal (GTEx Analysis Release V8; The GTEx Consortium, 2020) to look for eQTL associations. We used plinkv1.9 (Chang *et al.,* 2015) to identify ROHs using the option --homozyg. Genotype calling, filtering and Y haplogroup prediction were run as performed in Hallast *et al.,* 2020. To identify African haplotypes within our samples, we used RFMix v2.03 (Maples *et al.,* 2013) using 105 samples from the HGDP as references: 41 Druze and 64 Africans. f4 and F_{ST} statistics, qpAdm, qpGraph were run using the ADMIXTOOLS package (Patterson *et al.,* 2012).

Chapter 4: Future Directions

As the specific results of each chapter have already been discussed, in this final chapter I discuss general future directions in these areas.

4.1 Future directions for the analysis of structural variation.

Large-scale datasets of SVs have been published recently, much larger than the HGDP, based on short-read data. One is the gnomAD SV release (Collins *et al.,* 2020), which analyzed 14,891 genomes at high coverage (32x). Although the dataset claims to study diverse populations, almost half of the samples are of European ancestry, while the African samples in the dataset (35%), are mostly African-American, which will generally encompass admixed genomes with some West African ancestry. Another recent study, Abel *et al.,* 2020, examined 17,795 genomes at medium coverage (20x), but also had half of the dataset composed of European individuals. Moreover, for both these studies, samples are aggregated from different projects, mostly medical-based, and are subsequently assigned a continental-level ancestry after genetic analysis. These projects are useful in providing resources of allele frequencies of SVs at a broad continental-level, and have also identified rare and large SVs that affect coding sequences. In agreement with previous studies, such large SVs appear to be under purifying selection (Sudmant *et al.,* 2015a; Sudmant *et al.,* 2015b). However, the lack of population-level labels limits genetic analysis. For example, if the HGDP dataset was composed of only continental-level labels, we would do not have identified that specific groups have high frequencies of variants that show evidence of positive selection, that are rare in the wider region, or even private to the population. In our analysis, this is most evident in populations that have high diversity, such as African groups, and ones that show long-term isolation or private introgression events such as Oceanians. Another limitation is the restrictions imposed on the data, as both of these large-scale projects do not provide genotype-level data, just site-frequency data. This collectively illustrates the importance of projects such as the HGDP, and in addition the 1000GP, with their carefully-sampled populations and essentially open-access nature. While the HGDP has a large number of populations, 54, many populations have relatively-small sample sizes (27 populations have 12 samples or less). Nevertheless we find many examples of population-specific variants that appear to be medically and evolutionarily important. This suggests that additional projects with larger population-level sampling are

needed, to further understand the structure and stratification of SVs globally, especially for low-frequency and rare variants.

The 1000GP is a resource containing 2504 samples from 26 populations, with an appreciable number of samples per population (~100). However, its published SV callset is composed of ~68K variants, mostly deletions, much smaller than our HGDP dataset despite being more than double in sample size. A limitation of the 1000GP is the low-coverage of the 1000GP (~7x) and the shorter reads used (~100bp). A recent project, as yet unpublished, has sequenced the entire 1000GP samples at high-coverage, and has made the data available for the scientific community. It will be important to perform an SV analysis of this new dataset and compare it with other published datasets. Other recent studies with population-level sampling, although not open-access, are the GenomeAsia100K project (GenomeAsia100K Consortium, 2019) and a high-coverage dataset of 50 ethnolinguistic groups in Africa (Choudhury *et al.,* 2020). Although important datasets from understudied groups, it is disappointing that these projects did not analyse SVs. This demonstrates how they remain understudied in comparison to SNVs, despite their importance in genome evolution and disease susceptibility. Analyses of these datasets should be extended to include SVs.

The current state-of-the-art algorithms for the identification of SVs based on short-read sequencing studies identify 4-7 thousand variants per sample, substantially less than the >20 thousand identified using long-read and multi-platform technologies (Ho *et al.,* 2019). This demonstrates that the majority of SVs within a genome are undetected using widespread short-read technology. From a medical and complex trait perspective this is important, as SVs may be associated with diseases and traits but are missed by short-read technologies and subsequently not included in further analysis. For example, in the Deciphering Developmental Disorders project which attempts to pinpoint mutations associated with developmental disorders, a majority of patients, almost 70%, have not had a pathogenic mutation identified (Personal Communication, Hilary Martin). Additionally, from an evolutionary perspective, such undiscovered SVs may be involved in adaptation and positively selected. One cost-effective approach proposed is sequencing a small number of samples using long-read technologies in a discovery-phase, and then re-genotyping these variants in a larger short-read based cohort using a graph-based approach. This has recently been shown to have relatively good accuracy in genotyping different classes of SVs (Chen *et al.,* 2019; Eggertson *et al.,* 2019; Hickey *et al.,* 2020), as although short-reads cannot discover these variants, they provide some information to

genotype them. However, this approach will only include in analysis variants identified in the discovery-phase, which will miss many low-frequency and rare variants. Moreover, despite advances presented by these methods, they do not seem to perform well in genotyping variants in repetitive regions, complex variants, or large insertions, especially ones that are not completely assembled. Accurate genotyping is essential to test for associations, such as ones now routinely performed in SNV-based GWAS. Thus the development of accurate SV genotyping algorithms, for all classes of SVs and across the allele-frequency spectrum, will be an important advance for the inclusion of SVs in future association studies. A recent pre-print developed a graph-based method that exploits haplotype information from a reference of *de novo* assemblies to genotype SVs in short-read sequenced samples (Ebler *et al.,* 2020). This approach appears to provide high accuracy in genotyping different classes of SVs, even in repetitive regions, and is a promising method addressing the issues faced by previous algorithms.

The development of new technologies such as Strand-seq, which allows the sequencing of the paternal and maternal haplotype separately, coupled with advances in long-read sequencing, is now allowing the generation of chromosome-scale phased *de novo* assemblies (Porubsky *et al.,* 2020). This will allow balanced rearrangements, such as inversion and translocations, to be identified and included in analysis. Such structural variants are commonly excluded in population-scale analysis, due to the challenges in their discovery and genotyping. Inversions are an important class of structural variants, as they suppress recombination at heterozygous sites. For example, a 900kb inversion has been reported to be under positive selection in Europeans and is associated with higher fertility and recombination rates (Stefansson *et al.,* 2005). Translocations are known to associated with cancer, such as the relatively large translocation between chromosome 9 and 22 (‘Philadelphia chromosome’) which contributes to chronic myelogenous leukaemia (Nowell and Hungerford, 1960). Moreover these technologies will allow the analysis of complex SVs, which are unresolvable using short-reads. An extreme case is chromothripsis, a mutational event where thousands of clustered rearrangements occur in one or multiple chromosomes which are common across different cancers (Cortés-Ciriano *et al.,* 2020). The chromosome-scale phasing would also make genotyping much simpler, as each haplotype is genotyped separately.

These advances will also allow an important metric to be estimated, the mutation rate of SVs. Although it is known that indels and smaller repeats, such as STRs, have a higher mutation rate

than SNVs, the mutation rates of other classes of SVs are still unclear. Trio-based studies using long-read data coupled with accurate SV calling and genotyping is a promising approach to estimate their mutation rate. Moreover, this analysis can be extended to different primates allowing comparisons between different species. This will also assist in understanding the population structure of SVs. We have shown in the HGDP that all classes of SVs show population structure, although variable in degree, and we have suggested that this may be due to their varying mutation rates. It is becoming evident that even subtle population structure can confound association signals of SNVs, a class of variation that we have a relatively accurate estimate of mutation rate for (Jónsson *et al.,* 2017; Fu *et al.,* 2014). Thus it is unclear how to control for population structure of SVs in association studies, given their higher and variable mutation rates between classes. While SNVs can be in LD with SVs and be used to tag them in GWAS, SVs that are highly mutable, such as multiallelic variants, or ones that are located in repetitive regions, are unlikely to be in high LD with commonly used tag SNVs.

One of the 1000GP goals was to catalogue common global genetic diversity, especially from metropolitan populations, and they have mostly been able to do this successfully. Given the essentially open-access nature of the 1000GP and HGDP, and the availability of almost unlimited high-quality DNA, an important future project would be to sequence these samples using long-read and multiplatform technologies to create a comprehensive catalogue of global SVs, which crucially includes variants missed from short-read studies. A limitation of these projects is the lack of phenotype data available to look for associations. In contrast, the UK biobank, with around 500,000 samples, has extensive phenotyping data and will in the near future release full genome-data available for researchers, based on short-read technology. Although sequencing of such a large dataset using long-read technologies is economically-unfeasible today, the continued decreasing costs of such methods will hopefully make such a project more practical in the future. However, it should be noted that high molecular weight DNA needs to be extracted for such future projects. An SV-based GWAS analysis of this dataset will likely be an important resource to understand the functional effects of SVs on complex and disease traits. However, as the UK biobank is mostly of British European ancestry, such projects need to be initiated across the world to study human genetic diversity. In particular, a biobank-based project of populations with significant Denisovan introgression, such as Oceanians, would be instrumental in understanding the effect of introgressed sequences on complex and disease traits, potentially explaining the signals of adaptive introgression we find in our study.

The analysis of SVs, and even smaller classes of variation such as SNVs and indels, is affected by the reference genome used. Due to the limitations of a mostly linear human reference such as GRCh38, which does not encompass human genetic diversity, it is likely that future analyses will use a different reference format to account for the sequence variation found globally. Advances in technology are now generating assemblies that rival, and even exceed, the quality of GRCh38, are uncovering some of the most complex regions of the genome for functional study (Miga *et al.,* 2020). Moreover, many countries are generating population-specific reference genomes that offer a better representation of variation in their population (Cho *et al.,* 2016; Maretty *et al.,* 2017). While graph-based references are suggested as a solution to the limitations of a linear-based reference, there is currently no consensus on a standard that will integrate the variation found in multiple reference-quality genomes. Ideally, a human reference pangenome graph will maintain one coordinate system, while integrating variation supplied from many reference genomes. Encouragingly, a recent study has proposed a new format that can generate a human pangenome reference panel, although it is unclear if it can handle a large number of genomes (Li *et al.,* 2020). However, even if a reference standard is agreed upon which will provide a much better representation of human genetic diversity, it is uncertain when the research community will move on to using a newer reference genome. This is illustrated by many groups still using the GRCh37 reference in their analyses, despite GRCh38, which was released in 2013, being a substantially better-quality assembly (Schnieder *et al.,* 2017). Surprisingly, even high-profile large-scale genome sequencing projects are still being published using GRCh37, most recently in a large-scale analysis of African whole-genomes (Choudhury *et al.,* 2020).

4.2 Future directions for the population genomics of the Middle East

One of the main issues affecting the study of Middle Eastern populations is the lack of available genomic data, especially ones that are open-access. Surprisingly, there have been more published ancient genomes from the region than modern-day open-access whole-genomes. In work presented in this thesis, some progress has been achieved towards reducing the lack of representation of Middle Eastern genomes in global projects. The sequenced HGDP dataset analysed in chapter 2 has 134 high-coverage Levantine samples, while the Middle Eastern dataset presented in chapter 3 sequenced 137 high-coverage samples, using linked-read sequencing, from Arabia, the Levant and Iraq. Recently, in a positive development, the

gnomaAD allele frequency database has included the whole-genome sequenced HGDP populations in their online browser, allowing variant frequencies to be easily viewed in the three Levantine Middle Eastern groups: Bedouins, Palestinians and Druze. Although a warning is presented that due to their relatively small sample size, allele frequencies should be viewed with caution. In relatively-wealthy countries from Arabia, thousands of whole-genomes and exomes have been sequenced by ongoing national projects; however, the data are not publicly available, even for summary statistics such as allele frequencies (Kaiser, 2016). It is unclear if this data will be available to outside researchers, limiting the usefulness of such resources to the wider scientific community. In our analysis we show that Arabian populations have diversified recently, within the past 3ky. At the very least, regional national projects should collaborate to generate a reference panel that will be extremely useful for phasing and imputation of variants for GWAS. Privacy concerns regarding data of sequenced participants can be addressed by creating a web-based imputation service, such as the Michigan imputation server (Das *et al.,* 2016), that removes any issues with data access agreements and provides a fast and user-friendly method for imputation. While some Arabian countries are making progress in national biobanks, other countries in the region are unfortunately affected with political instabilities, making the planning and execution of large-scale national sequencing projects difficult.

An interesting follow-up study for the population structure of the Middle East is the careful sampling of individuals whose grandparents were all born in the same area, as has been done in the People of the British Isles project (Leslie *et al.,* 2015), and more recently in Spanish groups (Bycroft *et al.,* 2019). These projects have found strong substructure within these populations, with multiple genetically-differentiated sub-populations, even ones geographically located near each other, and have linked these patterns to historical events. In our analysis, we find strong substructure within the Middle East, concordant with geography, even within a small area. This should be further explored using a larger dataset with more comprehensive sampling across the region. In addition to providing insights into population history, such information would be important in designing and interpreting medical studies, as it has been recently demonstrated that including fine-scale haplotype information in GWAS reduces confounding caused by population structure in comparison standard SNV-based methods (Byrne *et al.,* 2020).

Ancient DNA from the region has provided unprecedented insights into the history of the Middle East, with almost all of the studies focused on the Levant, Iran and Anatolia. However, a nearby region that is surprisingly not sampled so far at the time of writing this thesis is Iraq, or historical Mesopotamia. One of the first civilizations developed in this region, Sumer, during the Chalcolithic and early Bronze Age (~7 kya). Moreover, one of the first written languages was Sumerian, a language isolate, which was written in cuneiform script. The language was gradually replaced by Akkadian, an East Semitic language during the rise of the Akkadian empire in the region. Who were the people living in these civilizations and speaking these ancient languages? Was the gradual language replacement accompanied by a turnover of ancestry? How did the establishment of the first civilizations and empires affect the population structure of the region? Many interesting questions can be addressed if ancient DNA is successfully extracted and analyzed from the region. The study of Mesopotamia also has important relevance for Arabian population history. The ancient civilization of Magan, thought to be in present-day Oman and United Arab Emirates, was referred to in Sumerian cuneiform texts as a source of copper trade to Mesopotamia. Another Bronze Age ancient civilization with strong trade links to Mesopotamia in Eastern Arabia was Dilmun, thought to be modern-day Bahrain. Whether or not these frequent contacts between different populations resulted in gene-flow is an open question. Moreover, parts of Arabia were conquered by different groups, and it is unclear whether this facilitated in population movements. Studies in the Levant have shown broad continuity in ancestry in the period after the Bronze Age until the present-day (Haber *et al.,* 2020); despite the region being under the rule of a large number of different groups at different times, such as the Ancient Egyptians, Babylonians, Assyrians, Persians, Greeks, Romans, Crusaders, Arabs, and Ottomans. These periods are associated with large cultural, linguistic and religious changes, but interestingly, do not seem to be paralleled with large changes in ancestry. Ancient DNA studies from Arabia are needed to investigate if similar patterns occurred there. Moreover, ancient DNA is important to uncover transient pulses of admixture that are not apparent in modern-day populations. An example is how the Crusades introduced European ancestry into the Levant, as they admixed with the local population during the military expeditions (Haber *et al.,* 2019). These admixture events appear to have been limited and did not survive in modern-day populations (except in one Y chromosome lineage; Zalloua *et al.,* 2008), and prior to ancient DNA results were largely unknown.

We find that the admixture dates of an ancient Iranian population in the region are correlated with the spread of Semitic languages estimated by linguistic data. This population replaced

around half of the local ancestry present at the time, and such a large turnover is often associated with technological innovations. For example, the domestication of the horse and invention of the wheel are thought to have assisted the spread of steppe pastoralists into Europe, where they replaced around half of the ancestry and may have introduced Indo-European languages. In the Middle East, we find that this population admixed first in the Levant (4-6 kya), and subsequently in Arabia (2-4 kya). While the invention of farming is known to have resulted in population movements and admixture, it is unclear how beneficial this technology was to movements into Arabia since the Arabian environment was already mostly desert at that time. This is also illustrated by the much smaller historical population size history of Arabians, around an order of magnitude lower than the Levant, illustrating that these population did not expand during and soon after the Neolithic period. This admixture and replacement of ancestry should be further studied, and additionally explored to investigate if it was sex-biased. For example, while the steppe ancestry movement into Iberia replaced around 40% of the local ancestry at the time, it replaced almost 100% of the local Y-chromosomes (Olalde *et al.,* 2019).

In addition to further work based on single nucleotide variant analyses discussed above, SVs need to be analyzed in Middle Eastern populations. Other than the dataset generated and analysed in the HGDP, there hasn't been a thorough investigation of SVs in Middle Eastern populations. As all the samples from our Middle Eastern dataset have been sequenced using linked-reads, having such a large number of *de novo* assemblies offers an opportunity to characterize SVs at relatively high resolution, overcoming limitations of standard short-read technology. However, to fully investigate the landscape of SVs, and it's implication in disease and adaptation, reference-quality assemblies need to be generated from the region.

4.3 Concluding Remarks

The availability of a large, and increasing, number of whole-genomes, including from diverse populations, offers an exciting opportunity to study human evolution and adaptation from a genetic perspective. While it is apparent that single variants with large effect sizes on traits are uncommon, I expect that regions of the genome that are inaccessible to standard short-reads will be an important source of genetic variation affecting complex traits, potentially including ones with strong effects. Such regions are repetitive and complex, with a higher mutation rate resulting in complex rearrangements, generating a pool of variation accessible to natural selection. The maturity of long-read technology, and the availability of extensive phenotype data from large biobanks, will allow the characterization of these variants and their function to be

elucidated. In addition, such resources are important in understanding the effects of polygenic selection, which, coupled with the developments of robust methods in their analysis, will provide a more comprehensive understanding of the role of natural selection in shaping human genomes. Finally, as we are now in an era where current and future generations will likely be sampled and sequenced by biobanks, we can for the first time witness and study human evolution in real-time. It is remarkable that such progress has been achieved since the publication of human genome sequence just twenty years ago.

Bibliography

1000 Genomes Project Consortium *et al.* (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), pp. 1061–1073.

1000 Genomes Project Consortium *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74.

Abel, H. J. *et al.* (2020) 'Mapping and characterization of structural variation in 17,795 human genomes', *Nature*, 583(7814), pp. 83–89.

Abi-Rached, L. et al. (2011) 'The shaping of modern human immune systems by multiregional admixture with archaic humans', *Science*, *334*(6052), pp. 89-94.

Abu-Amero, K. K. *et al.* (2008) 'Mitochondrial DNA structure in the Arabian Peninsula', *BMC evolutionary biology*, 8, p. 45.

Abu-Amero, K. K. *et al.* (2009) 'Saudi Arabian Y-Chromosome diversity and its relationship with nearby regions', *BMC genetics*, 10, p. 59.

Akay, A. *et al.* (2017) 'The Helicase Aquarius/EMB-4 Is Required to Overcome Intronic Barriers to Allow Nuclear RNAi Pathways to Heritably Silence Transcription', *Developmental cell*, 42(3), pp. 241–255.e6.

Akkaya, M. and Barclay, A. N. (2013) 'How do pathogens drive the evolution of paired receptors?', *European journal of immunology*, 43(2), pp. 303–313.

Alexander, D. H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals', *Genome research*, 19(9), pp. 1655–1664.

Allentoft, M. E. *et al.* (2015) 'Population genomics of Bronze Age Eurasia', *Nature*, 522(7555), pp. 167–172.

Almarri, M. A. *et al.* (2020a) 'Population Structure, Stratification, and Introgression of Human Structural Variation', *Cell*, 182(1), pp. 189–199.e15.

Almarri, M. A. *et al.* (2020b) ‘The Genomic History of the Middle East’. *biorxiv* doi: 10.1101/2020.10.18.342816.

Angata, T. *et al.* (2006) ‘Discovery of Siglec-14, a novel sialic acid receptor undergoing concerted evolution with Siglec-5 in primates’, *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 20(12), pp. 1964–1973.

Ardelean, C. F. *et al.* (2020) ‘Evidence of human occupation in Mexico around the Last Glacial Maximum’, *Nature*, 584(7819), pp. 87–92.

Armitage, S. J. *et al.* (2011) ‘The southern route “out of Africa”: evidence for an early expansion of modern humans into Arabia’, *Science*, 331(6016), pp. 453–456.

Audano, P. A. *et al.* (2019) ‘Characterizing the Major Structural Variant Alleles of the Human Genome’, *Cell*, 176(3), pp. 663–675.e19.

Axelsson, E. *et al.* (2013) ‘The genomic signature of dog domestication reveals adaptation to a starch-rich diet’, *Nature*, 495(7441), pp. 360–364.

Barrett, R. D., & Schluter, D. (2008) ‘Adaptation from standing genetic variation’. *Trends in ecology & evolution*, *23*(1), pp. 38-44.

Benazzi, S. *et al.* (2011) ‘Early dispersal of modern humans in Europe and implications for Neanderthal behaviour’, *Nature*, 479(7374), pp. 525–528.

Berg, J. J. *et al.* (2019) ‘Reduced signal for polygenic adaptation of height in UK Biobank’, *eLife*, 8. doi: 10.7554/eLife.39725.

Bergström, A. *et al.* (2017) ‘A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea’, *Science*, 357(6356), pp. 1160–1163.

Bergström, A. *et al.* (2020) ‘Insights into human genetic variation and population history from 929 diverse genomes’, *Science*, 367(6484). doi: 10.1126/science.aay5012.

Berisa, T. and Pickrell, J. K. (2016) ‘Approximately independent linkage disequilibrium blocks in human populations’, *Bioinformatics* , 32(2), pp. 283–285.

Beyter, D. *et al.* (2020) ‘Long read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits’. *biorxiv* doi: 10.1101/848366.

Bhatia, G. et al. (2013) ‘Estimating and interpreting FST: the impact of rare variants’, *Genome research*, *23*(9), pp. 1514-1521.

Bittles, A.H. and Black, M.L. (2010) ‘Consanguinity, human evolution, and complex diseases’. *Proceedings of the National Academy of Sciences*, 107(suppl 1), pp.1779-1786.

Browning, B. L. and Browning, S. R. (2016) ‘Genotype Imputation with Millions of Reference Samples’, *American journal of human genetics*, 98(1), pp. 116–126.

Browning, S. R. *et al.* (2018) ‘Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture’, *Cell*, 173(1), pp. 53–61.e9.

Bycroft, C. *et al.* (2019) ‘Patterns of genetic differentiation and the footprints of historical migrations in the Iberian Peninsula’, *Nature communications*, 10(1), p. 551.

Byrne, R. P. *et al.* (2020) ‘Dutch population structure across space, time and GWAS design’, *Nature Communications*. doi: 10.1038/s41467-020-18418-4.

Cameron, D. L., Di Stefano, L. and Papenfuss, A. T. (2019) ‘Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software’, *Nature communications*, 10(1), p. 3240.

Canela-Xandri, O., Rawlik, K. and Tenesa, A. (2018) ‘An atlas of genetic associations in UK Biobank’, *Nature genetics*, 50(11), pp. 1593–1599.

Cann, H. M. *et al.* (2002) ‘A human genome diversity cell line panel’, *Science*, 296(5566), pp. 261–262.

Cann, R. L., Stoneking, M. and Wilson, A. C. (1987) ‘Mitochondrial DNA and human evolution’, *Nature*, 325(6099), pp. 31–36.

Carvalho, C. M. (2013) ‘Replicative mechanisms for CNV formation are error prone. *Nature genetics*, *45*(11), 1319.

Carvalho, C. M., & Lupski, J. R. (2016) ‘Mechanisms underlying structural variant formation in genomic disorders’, *Nature Reviews Genetics*, *17*(4), 224.

Ceballos, F. C. *et al.* (2018) ‘Runs of homozygosity: windows into population history and trait architecture’, *Nature reviews. Genetics*, 19(4), pp. 220–234.

Chacón-Duque, J.-C. *et al.* (2018) ‘Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance’, *Nature communications*, 9(1), p. 5388.

Chaisson, M. J. P. *et al.* (2015) ‘Resolving the complexity of the human genome using single-molecule sequencing’, *Nature*, 517(7536), pp. 608–611.

Chaisson, M. J. P. *et al.* (2019) ‘Multi-platform discovery of haplotype-resolved structural variation in human genomes’, *Nature communications*, 10(1), p. 1784.

Chan, E. K. F. *et al.* (2019) ‘Human origins in a southern African palaeo-wetland and first migrations’, *Nature*, 575(7781), pp. 185–189.

Chang, C. C. *et al.* (2015) ‘Second-generation PLINK: rising to the challenge of larger and richer datasets’, *GigaScience*, 4, p. 7.

Chen, F. e*t al.* (2019) ‘A late middle pleistocene denisovan mandible from the tibetan plateau’, *Nature* 569, no. 7756 (2019): 409-412.

Chen, K. *et al.* (2009) ‘BreakDancer: an algorithm for high-resolution mapping of genomic structural variation’, *Nature methods*, 6(9), pp. 677–681.

Chen, L. *et al.* (2020) ‘Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals’, *Cell*, 180(4), pp. 677–687.e16.

Chen, S. *et al.* (2019) ‘Paragraph: a graph-based structural variant genotyper for short-read sequence data’, *Genome biology*, 20(1), p. 291.

Chen, X. *et al.* (2016) ‘Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications’, *Bioinformatics* , 32(8), pp. 1220–1222.

Chiaroni, J. *et al.* (2010) ‘The emergence of Y-chromosome haplogroup J1e among Arabic-speaking populations’, *European journal of human genetics: EJHG*, 18(3), pp. 348–353.

Cho, Y. S. *et al.* (2016) ‘An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes’, *Nature communications*, 7, p. 13637.

Choudhury, A. *et al.* (2020) ‘High-depth African genomes inform human migration and health’, *Nature*, 586(7831), pp. 741–748.

Clarkson, C. *et al.* (2017) ‘Human occupation of northern Australia by 65,000 years ago’, *Nature*, 547(7663), pp. 306–310.

Collins, R. L. *et al.* (2020) ‘A structural variation reference for medical and population genetics’, *Nature*, 581(7809), pp. 444–451.

Coop, G. *et al.* (2009) ‘The role of geography in human adaptation’, *PLoS genetics*, 5(6), p. e1000500.

Cortés-Ciriano, I. *et al.* (2020) ‘Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing’, *Nature Genetics*, pp. 331–341. doi: 10.1038/s41588-019-0576-7.

Crassard, R. *et al.* (2013) ‘Beyond the Levant: first evidence of a pre-pottery Neolithic incursion into the Nefud Desert, Saudi Arabia’, *PloS one*, 8(7), p. e68061.

Cruz, D. F., Farinha, C. M. and Swiatecka-Urban, A. (2019) ‘Unraveling the Function of Lemur Tyrosine Kinase 2 Network’, *Frontiers in pharmacology*, 10, p. 24.

Darwin, C. (1871) ‘*The Descent of Man, and Selection in Relation to Sex’*, Murray, London, UK:

Das, S. *et al.* (2016) ‘Next-generation genotype imputation service and methods’, *Nature genetics*, 48(10), pp. 1284–1287.

Dillehay, T. D. et al. (2015) ‘New archaeological evidence for an early human presence at Monte Verde, Chile’, *PloS one*, *10*(11), e0141923.

Drechsler, P. (2009) *The Dispersal of the Neolithic Over the Arabian Peninsula*. British Archaeological Reports Limited.

Durvasula, A. and Sankararaman, S. (2020) ‘Recovering signals of ghost archaic introgression in African populations’, *Science advances*, 6(7), p. eaax5097.

Eaaswarkhanth, M. *et al.* (2020) ‘Genome-Wide Selection Scan in an Arabian Peninsula Population Identifies a TNKS Haplotype Linked to Metabolic Traits and Hypertension’, *Genome biology and evolution*, 12(3), pp. 77–87.

Ebler, J. *et al.,* (2020) ‘Pangenome-based genome inference’. *bioRxiv*. doi: https://doi.org/10.1101/2020.11.11.378133

Eccleston, J. L. et al. (2012) ‘An apparent homozygous deletion in maltase-glucoamylase, a lesson in the evolution of SNP arrays’, *Molecular genetics and metabolism*, *107*(4), pp. 674-678.

Eggertsson, H. P. *et al.* (2019) ‘GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs’, *Nature communications*, 10(1), p. 5402.

Enattah, N. S. *et al.* (2002) ‘Identification of a variant associated with adult-type hypolactasia’, *Nature genetics*, 30(2), pp. 233–237.

Enattah, N. S. *et al.* (2008) ‘Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture’, *American journal of human genetics*, 82(1), pp. 57–72.

Fan, S. et al. (2016) ‘Going global by adapting local: A review of recent human adaptation’, *Science*, *354*(6308), pp. 54-59.

Flint, J. *et al.* (1986) ‘High frequencies of alpha-thalassaemia are the result of natural selection by malaria’, *Nature*, 321(6072), pp. 744–750.

Franco, J. R. *et al.* (2014) ‘Epidemiology of human African trypanosomiasis’, *Clinical epidemiology*, 6, pp. 257–275.

Fu, Q. *et al.* (2014) ‘Genome sequence of a 45,000-year-old modern human from western Siberia’, *Nature*, 514(7523), pp. 445–449.

Fu, Q. *et al.* (2016) ‘The genetic history of Ice Age Europe’, *Nature*, 534(7606), pp. 200–205.

Gallego Llorente, M. *et al.* (2015) ‘Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent’, *Science*, 350(6262), pp. 820–822.

Garrison, E. *et al.* (2018) ‘Variation graph toolkit improves read mapping by representing genetic variation in the reference’, *Nature biotechnology*, 36(9), pp. 875–879.

GenomeAsia100K Consortium (2019) ‘The GenomeAsia 100K Project enables genetic discoveries across Asia’, *Nature*, 576(7785), pp. 106–111.

Gittelman, R. M. *et al.* (2016) ‘Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments’, *Current biology: CB*, 26(24), pp. 3375–3382.

Goebel, T. et al. (2008) ‘The late Pleistocene dispersal of modern humans in the Americas’, *Science*, *319*(5869), pp. 1497-1502.

Gordon, D. *et al.* (2016) ‘Long-read sequence assembly of the gorilla genome’, *Science*, 352(6281), p. aae0344.

Green, R. E. *et al.* (2010) ‘A draft sequence of the Neandertal genome’, *Science*, 328(5979), pp. 710–722.

Gronau, I. *et al.,* (2011). ‘Bayesian inference of ancient human demography from individual genome sequences’. *Nature genetics*, *43*(10), 1031.

Groucutt, H. S. *et al.* (2018) ‘Homo sapiens in Arabia by 85,000 years ago’, *Nature Ecology & Evolution*, pp. 800–809. doi: 10.1038/s41559-018-0518-2.

Haak, W. *et al.* (2015) ‘Massive migration from the steppe was a source for Indo-European languages in Europe’, *Nature*, 522(7555), pp. 207–211.

Haber, M. *et al.* (2013) ‘Genome-wide diversity in the levant reveals recent structuring by culture’, *PLoS genetics*, 9(2), p. e1003316.

Haber, M. *et al.* (2017) ‘Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences’, *American journal of human genetics*, 101(2), pp. 274–282.

Haber, M. *et al.* (2019) ‘A Transient Pulse of Genetic Admixture from the Crusaders in the Near East Identified from Ancient Genome Sequences’, *American journal of human genetics*, 104(5), pp. 977–984.

Haber, M. *et al.* (2019) ‘Insight into the genomic history of the Near East from whole-genome sequences and genotypes of Yemenis. *BioRxiv*, p.749341. doi: https://doi.org/10.1101/749341

Haber, M. *et al.* (2020) ‘A Genetic History of the Near East from an aDNA Time Course Sampling Eight Points in the Past 4,000 Years’, *American journal of human genetics*, 107(1), pp. 149–157.

Hammer, M. F. *et al.* (2011) ‘Genetic evidence for archaic admixture in Africa’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(37), pp. 15123–15128.

Handsaker, R. E. *et al.* (2015) ‘Large multiallelic copy number variations in humans’, *Nature genetics*, 47(3), pp. 296–303.

Hebbring, S. J., Moyer, A. M. and Weinshilboum, R. M. (2008) ‘Sulfotransferase gene copy number variation: pharmacogenetics and function’, *Cytogenetic and genome research*, 123(1-4), pp. 205–210.

Hellenthal, G. *et al.* (2014) ‘A genetic atlas of human admixture history’, *Science*, 343(6172), pp. 747–751.

Hershkovitz, I. *et al.* (2018) ‘The earliest modern humans outside Africa’, *Science*, 359(6374), pp. 456–459.

Hickey, G. *et al.* (2020) ‘Genotyping structural variants in pangenome graphs using the vg toolkit’, *Genome biology*, 21(1), p. 35.

Higham, T. *et al.* (2011) ‘The earliest evidence for anatomically modern humans in northwestern Europe’, *Nature*, 479(7374), pp. 521–524.

Hilbert, Y. H. et al. (2015) ‘Archaeological evidence for indigenous human occupation of Southern Arabia at the Pleistocene/ Holocene transition: The case of al-Hatab in Dhofar, Southern Oman’, *Paléorient,* pp. 31–49. doi: 10.3406/paleo.2015.5674

Hirschfeld, L. and Hirschfeld, H. (1919) ‘SEROLOGICAL DIFFERENCES BETWEEN THE BLOOD OF DIFFERENT RACES.: THE RESULT OF RESEARCHES ON THE MACEDONIAN FRONT’. *The Lancet*, *194*(5016), pp.675-679.

Ho, S. S., Urban, A. E. and Mills, R. E. (2020) ‘Structural variation in the sequencing era’, *Nature reviews. Genetics*, 21(3), pp. 171–189.

Homburger, J. R. *et al.* (2015) ‘Genomic Insights into the Ancestry and Demographic History of South America’, *PLoS genetics*, 11(12), p. e1005602.

Hsieh, P. *et al.* (2019) ‘Adaptive archaic introgression of copy number variants and the discovery of previously unknown human genes’, *Science*, 366(6463). doi: 10.1126/science.aax2083.

Huerta-Sánchez, E. *et al.* (2014) ‘Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA’, *Nature*, 512(7513), pp. 194–197.

Imtiaz, F. *et al.* (2007) ‘The T/G 13915 variant upstream of the lactase gene (LCT) is the founder allele of lactase persistence in an urban Saudi population’, *Journal of medical genetics*, 44(10), p. e89.

International HapMap Consortium (2005) ‘A haplotype map of the human genome’, *Nature*, 437(7063), pp. 1299–1320.

Jablonski, N. G. and Chaplin, G. (2017) ‘The colours of humanity: the evolution of pigmentation in the human lineage’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 372(1724). doi: 10.1098/rstb.2016.0349.

Jacobs, G. S. *et al.* (2019) ‘Multiple Deeply Divergent Denisovan Ancestries in Papuans’, *Cell*, 177(4), pp. 1010–1021.e32.

Jakobsson, M. *et al.* (2008) ‘Genotype, haplotype and copy-number variation in worldwide human populations’, *Nature*, 451(7181), pp. 998–1003.

Jenkins, D. L et al. (2012) ‘Clovis age Western Stemmed projectile points and human coprolites at the Paisley Caves’, *Science*, *337*(6091), pp. 223-228.

Johnstone, R. W., Frew, A. J., & Smyth, M. J. (2008). ‘The TRAIL apoptotic pathway in cancer onset, progression and therapy’. *Nature Reviews Cancer*, *8*(10), 782-798.

Jones, E. R. *et al.* (2015) ‘Upper Palaeolithic genomes reveal deep roots of modern Eurasians’, *Nature communications*, 6, p. 8912.

Jónsson, H. *et al.* (2017) ‘Parental influence on human germline de novo mutations in 1,548 trios from Iceland’, *Nature*, 549(7673), pp. 519–522.

Joseph, T. A. and Pe’er, I. (2019) ‘Inference of Population Structure from Time-Series Genotype Data’, *American journal of human genetics*, 105(2), pp. 317–333.

Kaiser, J. (2016) ‘Qatar’s genome effort slowly gears up’, *Science*, 354(6317), p. 1220.

Kawa, S. *et al.* (2006) ‘Azoospermia in mice with targeted disruption of the Brek/Lmtk2 (brain-enriched kinase/lemur tyrosine kinase 2) gene’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(51), pp. 19344–19349.

Kehr, B. *et al.* (2017) ‘Diversity in non-repetitive human sequences not found in the reference genome’, *Nature genetics*, 49(4), pp. 588–593.

Kitchen, A. *et al.* (2009) ‘Bayesian phylogenetic analysis of Semitic languages identifies an Early Bronze Age origin of Semitic in the Near East’, *Proceedings. Biological sciences / The Royal Society*, 276(1668), pp. 2703–2710.

König, R. *et al.* (2008) ‘Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication’, *Cell*, 135(1), pp. 49–60.

Kosugi, S. *et al.* (2019) ‘Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing’, *Genome biology*, 20(1), p. 117.

Kronenberg, Z. N. *et al.* (2018) ‘High-resolution comparative analysis of great ape genomes’, *Science*, 360(6393). doi: 10.1126/science.aar6343.

Kuhlwilm, M. et al. (2016) ‘Ancient gene flow from early modern humans into Eastern Neanderthals’, *Nature*, *530*(7591), pp. 429-433.

Kuijpers, T. W. *et al.* (2010) ‘CD20 deficiency in humans results in impaired T cell-independent antibody responses’, *The Journal of clinical investigation*, 120(1), pp. 214–222.

Kwiatkowski, D. P. (2005) ‘How malaria has affected the human genome and what human genetics can teach us about malaria’, *The American Journal of Human Genetics*, *77*(2), 171-192.

Lamason, R. L. et al. (2005) ‘SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans’, *Science*, *310*(5755), pp. 1782-1786.

Lawson, D. J. *et al.* (2012) ‘Inference of population structure using dense haplotype data’, *PLoS genetics*, 8(1), p. e1002453.

Lazaridis, I. *et al.* (2014) ‘Ancient human genomes suggest three ancestral populations for present-day Europeans’, *Nature*, 513(7518), pp. 409–413.

Lazaridis, I. *et al.* (2016) ‘Genomic insights into the origin of farming in the ancient Near East’, *Nature*, 536(7617), pp. 419–424.

Lek, M. *et al.* (2016) ‘Analysis of protein-coding genetic variation in 60,706 humans’, *Nature*, 536(7616), pp. 285–291.

Leslie, S. *et al.* (2015) ‘The fine-scale genetic structure of the British population’, *Nature*, 519(7543), pp. 309–314.

Li, H. (2011) ‘A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data’, *Bioinformatics* , 27(21), pp. 2987–2993.

Li, H. (2015) ‘FermiKit: assembly-based variant calling for Illumina resequencing data’, *Bioinformatics* , 31(22), pp. 3694–3696.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics* , 25(14), pp. 1754–1760.

Li, H. and Durbin, R. (2011) ‘Inference of human population history from individual whole-genome sequences’, *Nature*, 475(7357), pp. 493–496.

Li, H., Feng, X. and Chu, C. (2020) ‘The design and construction of reference pangenome graphs with minigraph’, *Genome biology*, 21(1), p. 265.

Li, J. Z. *et al.* (2008) ‘Worldwide human relationships inferred from genome-wide patterns of variation’, *Science*, 319(5866), pp. 1100–1104.

Lindo, J. *et al.* (2018) ‘The genetic prehistory of the Andean highlands 7000 years BP though European contact’, *Science advances*, 4(11), p. eaau4921.

Lipson, M. *et al.* (2020) ‘Ancient West African foragers in the context of African population history’, *Nature*, 577(7792), pp. 665–670.

Liu, W. *et al.* (2015) ‘The earliest unequivocally modern humans in southern China’, *Nature*, 526(7575), pp. 696–699.

Loh, P.-R. *et al.* (2013) ‘Inferring admixture histories of human populations using linkage disequilibrium’, *Genetics*, 193(4), pp. 1233–1254.

Loh, P.-R., Palamara, P. F. and Price, A. L. (2016) ‘Fast and accurate long-range phasing in a UK Biobank cohort’, *Nature genetics*, 48(7), pp. 811–816.

Lübbers, J. et al. (2018). ‘Modulation of immune tolerance via siglec-sialic acid interactions’. *Frontiers in immunology*, *9*, p.2807.

Malaspinas, A.-S. *et al.* (2016) ‘A genomic history of Aboriginal Australia’, *Nature*, 538(7624), pp. 207–214.

Malik, M. *et al.* (2005) ‘Glucose intolerance and associated factors in the multi-ethnic population of the United Arab Emirates: results of a national survey’, *Diabetes research and clinical practice*, 69(2), pp. 188–195.

Mallick, S. *et al.* (2016) ‘The Simons Genome Diversity Project: 300 genomes from 142 diverse populations’, *Nature*, 538(7624), pp. 201–206.

Manga, P. *et al.* (2001) ‘In Southern Africa, brown oculocutaneous albinism (BOCA) maps to the OCA2 locus on chromosome 15q: P-gene mutations identified’, *American journal of human genetics*, 68(3), pp. 782–787.

Manolio, T. A. *et al.* (2009) ‘Finding the missing heritability of complex diseases’, *Nature*, 461(7265), pp. 747–753.

Maples, B. K. *et al.* (2013) ‘RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference’, *American journal of human genetics*, 93(2), pp. 278–288.

Maretty, L. *et al.* (2017) ‘Sequencing and de novo assembly of 150 genomes from Denmark as a population reference’, *Nature*, 548(7665), pp. 87–91.

Marks, P. *et al.* (2019) ‘Resolving the full spectrum of human genome variation using Linked-Reads’, *Genome research*, 29(4), pp. 635–645.

Marshall, M. J. E., Stopforth, R. J. and Cragg, M. S. (2017) ‘Therapeutic Antibodies: What Have We Learnt from Targeting CD20 and Where Are We Going?’, *Frontiers in immunology*, 8, p. 1245.

Mathieson, S. and Mathieson, I. (2018) ‘FADS1 and the Timing of Human Adaptation to Agriculture’, *Molecular biology and evolution*, 35(12), pp. 2957–2970.

McCarthy, S. *et al.* (2016) ‘A reference panel of 64,976 haplotypes for genotype imputation’, *Nature genetics*, 48(10), pp. 1279–1283.

McColl, H. *et al.* (2018) ‘The prehistoric peopling of Southeast Asia’, *Science*, 361(6397), pp. 88–92.

McDougall, I., Brown, F. H. and Fleagle, J. G. (2005) ‘Stratigraphic placement and age of modern humans from Kibish, Ethiopia’, *Nature*, 433(7027), pp. 733–736.

McInnes, L. *et al.* (2018) ‘UMAP: Uniform Manifold Approximation and Projection’, *Journal of Open Source Software*, p. 861. doi: 10.21105/joss.00861.

McKenna, A. *et al.* (2010) ‘The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data’, *Genome research*, 20(9), pp. 1297–1303.

McLaren, W. *et al.* (2016) ‘The Ensembl Variant Effect Predictor’, *Genome biology*, 17(1), p. 122.

Mendez, F. L. *et al.* (2013) ‘An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree’, *American journal of human genetics*, 92(3), pp. 454–459.

Meyer, M. *et al.* (2012) ‘A high-coverage genome sequence from an archaic Denisovan individual’, *Science*, 338(6104), pp. 222–226.

Miga, K. H. *et al.* (2020) ‘Telomere-to-telomere assembly of a complete human X chromosome’, *Nature*, 585(7823), pp. 79–84.

Miller, L. H. *et al.* (1976) ‘The resistance factor to Plasmodium vivax in blacks: the Duffy-blood-group genotype, FyFy’, *New England Journal of Medicine*, *295*(6), 302-304.

Mirzai, B. A., Montana, I. M. and Lovejoy, P. E. (2009) ‘*Slavery, Islam and Diaspora’*. Africa Research and Publications.

Moorjani, P. *et al.* (2016) ‘A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years’, *Proceedings of the National Academy of Sciences of the United States of America*, 113(20), pp. 5652–5657.

Moreno-Estrada, A. *et al.* (2014) ‘Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits’, *Science*, 344(6189), pp. 1280–1285.

Myers, S. *et al.* (2008) ‘A common sequence motif associated with recombination hot spots and genome instability in humans’, *Nature genetics*, 40(9), pp. 1124–1129.

Narasimhan, V. M. *et al.* (2019) ‘The formation of human populations in South and Central Asia’, *Science*, 365(6457). doi: 10.1126/science.aat7487.

Neph, S. *et al.* (2012) ‘BEDOPS: high-performance genomic feature operations’, *Bioinformatics* , 28(14), pp. 1919–1920.

Nichols, B. L. *et al.* (2003) ‘The maltase-glucoamylase gene: common ancestry to sucrase-isomaltase with complementary starch digestion activities’, *Proceedings of the National Academy of Sciences of the United States of America*, 100(3), pp. 1432–1437.

Nielsen, R. (2017) ‘Tracing the peopling of the world through genomics’, *Nature*, *541*(7637), pp. 302-310.

Novembre, J. *et al.* (2008) ‘Genes mirror geography within Europe’, *Nature*, 456(7218), pp. 98–101.

Nowell P. and Hungerford D (1960). ‘A minute chromosome in human chronic granulocytic leukemia’. *Science* 132:1497.

Offermanns, S. (2017) ‘Hydroxy-Carboxylic Acid Receptor Actions in Metabolism’, *Trends in endocrinology and metabolism: TEM*, 28(3), pp. 227–236.

Olalde, I. *et al.* (2019) ‘The genomic history of the Iberian Peninsula over the past 8000 years’, *Science*, 363(6432), pp. 1230–1234.

Pagani, L. *et al.* (2012) ‘Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool’, *American journal of human genetics*, 91(1), pp. 83–96.

Pagani, L. *et al.* (2016) ‘Genomic analyses inform on migration events during the peopling of Eurasia’, *Nature*, 538(7624), pp. 238–242.

Patterson, N. *et al.* (2012) ‘Ancient admixture in human history’, *Genetics*, 192(3), pp. 1065–1093.

Patterson, N., Price, A. L. and Reich, D. (2006) ‘Population structure and eigenanalysis’, *PLoS genetics*, 2(12), p. e190.

Pedersen, B. S. *et al.* (2017) ‘Indexcov: fast coverage quality control for whole-genome sequencing’, *GigaScience*. doi: 10.1093/gigascience/gix090.

Pedersen, M. W. et al. (2016) ‘Postglacial viability and colonization in North America’s ice-free corridor', *Nature*, *537*(7618), pp. 45-49.

Perry, G. H. *et al.* (2007) ‘Diet and the evolution of human amylase gene copy number variation’, *Nature genetics*, 39(10), pp. 1256–1260.

Petr, M. et al. (2019) ‘Limits of long-term selection against Neandertal introgression’, *Proceedings of the National Academy of Sciences*, *116*(5), pp. 1639-1644.

Petr, M. et al. (2020) ‘The evolutionary history of Neanderthal and Denisovan Y chromosomes’, *Science*, *369*(6511), pp.1653-1656.

Petraglia, M. D. *et al.* (2020) ‘Human responses to climate and ecosystem change in ancient Arabia’, *Proceedings of the National Academy of Sciences of the United States of America*, 117(15), pp. 8263–8270.

Pickrell, J. K. *et al.* (2014) ‘Ancient west Eurasian ancestry in southern and eastern Africa’, *Proceedings of the National Academy of Sciences of the United States of America*, 111(7), pp. 2632–2637.

Plagnol, V. and Wall, J. D. (2006) ‘Possible ancestral structure in human populations’, *PLoS genetics*, 2(7), p. e105.

Porubsky, D. *et al.* (2020) ‘Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads’, *Nature biotechnology*. doi: 10.1038/s41587-020-0719-5.

Porubsky, D. et al. (2021) ‘ Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads’, *Nature biotechnology*, *39*(3), pp. 302-308.

Posth, C. et al. (2017) ‘Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals’, *Nature communications*, *8*(1), pp.1-9.

Price, A. L. *et al.* (2006) ‘Principal components analysis corrects for stratification in genome-wide association studies’, *Nature genetics*, 38(8), pp. 904–909.

Pritchard, J. K., Pickrell, J. K., & Coop, G. (2010) ‘The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation’, *Current biology*, *20*(4), pp. 208-215.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000) ‘Inference of population structure using multilocus genotype data’, *Genetics*, 155(2), pp. 945–959.

Prüfer, K. *et al.* (2014) ‘The complete genome sequence of a Neanderthal from the Altai Mountains’, *Nature*, 505(7481), pp. 43–49.

Prüfer, K. *et al.* (2017) ‘A high-coverage Neandertal genome from Vindija Cave in Croatia’, *Science*, 358(6363), pp. 655–658.

Raghavan, M. *et al.* (2014) ‘Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans’, *Nature*, 505(7481), pp. 87–91.

Raghavan, M. *et al.* (2015) ‘POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans’, *Science*, 349(6250), p. aab3884.

Ranji, A. and Boris-Lawrie, K. (2010) 'RNA helicases: emerging roles in viral replication and the host innate response', *RNA biology*, 7(6), pp. 775–787.

Redon, R. *et al.* (2006) 'Global variation in copy number in the human genome', *Nature*, 444(7118), pp. 444–454.

Reich, D. *et al.* (2009) 'Reconstructing Indian population history', *Nature*, 461(7263), pp. 489–494.

Reich, D. *et al.* (2010) 'Genetic history of an archaic hominin group from Denisova Cave in Siberia', *Nature*, 468(7327), pp. 1053–1060.

Reich, D. et al. (2011) 'Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania', *The American Journal of Human Genetics*, *89*(4), pp. 516-528.

Richter, D. *et al.* (2017) 'The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age', *Nature*, 546(7657), pp. 293–296.

Rodriguez-Flores, J. L. *et al.* (2016) 'Indigenous Arabs are descendants of the earliest split from ancient Eurasian populations', *Genome research*, 26(2), pp. 151–162.

Rosenberg, N. A. (2006) 'Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives', *Annals of human genetics*, 70(Pt 6), pp. 841–847.

Rosenberg, N. A. *et al.* (2002) 'Genetic structure of human populations', *Science*, 298(5602), pp. 2381–2385.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5463–5467.

Sankararaman, S. *et al.* (2014) 'The genomic landscape of Neanderthal ancestry in present-day humans', *Nature*, 507(7492), pp. 354–357.

Sankararaman, S. *et al.* (2016) 'The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans', *Current biology: CB*, 26(9), pp. 1241–1247.

Scally, A. and Durbin, R. (2012) ‘Revising the human mutation rate: implications for understanding human evolution’, *Nature reviews. Genetics*, 13(10), pp. 745–753.

Scerri, E. M. L. *et al.* (2018) ‘Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter?’, *Trends in ecology & evolution*, 33(8), pp. 582–594.

Schiffels, S. and Durbin, R. (2014) ‘Inferring human population size and separation history from multiple genome sequences’, *Nature genetics*, 46(8), pp. 919–925.

Schlebusch, C. M. *et al.* (2017) ‘Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago’, *Science*, 358(6363), pp. 652–655.

Schneider, V. A. *et al.* (2017) ‘Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly’, *Genome research*, 27(5), pp. 849–864.

Scott, E. M. *et al.* (2016) ‘Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery’, *Nature genetics*, 48(9), pp. 1071–1076.

Seguin-Orlando, A. *et al.* (2014) ‘Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years’, *Science*, 346(6213), pp. 1113–1118.

Sharp, W. D. and Paces, J. B. (2018) ‘Comment on “The earliest modern humans outside Africa”’, *Science*. doi: 10.1126/science.aat6598.

Sherman, R. M. *et al.* (2019) ‘Assembly of a pan-genome from deep sequencing of 910 humans of African descent’, *Nature genetics*, 51(1), pp. 30–35.

Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) ‘The Missing Diversity in Human Genetic Studies’, *Cell*, 177(4), p. 1080.

Skoglund, P. *et al.* (2017) ‘Reconstructing Prehistoric African Population Structure’, *Cell*, 171(1), pp. 59–71.e21.

Skov, L. *et al.* (2018) ‘Detecting archaic introgression using an unadmixed outgroup’, *PLoS genetics*, 14(9), p. e1007641.

Smith, A. B. and Hajduk, S. L. (1995) ‘Identification of haptoglobin as a natural inhibitor of trypanocidal activity in human serum’, *Proceedings of the National Academy of Sciences of the United States of America*, 92(22), pp. 10262–10266.

Sohail, M. *et al.* (2019) ‘Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies’, *eLife*, 8. doi: 10.7554/eLife.39702.

Speidel, L. *et al.* (2019) ‘A method for genome-wide genealogy estimation for thousands of samples’, *Nature genetics*, 51(9), pp. 1321–1329.

Stefansson, H. *et al.* (2005) ‘A common inversion under selection in Europeans’, *Nature genetics*, 37(2), pp. 129–137.

Stern, A. J. *et al.* (2021) ‘Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies’, *American journal of human genetics*. doi: 10.1016/j.ajhg.2020.12.005.

Stern, A. J., Wilton, P. R. and Nielsen, R. (2019) ‘An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data’, *PLoS genetics*, 15(9), p. e1008384.

Stewart, M. *et al.* (2020) ‘Human footprints provide snapshot of last interglacial ecology in the Arabian interior’, *Science advances*, 6(38). doi: 10.1126/sciadv.aba8940.

Sudmant, P. H., Mallick, S. *et al.,* (2015b) ‘Global diversity, population stratification, and selection of human copy-number variation’, *Science*, 349(6253), p. aab3761.

Sudmant, P. H., Rausch, T. *et al.,* (2015a) ‘An integrated map of structural variation in 2,504 human genomes’, *Nature*, 526(7571), pp. 75–81.

Terhorst, J., Kamm, J. A. and Song, Y. S. (2017) ‘Robust and scalable inference of population history from hundreds of unphased whole genomes’, *Nature genetics*, 49(2), pp. 303–309.

The GTEx Consortium (2020) ‘The GTEx Consortium atlas of genetic regulatory effects across human tissues’. *Science*, *369*(6509), pp.1318-1330.

Thorvaldsdóttir, H., Robinson, J. T. and Mesirov, J. P. (2013) 'Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration', *Briefings in bioinformatics*, 14(2), pp. 178–192.

Tishkoff, S. A. *et al.* (2007) 'Convergent adaptation of human lactase persistence in Africa and Europe', *Nature genetics*, 39(1), pp. 31–40.

Turchin, M. C. *et al.* (2012) 'Evidence of widespread selection on standing variation in Europe at height-associated SNPs', *Nature genetics*, 44(9), pp. 1015–1019.

Uerpmann, H.-P., Potts, D. T. and Uerpmann, M. (2010) 'Holocene (Re-)Occupation of Eastern Arabia', *The Evolution of Human Populations in Arabia*, pp. 205–214. doi: 10.1007/978-90-481-2719-1_15.

van de Loosdrecht, M. *et al.* (2018) 'Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations', *Science*, 360(6388), pp. 548–552.

Watanabe, K. *et al.* (2019) 'A global overview of pleiotropy and genetic architecture in complex traits', *Nature genetics*, 51(9), pp. 1339–1348.

Weischenfeldt, J. et al. (2013) 'Phenotypic impact of genomic structural variation: insights from and for human disease', *Nature Reviews Genetics*, *14*(2), pp.125-138.

Weisenfeld, N. I. *et al.* (2017) 'Direct determination of diploid genome sequences', *Genome research*, 27(5), pp. 757–767.

Weiss, H. *et al.* (1993) 'The genesis and collapse of third millennium north mesopotamian civilization', *Science*, 261(5124), pp. 995–1004.

Weiss, L. A. *et al.* (2008) 'Association between microdeletion and microduplication at 16p11.2 and autism', *The New England journal of medicine*, 358(7), pp. 667–675.

Wilde, S. et al. (2014) 'Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y', *Proceedings of the National Academy of Sciences*, 111(13), pp. 4832-4837.

Wong, K. H. Y., Levy-Sakin, M. and Kwok, P.-Y. (2018) ‘De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations’, *Nature communications*, 9(1), p. 3040.

Yamanaka, M. *et al.* (2009) ‘Deletion polymorphism of SIGLEC14 and its functional implications’, *Glycobiology*, 19(8), pp. 841–846.

Yi, X *et al.* (2010) ‘Sequencing of 50 human exomes reveals adaptation to high altitude’, *Science*, *329*(5987), pp.75-78.

Zaidi, A. A. and Mathieson, I. (2020) ‘Demographic history mediates the effect of stratification on polygenic scores’, *eLife*, 9. doi: 10.7554/eLife.61548.

Zalloua, P. A. *et al.* (2008) ‘Y-chromosomal diversity in Lebanon is structured by recent historical events’, *American journal of human genetics*, 82(4), pp. 873–882.