THE UNIVERSITY OF CHICAGO

THEORY AND METHODS FOR INVESTIGATING THE SPATIO-TEMPORAL STRUCTURE OF HUMAN GENETIC DIVERSITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
ARJUN APPANNA BIDDANDA

CHICAGO, ILLINOIS
DECEMBER 2020

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

There are several fundamental challenges in modeling population genetic data that is distributed along spatial and temporal dimensions. Spatial population genetic data is increasingly commonplace across a wide number of organisms (Wasser et al., 2004; Pagani et al., 2016). Over the past decade, large global sequencing studies in humans have increasingly sampled a larger proportion of genetic diversity around the world. (e.g. Bergström et al., 2020; Auton et al., 2015). Improvements in the sequencing of degraded DNA from ancient human remains have also made it possible to directly consider population genetic data at different points in time (e.g. Skoglund and Mathieson, 2018). This expansion in the dimensionality of population genetic datasets poses challenges for data visualization and statistical inference.

Historical approaches to studying spatio-temporal patterns of human genetic diversity have largely focused on summary statistics or lower-dimensional representations of the data. For example, in modeling gene expression datasets using hierarchical clustering methods, Eisen et al. (1998) helped to show the utility of low-dimensional summaries for understanding biological processes. Summary statistics and current lower-dimensional representations of multi-dimensional data are attractive due to their simplicity, but carry their own challenges.

A popular approach to visualizing and reasoning about spatial population genetic datasets is to consider summary statistics of the multi-population genetic data. For the majority of this dissertation, we consider spatially distributed populations to be sufficiently described by multiple discrete (yet connected) populations. Many summary statistics are used in the context of multi-population datasets, such as F_{ST}(Bhatia et al., 2013). F_{ST} is a useful summary statistic in the multi-population setting as it is related to the shared genealogical branch lengths between samples (Slatkin, 1991) and is a measure of genetic differentiation between two populations (Bhatia et al., 2013). Summary statistics such as F_{ST} are useful

in the comparison of multiple populations, in that they are able to show which populations are more similar to one another averaged across all genetic variants. However, they are also limited by their simplicity. For example, using F_{ST} cannot reflect absolute patterns of variant frequencies between populations. Knowing the F_{ST} at a single variant between two populations only constrains the *relative* differentiation in frequency of the variant between two populations, and does not inform us on the absolute frequency of the variant in each population. This lack of correspondence between summary statistics and the absolute frequency of genetic variants is a feature of many summaries of human population genetic structure (e.g. principal component loadings).

A more detailed summary of multi-population genetic data is the joint site frequency spectrum (SFS). The one-population SFS is a histogram of the derived allele count across all polymorphic sites in a single sampled group. The joint SFS is analogously a P-dimensional histogram of the allele counts across P defined populations. The joint SFS is a rich lower-dimensional summary statistic of the multi-population genetic data and is frequently used for demographic inference of joint population histories (Gutenkunst et al., 2009; Kamm et al., 2016; Jouganous et al., 2017; Kamm et al., 2020). While the joint SFS does not adequately capture the effects of linkage disequilibrium, it is often considered a robust summary to capture the effects of demographic history and migration between populations (Kamm et al., 2020).

Latent genealogies are another framework to model multi-population genetic data. Briefly, coalescent theory is concerned with the stochastic process of ancestral relationships between a collection of samples (Wakeley, 2009). For a single locus, this is represented as a single binary tree, where the branches reflecting individual samples are brought together through *coalescence events* to reflect their shared common ancestry. This stochastic process of ancestral lineages brought together into specific lines of ancestry is known as Kingman's coalescent, or simply "the coalescent" (Kingman, 1982; Hudson, 1985, 1990). The coalescent is an attrac-

tive statistical model, since mutations placed on the branches of the latent genealogy directly reflect mutations observed in genetic data (Figure 1.1). Since the coalescent is concerned with modeling the ancestry of a *sample* rather than an entire *population*, it is a more direct model of the available data. Coalescent models for single non-recombining loci are quite flexible to several demographic extensions such as population subdivision, varying population size history, and serial sampling (Wakeley, 2009; Forsberg et al., 2005; Duforet-Frebourg and Slatkin, 2016).

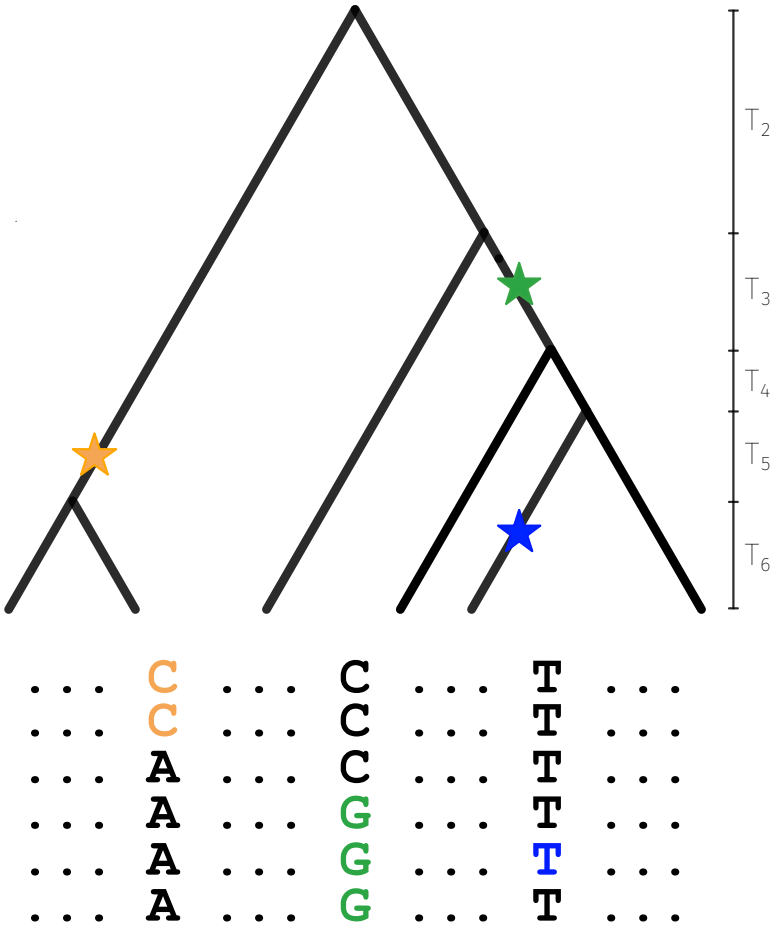

Figure 1.1: The coalescent process for $n = 6$ samples. Each mutation is placed on the coalescent genealogy and is assumed to lead to a unique new varying site in the locus (e.g. the infinite-sites assumption). The time-intervals T_k on the right reflect the amount of time during which there are k ancestral lineages present in the process.

However, these previous modeling approaches are not without their limitations in devel-

oping an understanding of spatial and temporal population structure. For example, summary statistics are not necessarily able to reflect patterns of absolute variation, as in the case of F_{ST}. More detailed summaries of genetic data such as the joint SFS are often high dimensional ($N_{pops} > 3$) making visual representations difficult to obtain, and limiting their utility for exploratory data analysis. For these reasons, we develop a new approach in Chapter 2 for the visualization of multi-population allele frequency data.

There are additional challenges imposed when data are temporally structured, as in the case with human ancient DNA datasets. Over the past decade, technological advances have enabled human genetics researchers to obtain endogenous DNA from historical remains, allowing for a direct view into the human past (Skoglund and Mathieson, 2018). This added temporal structure within population genetic data has led to an improved understanding of human population history, (Nielsen et al., 2017; Pickrell and Reich, 2014) and adaptive genetic variation (e.g. Mathieson et al., 2015). However, this new temporal dimension of the data has posed several challenges for previous theoretical models to investigate population genetic variation.

In the single-locus coalescent framework, theory has previously been developed for the analysis of serially sampled data. Specifically Rodrigo and Felsenstein (1999) focused on likelihood calculations of observed variation from coalescent trees with serially sampled HIV datasets, finding widespread use in the context of viral phylodynamics (e.g. Drummond et al., 2005). More recently, Forsberg et al. (2005) considered the effects of different serial sampling schemes on the expected joint properties of variants such as the number of variants private to a given ancient sample. To our knowledge, the most recent attempt to jointly model samples at multiple time-points is the method introduced by Kamm et al. (2020), which performs demographic inference on the joint SFS, while accounting for different sampling times analytically.

In contrast to single-locus coalescent models, multi-locus models of genealogies have not

been sufficiently developed to accommodate serial sampling. Two-locus models of genealogical ancestry have been previously used for inference of the population scaled recombination rate from patterns of linkage disequilibrium and the detection of recombination hotspots in humans (Hudson, 2001; McVean et al., 2004). There are sampling approaches to calculate two-locus likelihoods for data under varying demography (Kamm et al., 2016) and serial-sampling (Dialdestoro et al., 2016), but no attempt to incorporate serial sampling directly into the analytical theory of the two-locus ancestral process (Simonsen and Churchill, 1997; Richard Durrett, 2002). By directly incorporating serial samples, it is possible to understand the effects of time-separation on patterns of linkage disequilibrium and other summaries of the joint genealogies at two loci.

An alternative model of multi-locus genealogical ancestry is the haplotype-copying model proposed by Li and Stephens (2003). The haplotype-copying model is central to modern statistical genetics procedures such as genotype imputation (e.g. Howie et al., 2009) and haplotype phasing (e.g. Loh et al., 2016), and is connected to more formal models of genealogical ancestry (Paul et al., 2011). However, the inclusion of samples at multiple timepoints has not been explored in the context of the haplotype copying model.

This dissertation addresses several problems within the broader challenge of modeling spatio-temporal population genetic data. In Chapter 2, we develop a framework to show the absolute frequencies of genetic variation across global populations. We then turn to the development of two-locus coalescent models for serially sampled data in Chapter 3, with implications for patterns of linkage disequilibrium and imputation of ancient samples. Chapter 4 focuses on the population genetic history of the Kodava population in south western India, as a way to further contextualize population genetic diversity in south India, a region under-sampled in global human genetic datasets.

CHAPTER 2

GEOGRAPHIC PATTERNS OF HUMAN ALLELE FREQUENCY VARIATION: A VARIANT-CENTRIC PERSPECTIVE

2.1 Abstract[1]

A key challenge in human genetics is to describe and understand the distribution of human genetic variation. Often genetic variation is described by showing relationships among populations or individuals, in each case drawing inferences over a large number of variants. Here, we present an alternative representation of human genetic variation that reveals the relative abundance of different allele frequency patterns across populations. This approach allows viewers to easily see several key features of human genetic structure: (1) most variants are rare and geographically localized, (2) variants that are common in a single geographic region are vastly more likely to be shared across the globe than to be private to that region, (3) African populations have more diversity than other regions, and (4) when two individuals differ, it is most often at nucleotide sites carrying common variants across all global populations, regardless of whether the individuals are from the same region or different regions. By comparing the observations to theoretical models, we show that the main features of the data can be explained by the Recent-African-Origin model of modern human populations with subsequent gene flow. Overall, our visualizations clarify the major geographic patterns of human variation and the evolutionary history that shaped them.

1. Citation for chapter: **Geographic patterns of human allele frequency variation: a variant-centric perspective.** Arjun Biddanda, Daniel P. Rice, John Novembre *bioRxiv* 2020 (doi: https://doi.org/10.1101/2020.07.01.182311)

2.2 Introduction

Understanding human genetic variation, including its origins and its consequences, is one of the long-standing challenges of human biology. A first step is to learn the fundamental aspects of how human genomes vary within and between populations. For instance, how often do variants have an allele at high frequency in one narrow region of the world that is absent everywhere else? For answering many applied questions, we need to know how many variants show any particular geographic pattern in their allele frequencies. In order to answer such questions, one needs to measure the frequencies of many alleles around the world without the ascertainment biases that affect genotyping arrays and other probe-based technologies (Li et al., 2008; International HapMap Consortium, 2005). Recent whole-genome sequencing studies (Mallick et al., 2016; Bergström et al., 2020; Fairley et al., 2020) provide this data, and thus present an opportunity for new perspectives on human variation.

However, large genetic data sets present a visualization challenge: how does one show the allele frequency patterns of millions of variants? Plotting a joint site frequency spectrum (SFS) is one approach that efficiently summarizes allele frequencies and can be carried out for data from two or three populations (Gutenkunst et al., 2009). For more than three populations, one must resort to showing multiple combinations of two or three-population SFSs. This representation becomes unwieldy to interpret for more than three populations, and cannot represent information about the joint distribution of allele frequencies across all populations. Thus, we need visualizations that intuitively summarize allele frequency variation across several populations.

New visualization techniques also have the potential to improve population genetics education and research. Many commonly used analysis methods, such as principal components analysis (PCA) or admixture analysis, do a poor job of conveying absolute levels of differentiation (McVean, 2009; Lawson et al., 2018). Observing the genetic clustering of individuals into groups can give a misleading impression of "deep" differentiation between populations,

even when the signal comes from subtle allele frequency deviations at a large number of loci (Patterson et al., 2006; McVean, 2009; Novembre and Peter, 2016). Similar misconceptions can arise from observing how direct-to-consumer genetic ancestry tests apportion ancestry to broad continental regions. One may mistakenly surmise from the output of these methods that most human alleles must be sharply divided among regional groups, such that each allele is common in one continental region and absent in all others. Similarly, one might mistakenly conclude that two humans from different regions of the world differ mainly due to alleles that are restricted to each region. Such misconceptions can impact researchers and the broader public alike. All of these misconceptions potentially can be avoided with visualizations of population genetic data that make typical allele frequency patterns more transparent.

Here, we develop a new representation of population genetic data and apply it to the New York Genome Center deep coverage sequencing data from the 1000 Genomes Project (1KGP) samples (Auton et al., 2015; Fairley et al., 2020) In essence, our approach represents a multi-population joint SFS with coarsely binned allele frequencies. It trades precision in frequency for the ability to show several populations on the same plot. Overall, we aimed to create a visualization that is easily understandable and useful for pedagogy. As we will show, the visualizations reveal with relative ease many known important features of human genetic variation and evolutionary history. This work follows in the spirit of Rosenberg (2011) who used an earlier dataset of microsatellite variation to create an approachable demonstration of the major features in the geographic distribution of human genetic variation (as well as earlier related papers such as Lewontin (1972); Witherspoon et al. (2007)). Our results complement several recent analyses of single-nucleotide variants in whole-genome sequencing data from humans (Auton et al., 2015; Mallick et al., 2016; Bergström et al., 2020) We label the approach taken here a variant-centric view of human genetic variation, in contrast to representations that focus on individuals or populations and their relative levels of similarity.

2.3 Results

2.3.1 A variant-centric view of genetic diversity

To introduce the approach, we begin with considering 100 randomly chosen single nucleotide variants sampled from chromosome 22 of the 1KGP high coverage data (2.3.1,(Fairley et al., 2020)). 2.1 shows the allele frequency of each variant (rows) in each of the 26 populations of the 1KGP (columns, see Supplemental Table 1 for labels). As a convention throughout this paper, we use deeper colors of blue to represent higher allele frequency, and we keep track of the globally minor allele, i.e., the rarer (¡ 50 % frequency) allele within the full sample. The figure shows that variants seem to fall into a few major descriptive categories: variants with alleles that are localized to single populations and rare within them, and variants with alleles that are found across all 26 populations and are common among them.

To investigate whether such patterns hold genome-wide, we devise a scheme that allows us to represent the ~ 92 million single-nucleotide variants (SNVs) in the genome-wide data (see schematic, Figure 2). First, we follow the 1KGP study in grouping the samples from the 26 populations into five geographical ancestry groups: African (AFR), European (EUR), South Asian (SAS), East Asian (EAS), and Admixed American (AMR) (Figure 2.2A, Box Box 2.3.1). For clarity, we modify the original 1KGP groupings slightly for this project (by including several samples from the Americas in the AMR grouping, see 2.3.1). While human population structure continuously varies and can be dissected at much finer scales than these groups (e.g., (Leslie et al., 2015; Novembre and Peter, 2016)), the regional groupings we use are a practical and instructive starting point — as we will show, several key features of human evolutionary history become apparent, and many misconceptions about human differentiation can be addressed efficiently with this coarse approach (see 2.4). As any such groupings are necessarily arbitrary, we also show results without using regional groupings to calculate frequencies (see section 2.3.4).

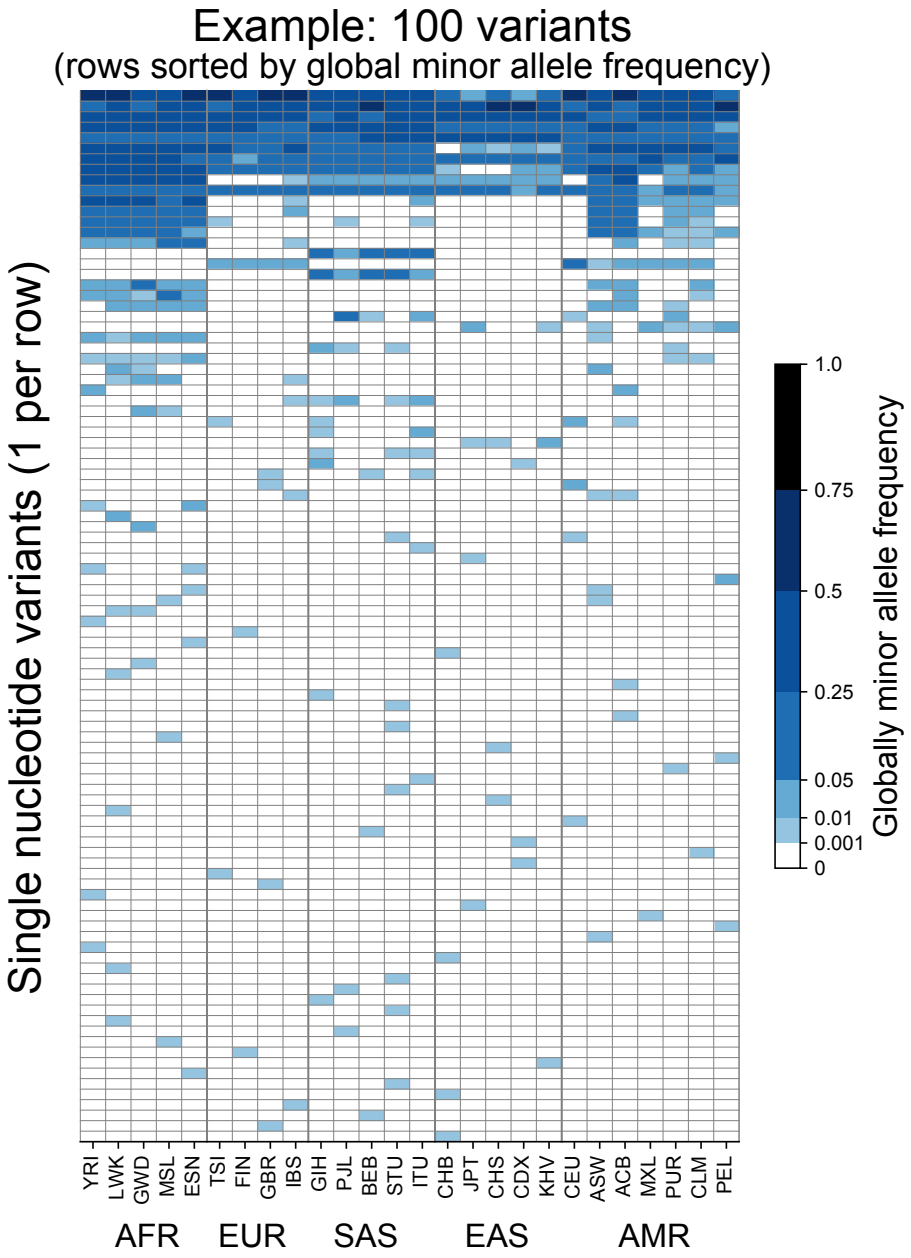

Figure 2.1: Frequencies of the globally minor allele across 26 populations from the 1KGP for 100 randomly chosen variants from Chromosome 22. Note that the allele frequency bin spacing is non-linear to capture variation at low as well as high frequencies.

To represent the geographic distributions of alleles compactly, we give every variant a five-letter code according to its allele frequencies across regions (Figure 2A). More precisely, for each bi-allelic single nucleotide variant, we identify the global rarer (minor) allele. Then

for each region, we code the allele's frequency as 'u', 'R', or 'C', based on whether the allele is "(u)ndetected," "(R)are," or "(C)ommon" (Figure 2B). Finally, we concatenate the allele's regional frequency codes in the fixed (and arbitrary) order: AFR, EUR, SAS, EAS, and AMR. This procedure generates a "geographic distribution code" per variant. For example, the code 'CCCCC' represents a variant that is common across every region, while 'uuRuu' represents a variant that is rare in South Asia and unobserved elsewhere (Figure 2.2C). This scheme requires a few choices. To distinguish between "rare" and "common" alleles, we used a threshold of 5% frequency. For comparison, we also show results using a 1% frequency threshold (Figure S1A). For 96.6% of variants in the dataset with high-quality ancestral allele calls (Box 1), the globally minor allele is the derived (younger) allele, and for comparison we also produced results tracking the derived rather than the globally minor allele (Figure S1C). Neither changing the frequency threshold to 1% nor tracking the derived allele meaningfully affects the basic observations that follow.

Next, we coded all 92 million biallelic SNVs in the dataset and tabulated the proportions of each geographic distribution code. We display the codes in a vertical stack from the most abundant code at the bottom to the least abundant at the top with the height of each code proportional to its abundance, so that the cumulative proportions of the rank-ordered codes are easily readable (Figure 3).

The distribution of codes is heavily concentrated, with 85% of variants falling into just eight codes out of the 242 (3^5-1) that are possible. Of the top eight codes, the top four codes represent rare variants that are localized in a single region. The fifth most abundant code , 'RuuuR', represents rare variants found in Africa and the Admixed Americas (which includes African-American individuals, for example). The sixth code is another set of localized rare variants ('uRuuu', i.e. variants rare in EUR). The seventh code is 'CCCCC' or "globally common variants." The eighth most abundant category 'uRuuR' represents rare variants found in Europe and the Admixed Americas. Conspicuously infrequent in the distribution

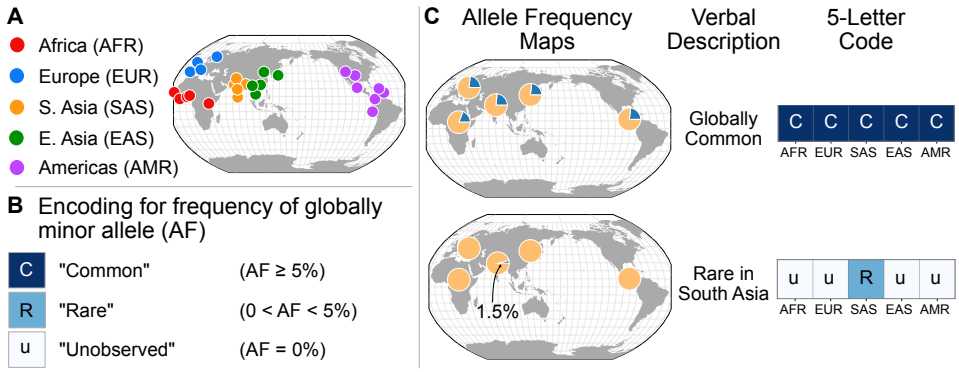

Figure 2.2: **A:** Regional groupings of the 26 populations in the 1KGP Project. **B:** Legend for minor allele frequency bins. **C:** Two examples of how a verbal description of an allele frequency map can be communicated equivalently with a 5-letter code (yellow signifies the major allele frequency, blue signifies the minor allele frequency in the pie charts).

are variants that are common in only one region outside of Africa and absent in others (e.g., 'uCuuu', 'uuCuu', 'uuuCu', 'uuuuC'). Instead, when a variant is found to be common (¿5% allele frequency) in one population, the modal pattern (37.3%) is that it is common across the five regions ('CCCCC'). Further, 63% of variants common in at least one region are also globally widespread, in the sense of being found across all five regions. This number rises to 82% for variants common in at least one region outside of Africa (Figure S2 and S3).

Singleton variants—alleles found in a single individual—are the most abundant type of variant in human genetic data and are necessarily found in just one geographic region. To focus on the distributions of non-singleton variants, we removed singletons and re-tallied the relative abundance of patterns (Figure 3C). Removing singletons reduces the absolute number of variants observed by 48.2% (91,784,637 vs. 44,290,364). Without singletons, we see more clearly the abundance of patterns that have rare variants shared between two or more regions (codes with two R's and one u, such as 'uuRRu' or 'RRuuu').

The patterns observed here are interpretable in light of some basic principles of population genetics. Rare variants are typically the result of recent mutations (Mathieson and McVean, 2014; Kiezun et al., 2013; Kimura and Ohta, 1973; Albers and McVean, 2020). Thus, we interpret the localized rare variants (such as 'Ruuuu'or 'uuuRu') as mostly young

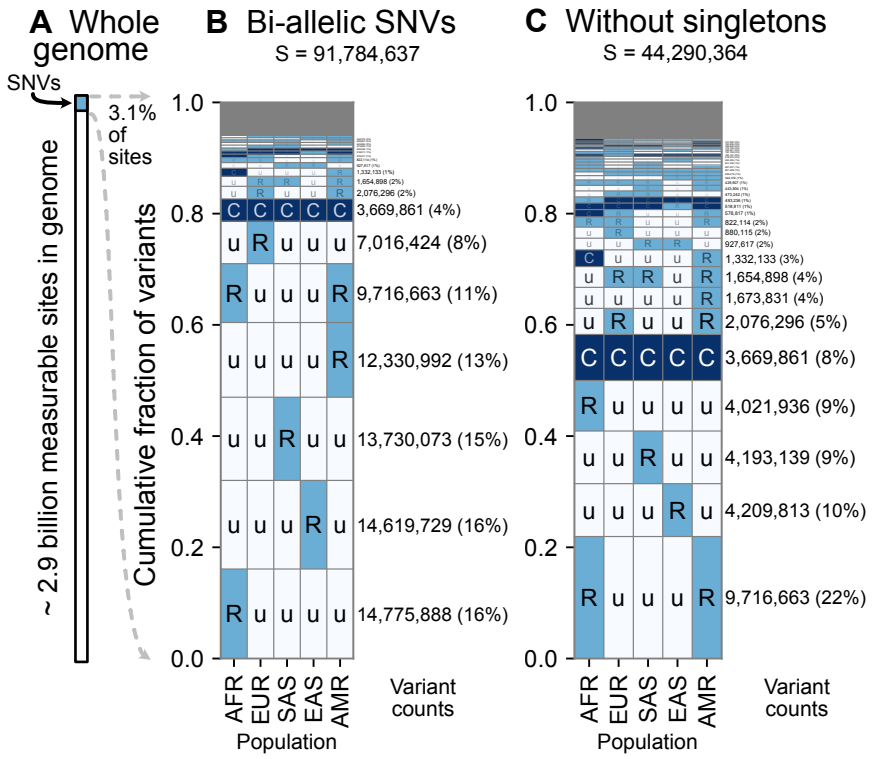

Figure 2.3: **A:** We observe variants at 3.1% of the measurable sites in the reference human genome (GRCh38). A measurable site is one at which it is possible to detect variation with current sequencing technologies (currently approximately 2.9 Gb out of 3.1 Gb in the human genome; see URLs). **B and C:** The relative abundance of different geographic distributions for 1KGP variants, **(B)** including singletons, and **(C)** excluding singletons. In panels **B** and **C**, the right-hand rectangles show the number and percentage of variants that fall within the corresponding geographic code on the left-hand side; distribution patterns are sorted by their abundance, from bottom-to-top. See Figure 2.2 for an explanation of the 5-letter 'u','R','C' codes. The proportion of the genome with variants that have a given geographic distribution code can be calculated from the data above (for example, with the 'Ruuuu' code, as $17\% \times 3.1\% = 0.53\%$).

mutations that have not had time to spread geographically. The code 'CCCCC' (globally common variants), likely comprises mostly older variants that arose in Africa and were spread globally during the Out-of-Africa diaspora and other dispersal events (see Box 2). The appearance of rare variants shared between two or more regions (codes with two R's and three u's, such as 'uuRRu' or 'RRuuu') is likely the signature of recent gene flow between those regions (2.3.1)(Platt et al., 2019; Mathieson and McVean, 2014; Gutenkunst et al., 2009). In

particular, the abundant 'RuuuR' and 'uRuuR' codes likely represent young variants that are shared between the Admixed Americas and Africa ('RuuuR') or Europe ('uRuuR') because of the population movements during the last 500 years that began with European colonization of the Americas and the subsequent slave trade from Africa. We interpret the 10th most abundant code ('CuuuR') as mostly variants that were lost in the Out-of-Africa bottleneck and subsequently carried to the Americas by African ancestors. There is a relative absence of variants that are common in only one region outside of Africa and absent across all others (e.g., 'uCuuu', 'uuCuu', 'uuuCu', 'uuuuC') – this is consistent with human populations having not diverged deeply, in the sense that there has not been sufficient time for genetic drift to greatly shift allele frequencies among them (Box 2). To help make this clear, consider the alternative scenario—in a deep, multiregional origins model (Wolpoff et al., 1984), one would expect many more variants to be common to one region and absent in others ('uCuuu' or 'uuuCu' for example, see 2.3.1). Overall, these results reflect a time-scale of divergence consistent with the Recent-African-Origin model of human evolution as well as subsequent gene flow among regions (Cann et al., 1987; Stringer and Andrews, 1988; Thomson et al., 2000; Ramachandran et al., 2005; Pickrell and Reich, 2014)

Box 1: Dataset Descriptions and Groupings

We use bi-allelic single nucleotide variants from the New York Genome Center high-coverage sequencing of the 1000 Genomes Project (1KGP) Phase 3 samples (Auton et al., 2015; Fairley et al., 2020) (see URLs, Accessed July 22nd, 2019, we include only variants with PASS in the vcf variant filter column). Most of the samples are from an ethnic group in an area (e.g., the "Yoruba of Ibadan," YRI, or the "Han Chinese from Beijing," CHB), so the sampling necessarily represents a simplification of the diversity present in any locale (e.g., Beijing is home to several ethnic groups beyond the Han Chinese). For each grouping, the 1KGP typically required each individual to have at least 3 out of 4 grandparents who

identified themselves as members of the group being sampled. The 1KGP further defined five geographical ancestry groups: African (AFR), European (EUR), South Asian (SAS), East Asian (EAS), and Admixed American (AMR). Differing from the 1KGP, we include in the "Admixed in the Americas" (AMR) regional grouping the following populations: "Americans of African Ancestry in SW USA", "African-Caribbeans in Barbados (ACB)", and the "Utah Residents (CEPH) with Northern and Western European Ancestry". We chose this grouping because it is a more straightforward representation of current human geography. We note challenges and caveats of these alternate decisions in the Discussion. Supplemental Table 1 provides a full list of the 26 populations and the grouping into five regions. Figure 7 and Figure S7 provide a complementary view to Figure 2 where the analysis is not based on the five groupings, but instead all 26 populations.

In Figure 5, we present results for variants differing between pairs of individuals from the Simons Genome Diversity Project (SGDP). We include only autosomal biallelic SNVs for variants that pass "filter level 1", which is the filtering procedure for the majority of analyses used by (Mallick et al., 2016) (see URLS). In Figure 6 we present results for variants found on 5 commercially available genotyping arrays: The Affymetrix 6.0 (Affy6) genotyping array, the Affymetrix Human Origins array (HumanOrigins), the Illumina HumanOmniExpress (OmniExpress) array, and the Illumina Omni2.5Exome (Omni2.5Exome), and the Illumina MEGA array (MEGA). We only include autosomal biallelic SNVs in our analysis. Variant lists for each array were downloaded from the manufacturer websites (see URLs). For assessing the impact of polarizing to ancestral or derived alleles, we downloaded human ancestral allele calls for GRCh38 based on an 8 primate EPO alignment from Ensembl (see URLs). We used only ancestral allele calls supported by at least two outgroup species for our downstream analysis.

Box 2: Theoretical Modeling

We can use theoretical models to estimate what our visualizations would look like for two populations in simple contrasting cases of "deep" divergence, "shallow" divergence", and "shallow" divergence with gene flow. The "shallow" case is calibrated to be qualitatively consistent with the Recent-African-Origin model with subsequent gene flow. The "deep" case mimics a multi-regional model of human evolution (Wolpoff et al., 1984). For each case, we computed the expected abundances of distribution codes in a simple model of population divergence: two modern populations of N individuals each that diverged T generations ago from a common population of N individuals (see Appendix for information about this calculation). We model gene flow by including recent admixture: individuals in Population A derive an average fraction α of their ancestry from Population B and vice versa. This simplified model neglects many of the complications of human population history, including population growth, continuous historical migration, and natural selection, but it captures the key features of common origins, divergence, and subsequent contact.

In this model, the key control parameter is T/2N, the population-scaled divergence time. Human pairwise nucleotide diversity ($\sim 10^{-3}$) and per-basepair per-generation mutation rate ($\sim 1.25 \times 10^{-8}$) imply a Wright-Fisher effective population size of $N = 2 \times 10^4$ individuals. The Out-of-Africa divergence is estimated to have occurred approximately 60,000 years ago (Nielsen et al., 2017). Assuming a 30-year generation time (Fenner, 2005) gives T/2N=0.05. We compare this scenario with $T/2N = 0.5$, corresponding to a deeper divergence of approximately 600,000 years ago.

Figure 2.4A shows the expected patterns in a sample of 100 individuals from each population for deep divergence ($T/2N = 0.5$), recent divergence ($T/2N = 0.05$) without admixture, and recent divergence with admixture ($\alpha = 0.02$). The recent divergence model with or without admixture reproduces the preponderance of Ru and CC mutations seen in the data, while the deep divergence model shows many more Cu and many fewer CC mutations. The case

with admixture shows a slight increase in variant sharing (RR alleles increase from 1.3% of variants to 4.2%; RC and CR alleles increase from 6% to 10%; CC alleles comprise 23% in both cases).

We can understand the relationship between the split time and geographic distribution abundances heuristically as follows. During an interval of t generations, the frequency of a neutral mutation starting at frequency f changes randomly by a typical amount of $f(1-f)2Nt$. Consider a mutation that is at 25% frequency, i.e., common, in the ancestral population at the time of the split (Figure 4B). At time $\frac{t}{2N} = 0.05$ after the split, the frequency of the mutation is likely to be in the interval (15%, 35%) in both populations and will be assigned the code CC. On the other hand, by time $\frac{t}{2N} = 0.5$ after the split, the mutation has a significant chance of going extinct in one or both populations (Figure 2.4C). Mutations that go extinct in one population but not the other will typically be assigned a code Cu or uC.

At the same time, new mutations are constantly entering the evolving populations. These new mutations will be private to one population (Ru or Cu) and the overwhelming majority will go extinct before reaching detectable frequencies. Conditional on non-extinction, the expected frequency of a neutral mutation increases linearly with time (see Appendix B). As a result, the frequencies of new mutations since the split time twill mostly be contained in a triangular envelope $f < \frac{t}{2N}$ (Figure 2.4B). For recent divergence, the new mutations will be assigned code Ru or uR, while in deeply diverged populations they may be categorized as Cu or uC.

2.3.2 The variants that differ between a pair of individuals

While Figure 3 illustrates genetic variants found in a large, global sampling of human diversity, it does not show what to expect for the variants that differ between pairs of individuals. Are the variants that differ between two individuals more often geographically widespread or

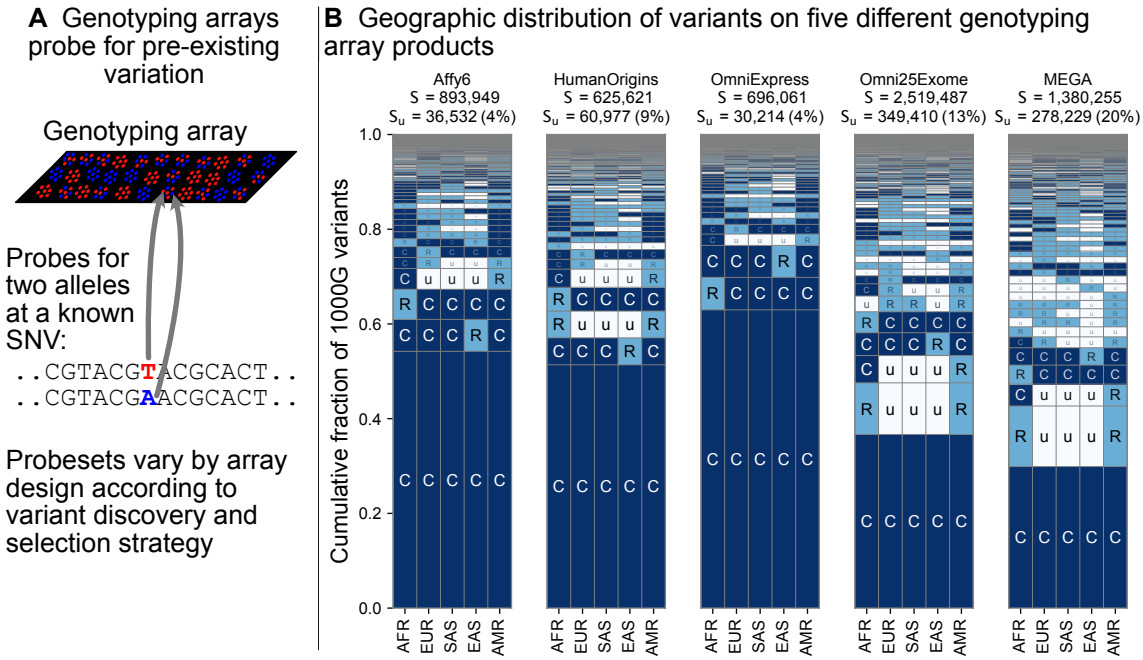

Figure 2.4: **A:** Expected geographic distribution code abundances in a sample of 100 diploid individuals from each of two populations, for deep divergence ($T/2N = 0.5, \alpha = 0$), recent divergence without admixture ($T/2N = 0.05, \alpha = 0$), and recent divergence with admixture ($T/2N = 0.05, \alpha = 0.02$). **B:** Simulated allele frequency time series for mutations starting at 25% frequency (blue) and new mutations entering the population since the split (orange). **C:** The probability of extinction of a mutation starting at 25% frequency (see Appendix B).

spatially localized? To address these questions, we considered the variants carried by pairs of individuals from the whole-genome sequencing data of the Simons Genome Diversity Project (SGDP) (Mallick et al., 2016) (Figure 2.5). The SGDP sampled 300 individuals from 142 diverse populations. We use the SGDP data to avoid ascertainment biases that might arise from looking at individuals within the same dataset we use to measure allele frequencies. Figure 5 shows a representative subset with 6 pairs chosen from 3 populations (Figure S6, shows a larger set of examples). For each pair we see some variants that were undiscovered in the 1KGP data (denoted S_u in the figure). These account for 17-20% of each set of pairwise SNVs and are likely rare variants. We see that the variants that differ between each pair of individuals are typically globally widespread (i.e., codes with no 'u's, with proportions out of the total S varying from 54%–76% for the pairs in Figure 2.5.) The observation of mostly

globally common variants in pairwise comparisons may seem counter-intuitive considering the abundance of rare, localized variants overall. However, precisely because rare variants are rare, they are not often carried by either individual in a pair. Instead, pairs of individuals mostly differ because one of them carries a common variant that the other does not; and as Figure 2.3 already showed, common variants in any single location are often common throughout the world (also see Figures 2.7 and S1).

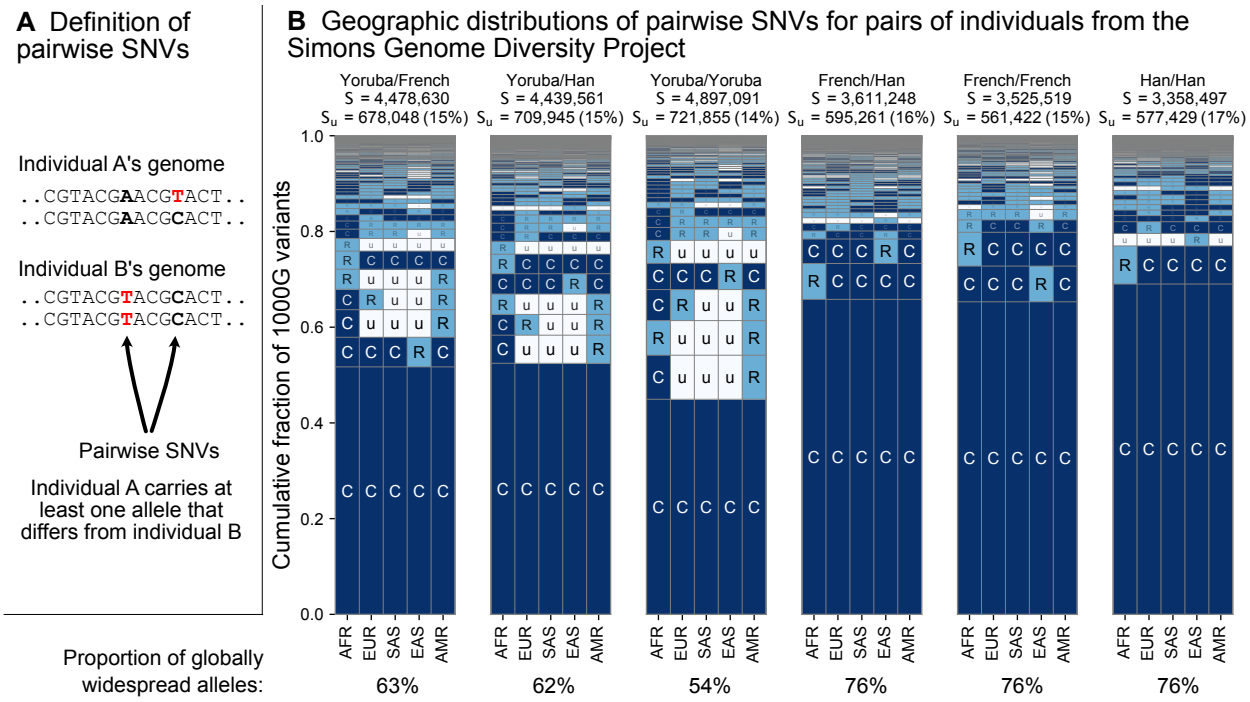

Figure 2.5: **A:** Definition of a pairwise SNV. **B:** The abundance of geographic distribution codes for different pairs of individuals from the SGDP dataset. Above each plot we show the total number of variants that differ between each individual (S) and the number that were unobserved completely in the 1KGP data (S_U). Across the bottom we show the proportion of variants with globally widespread alleles for each pair. We calculate this as the fraction of variants with no 'u' encodings over the total number of variants (S). (Note: by doing so, we make the assumption that if a variant is not found in the 1KGP data it is not globally widespread).

From the example pairwise comparisons (Figure 5, and Figure S6), one also observes evidence for higher diversity in Africa, which is typically interpreted in terms of founder effects reducing diversity outside of Africa (Cann et al., 1987; Harpending and Rogers, 2000;

Ramachandran et al., 2005; Prugnolle et al., 2005); though other models, especially ones including substantial subsequent admixture, can also produce this pattern (DeGiorgio et al., 2009; Pickrell and Reich, 2014). For example, the two Yoruba individuals have more pairwise SNVs (S = 4,897,091) than the French/French (S = 3,525,519) and Han/Han (S = 3,358,497) pairs. Pairs involving one or both of the sample Yoruba individuals have more variants with alleles common in Africa and rare or absent elsewhere (e.g., 'CuuuR', 'RuuuR'). Finally, a more subtle, but expected, impact of founder effects is that the sample Yoruba/Yoruba comparison is expected to have higher numbers of pairwise variants than the sample Yoruba/Han or Yoruba/French comparison, which we observe.

2.3.3 The geographic distributions of variants typed on genotyping arrays

Targeted genotyping arrays are a cost-effective alternative to whole-genome sequencing. The geographic distribution of the variants on genotyping arrays affects genotype imputation and genetic risk prediction (Howie et al., 2012; Martin et al., 2017). In contrast to whole-genome sequencing, genotyping arrays use targeted probes to measure an individual's genotype only at preselected variant sites. The process of discovering and selecting these target sites typically enriches the probe sets towards common variants (Clark et al., 2005) and under-represents geographically localized variants (Albrechtsen et al., 2010; Lachance and Tishkoff, 2013).

Figure 2.6 shows the geographic distributions of bi-allelic SNVs included on five popular array products in the 1KGP data. In stark contrast with the SNVs identified by whole-genome sequencing (Figure 3B), a large fraction of the variants on genotyping arrays are globally common, especially for the Affy6, Human Origins, and OmniExpress arrays which were designed primarily to capture common variants. The Omni2.5Exome and MEGA arrays in contrast exhibit many more rare variants. In both of these arrays, the second and third most abundant codes are 'CuuuR' and 'RuuuR' variants. The MEGA array was

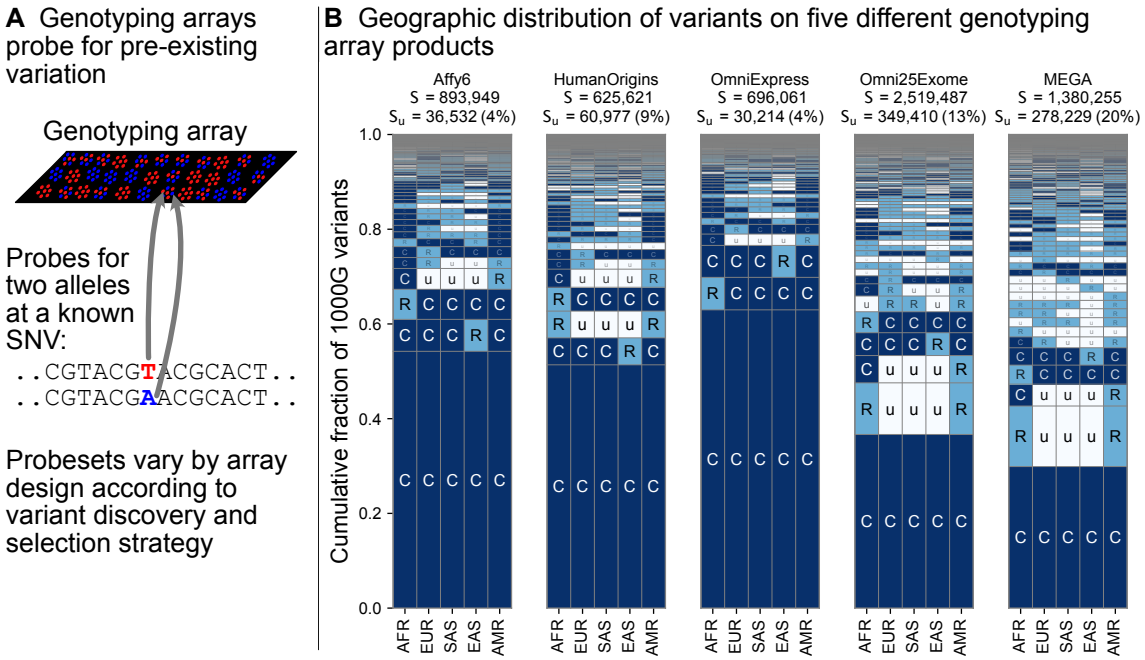

Figure 2.6: **A:** Genotyping arrays consist of probes for a fixed set of variants chosen during the design of the array product. **B:** For each array product, we extracted the genomic position of variants found on the array and kept variants that are also found within the 1KGP to highlight their geographic distributions.

uniquely designed to capture rare variation in undersampled continental groups, including African ancestries (Bien et al., 2016, 2019).(Wojcik et al., 2019) found that this design improved African and African-American imputation accuracy, leading to greater power to map population-specific disease risk.

2.3.4 Finer-scale resolution of variant distributions

While the use of 5 regional groupings above allows us to describe variant distributions compactly with a 5-digit encoding, the basic principle of grouping allele frequencies can be extended to build a 26-digit encoding for the 1KGP variants. Doing so, we find a consistent pattern with Figure 2.3B, in that the majority of variants are seen to be rare and geographically localized (1 'R', and the remainder 'u's), and when a variant is common in any one population, it is typically common across the full set of populations (Figure 7, pattern with

all ‘C’s). This view reveals that the 5-digit encodings with 1 ‘R’ and 4 ‘u’s are often due to variants that are rare even within a single population. This is not unexpected given many of them are singletons. When we remove singletons (Supp Fig. 7), we again see more clearly rare allele sharing indicative of recent gene flow, though at finer-scale resolution.

2.4 Discussion

By encoding the geographic distributions of the ∼92 million biallelic SNVs in the 1KGP data and tallying their abundances, we have provided a new visualization of human genetic diversity. We term our figures “GeoVar” plots as they help reveal the geographic distribution of variant sets. GeoVar plots can complement other methods of visualizing population structure, including: plots of pairwise genetic distance, dimensionality-reduction approaches such as PCA, admixture proportion estimates such as STRUCTURE, and explicitly spatial methods that use the sampling locations of individuals (Guillot et al., 2009; Novembre and Peter, 2016; Bradburd and Ralph, 2019). These previously developed methods help reveal population structure, infer genetic ancestry, and measure historical migration patterns. However, they do a poor job of showing how alleles are distributed geographically. To minimize confusion about levels of differentiation among populations, researchers and educators can consider complementing PCA or STRUCTURE outputs with a variant-centric visualization like the ones presented here. To that end, we provide source code to replicate our figures and to generate similar plots for other datasets (the “GeoVar” software package; see URLs).

A goal of our work was to build a visualization that can help correct common misconceptions about human genetic variation. First, because many existing methods to describe population structure emphasize between-group or between-individual differentiation, they can convey a misleading impression of “deep” divergence between populations when it may not exist. Comparing Figure 2.1 to outputs of models with “deep” or “shallow” divergence can help teach how patterns of human variation are consistent with shallow divergence and

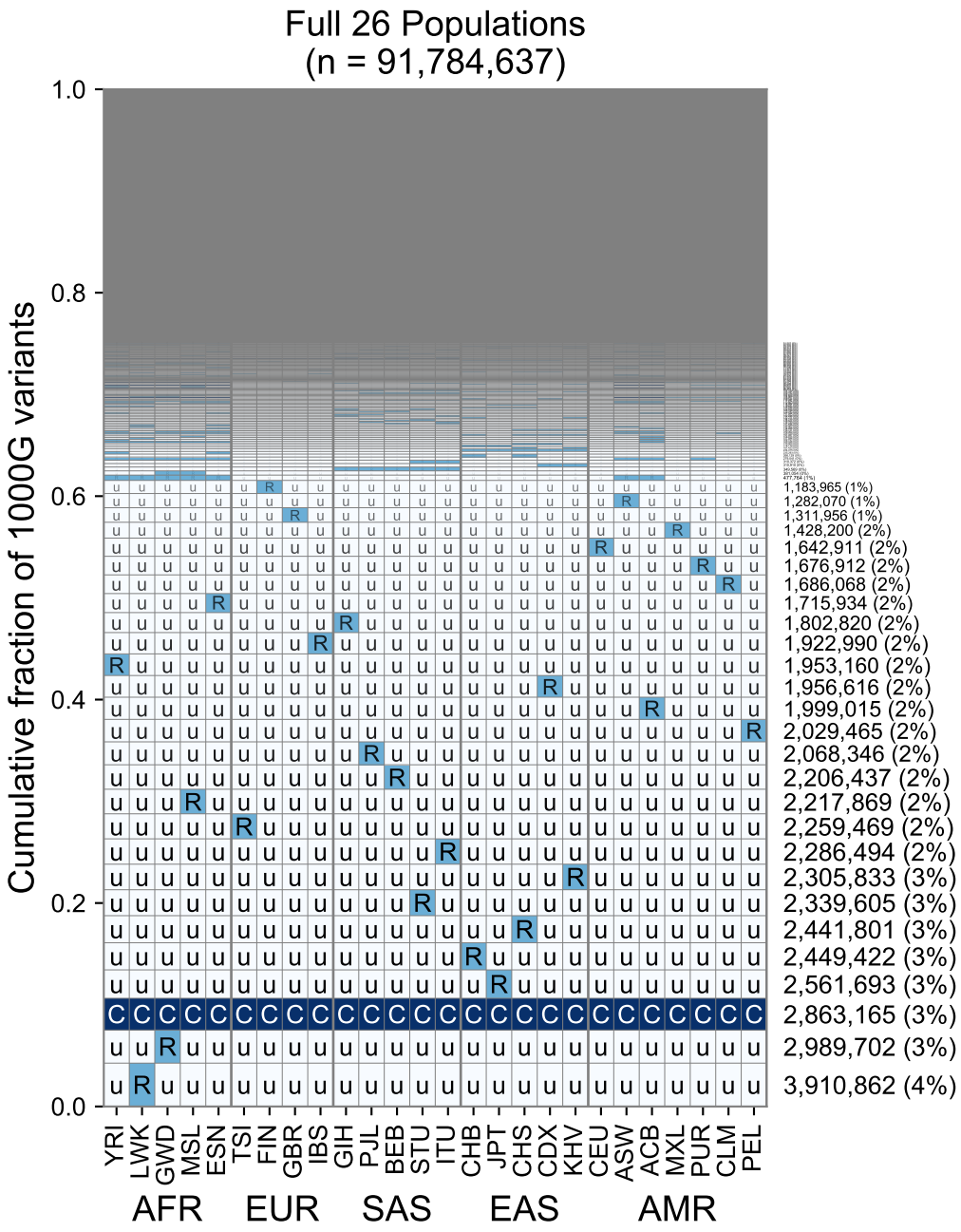

Figure 2.7: This plot is the analogous plot to Figure 2.3B but rather than calculating frequencies with the 5 regional groupings, we compute them within each of the 26 1KGP populations. The total number of variants represented is the same as in Figure 2.3B ($S = 91,784,367$). See Figure 2.2 for an explanation of the 'u','R','C' codes.

the Recent African Origins model (Box 2.3.1). Second, because personal ancestry tests can identify ancestry to broad continental regions, it is possible to incorrectly conclude human

alleles are typically found exclusively in a single region and at high frequency within that region (e.g, patterns such as "uuCuu") As our figures show, this is not the case. Rather, it should be kept in mind that most fine-scale personal ancestry tests work using genotyping arrays and combining evidence from subtle fluctuations in the allele frequencies of many common variants (Novembre and Peter, 2016). Finally, another related misconception is that two humans from different regions of the world differ mainly due to alleles that are typical of each region. As we show in Figure 2.5, most of the variants that differ between two individuals are variants with alleles that are globally widespread.

Our method requires computing allele frequencies within pre-defined groupings. Grouping and labeling strategies vary between genetic studies, and are determined by the goals and constraints of a particular study (Race, Ethnicity, and Genetics Working Group, 2005; Panofsky and Bliss, 2017; Mathieson and Scally, 2020). While we chose deliberately coarse grouping schemes to address the misconceptions described above, the key facts we derive about human genetic variation are robust, and appear in finer-grained 26-population versions of the plot (Figure 2.7). We recommend that any application of the GeoVar approach needs to be interpreted with the choice of groupings in mind. . The visualization method developed here is also useful for comparing the geographic distributions of different subsets of variants, (e.g., Figures 2.5 and 2.6). For example, when applied to the list of variants targeted by a genotyping array (Figure 2.6), the approach quickly reveals the relative balance of common versus rare variants and the geographical patterns of those variants. Interpreting the results of this visualization approach does have some caveats. First, we estimate the frequency of alleles from samples of local populations. We expect that as sample sizes increase many alleles called as unobserved "u" will be reclassified as rare "R". The average sample size across all of our geographic regions is approximately 500 individuals (AFR: 504, EUR: 404, SAS: 489, EAS: 504, AMR: 603). Assuming regions are internally well-mixed, we have $\sim$80% power to detect alleles with a frequency of $\sim$0.2% in a region (Figure 2.11). For al-

leles with lower frequencies, we would require larger sample sizes to ensure similar detection power. An implication is that in large samples, we should observe more rare variant sharing. Thus, we expect the figures here to under-represent the levels of rare variant sharing between human populations.

A second caveat is that our encoding groups a wide range of variants into the "(C)ommon" category (i.e., all variants where the frequency of the globally minor allele is greater than 5%). For some applications, such as population screening for carriers, it may be enough to know a variant falls in the "rare" or "common" bins we have described, and more detail is inconsequential. For other applications, the detailed fluctuations in allele frequency across populations are relevant—for example, differences in allele frequencies at common variants (Figure S5) are regularly used to infer patterns of population structure and relatedness (Li et al., 2008; Pickrell and Pritchard, 2012; Patterson et al., 2012).

Third, one must interpret our results with the sampling design of the 1KGP study design in mind. In particular, the 1KGP filtered for individuals of a single ethnicity within each locale. However, in our current cosmopolitan world, the genetic diversity in any location or broad-based sampling project will be considerably higher than implied by the geographic groupings above. For example, the UK Biobank, while predominantly of European ancestry, has representation of individuals from each of the five regions used here (Bycroft et al., 2018). The 1KGP also sampled South Asian ancestry from multiple locations outside of South Asia, and whether those individuals show excess allele sharing due to recent admixture in those contexts is unclear. While we expect overall similar patterns to those seen here using emerging alternative datasets (Bergström et al., 2020), there may be subtle differences due to sampling and study design considerations. Despite these caveats, the results of the visualizations provided here help reinforce the conclusions of a long history of empirical studies in human genetics (Lewontin, 1972; Ramachandran et al., 2005; Conrad et al., 2006; Li et al., 2008; Auton et al., 2015; Mallick et al., 2016; Bergström et al., 2020). The results

show how the human population has an abundance of localized rare variants and broadly shared common variants, with a paucity of private, locally common variants. Together these are footprints of the recent common ancestry of all human groups. As a consequence, human individuals most often differ from one another due to common variants that are found across the globe. Finally, though not examined explicitly above, the large abundance of rare variants observed here is another key feature of human variation and a consequence of recent human population growth (Slatkin and Hudson, 1991; Di Rienzo and Wilson, 1991; Keinan and Clark, 2012; Nelson et al., 2012; Tennessen et al., 2012)

The well-established introgression of archaic hominids (e.g., Neandertals, Denisovans) into modern human populations (Wolf and Akey, 2018) is not apparent in the GeoVar plots we produced. We believe that there are two broad reasons for this: (1) The clearest signal of introgression will come from sites where archaic hominids differed from modern humans, and we expect that these sites are only a very small fraction of variants found in humans today. The average human-Neandertal and human-Denisovan sequence divergence are both less than 0.16% (using observations from Prüfer et al. (2014)), and a recent study estimates that there are fewer than 70 Mb (2.3% of the genome) of Neanderthal introgressed segments per individual for all individuals in the 1KGP (Chen et al., 2020) (2) We do not expect SNVs from archaic introgression to be concentrated in a single GeoVar category. For example, introgressed variants occupy a wide range of allele frequencies (Bergström et al., 2019). Archaic introgression events are believed to be old: >30,000 years ago, allowing time for substantial genetic drift and admixture among human populations(Chen et al., 2020). Negative selection (Harris and Nielsen, 2016; Juric et al., 2016), and in some cases, strong positive selection(Racimo et al., 2015) have also shaped the patterns of introgressed SNVs. For these reasons, we expect low levels of archaic introgression not to create a striking visual deviation in our GeoVar plots from the background patterns of a Recent African Origin model with subsequent migration (Box 2.3.1). To highlight the contributions of archaic hominids

to human variation, more targeted approaches are needed (e.g. Green et al., 2010; Durand et al., 2011). Future work could also naturally extend the approach here to include archaic sequence data.

The geographic distributions of genetic variants visualized here are relevant for a number of applications, including studying geographically varying selection (Yi et al., 2010; Key et al., 2018), human demographic history (Gutenkunst et al., 2009), and the genetics of disease risk. For instance, due to ascertainment bias in arrays (Figure 2.6) and power considerations, common variants are often found in genome-wide association studies of disease traits (Manolio et al., 2009). The patterns shown above make it clear that most common variants are shared across geographic regions. Indeed, many common variant associations replicate across populations (Marigorta and Navarro, 2013) (though see (Bomba et al., 2017)). As our work here emphasizes, rare variants are likely to be geographically restricted, and so one can expect the rare variants found in one population will not be useful for explaining trait variation in other populations, though they may identify relevant biological pathways that are shared across populations.

A future direction for the work here would be to apply our approach to other forms of variation such as insertions, deletions, microsatellites, and structural variants. We note that in studies with sample sizes similar to or smaller than the 1KGP, nearly all SNVs arise from single mutation events. For other variants that arise from single mutation events (e.g., indels that arise from single mutations), we expect similar patterns to those observed for SNVs here. In contrast, for highly mutable loci, such as microsatellites, we expect alleles will be distributed in disjoint regions of the world due to multiple mutational origins (Ralph and Coop, 2013; Mathieson and McVean, 2014; Phillips et al., 2020).

Another future direction would be to shift from visualizing patterns of allele sharing to the patterns of sharing of ancestral lineages in coalescent genealogies. Recent advances in the inference of genome-wide tree sequences (Kelleher et al., 2019; Speidel et al., 2019) and

allele ages (Albers and McVean, 2020) allow for quantitative summaries of ancestral lineage sharing. Such quantities have a close relationship to the multi-population SFS properties that are studied here, yet are more fundamental in a sense and less subject to the stochasticity of the mutation process. That said, the conceptual simplicity of visualizing allele frequency patterns may be an advantage in educational settings.

Most importantly, future applications of the approach will ideally use datasets that include a greater sampling of the world's genetic diversity (Bustamante et al., 2011; Popejoy and Fullerton, 2016; Martin et al., 2017; Peterson et al., 2019). A related point is that the application of our method to genotyping array variants (Fig. 2.6) reinforces the importance of considering the ancestry of study populations in genotype array design and selection (Peterson et al., 2019).

Overall, the visualizations produced here provide an interpretable way to depict geographic patterns of human genetic variation. With personal genomic technologies and ancestry testing becoming commonplace, there is increasing importance in fostering the understanding of human population genetics. To this end, human genetics researchers must develop interpretable materials on patterns of genetic variation for use in educational and outreach settings (Donovan et al., 2019). The variant-centric approach detailed here complements existing visualizations of population structure, facilitating a clearer understanding of the major patterns of human genetic diversity.

2.5 Acknowledgments

The 1000 Genomes data used here were generated at the New York Genome Center with funds provided by NHGRI Grant 3UM1HG008901-03S1. We thank members of the Novembre Lab, especially as this project was initiated in a group hackathon with contributions from Hussein Al-Asadi, Kushal Dey, Evan Koch, Joe Marcus, Ben Peter, Mark Reppell, and Joel Smith. We also thank Jeremy Berg, Jedidiah Carlson, Anna Di Rienzo, Joe Marcus, Aaron

Panofsky, Molly Przeworski, Harald Ringbauer, Mashaal Sohail, Matthias Steinrücken, and Xin He for comments on the manuscript draft, and Paul Strode and Brian Donovan for helpful conversations. This work was completed in part with resources provided by the University of Chicago's Research Computing Center and was supported by NIH training grant T32 GM07197 (AB), the University of Chicago "Chicago Fellows" program (DPR), and NIH grant R01 GM132383.

2.5.1 Supplementary figures

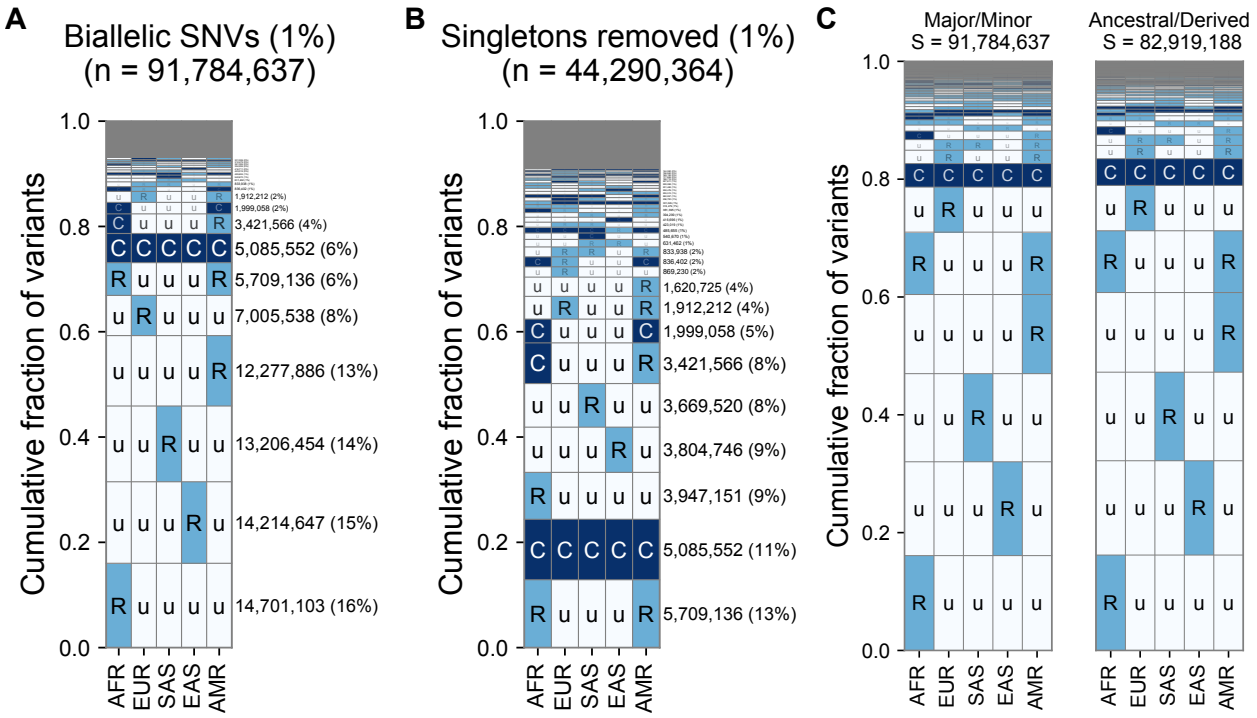

Figure 2.8: **A:** The relative abundance of geographic distribution codes within the $\sim$ 92 million variants when using an MAF of 1% as the distinction between "common" ('C'), and "rare" ('R'). The right-hand panel shows the percentage of variants that fall within the geographic code represented on the left-hand side; distribution patterns are sorted by their abundance, from bottom-to-top. **B:** The abundance of geographic distribution codes for $\sim$ 44 million non-singleton variants using an MAF of 1% as the boundary between "common" ('C'), and "rare" ('R'). **C:** Comparison for the abundance of geographic distribution codes when polarizing to the ancestral and derived allele (using build 38) versus major/minor allele. We only include positions where an ancestral allele is supported by at least two outgroups. At 96.6% of variants (80,068,013 / 82,919,198), the minor allele is also the derived allele.

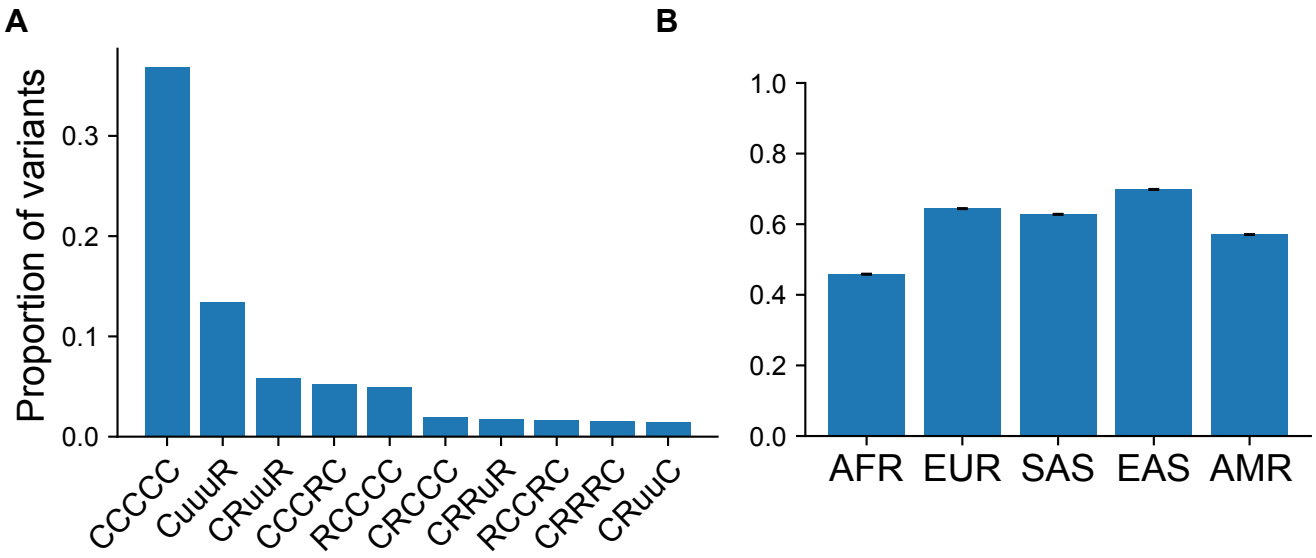

Figure 2.9: **A:** Top 10 categories when conditioning on the variant being "common" (MAF $> 5\%$) in at least one population. Conditioned on a variant being common in a single region, 37.3% of variants are categorized as "globally common" or "CCCCC". **B:** The proportion of variants that fall within the "globally common" or "CCCCC" geographic distribution code conditional on the variant being common (MAF $> 5\%$) in the specific continental group.

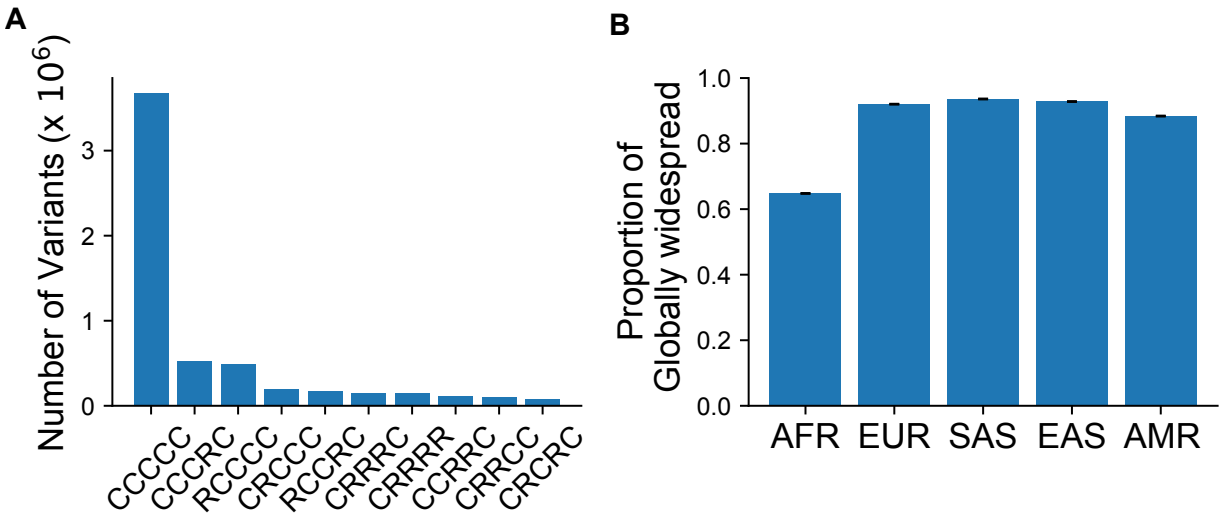

Figure 2.10: **A:** The proportion of variants that fall within a given geographic distribution code conditional on the variant being "globally widespread", i.e. a category that has no unobserved ("u") codes. We note that 55.6% of variants conditioned on being globally widespread are also globally common ("CCCCC"). In terms of absolute numbers, variants that are common in at least one population ($S = 9,958,838$) that are also globally widespread ($S = 6,322,767$) comprise $\sim$ 63% of the total when conditioning on being common in at least one population. When conditioning on variants common only in regions outside Africa ($S = 7,544,648$), the percentage of globally widespread variants ($S = 6,179,781$) increases to $\sim$ 82%. **B:** The proportion of variants that fall within a "globally present" category, defined as categories that contain no unobserved ("u") codes, conditional on the variant being common (MAF $> 5\%$) in the specific continental group.

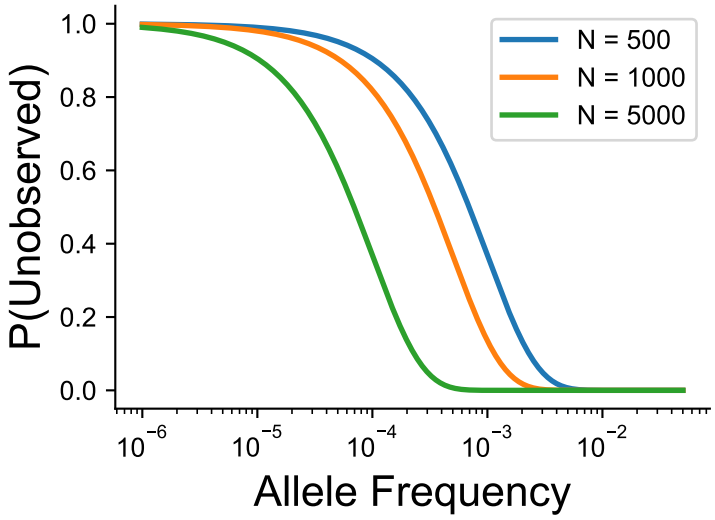

Figure 2.11: Probability of not observing a variant at a given allele frequency and sample size in number of individuals. We have assumed that allele frequencies follow Hardy-Weinberg equilibrium, and the probability of no observations of an allele is calculated using the binomial distribution.

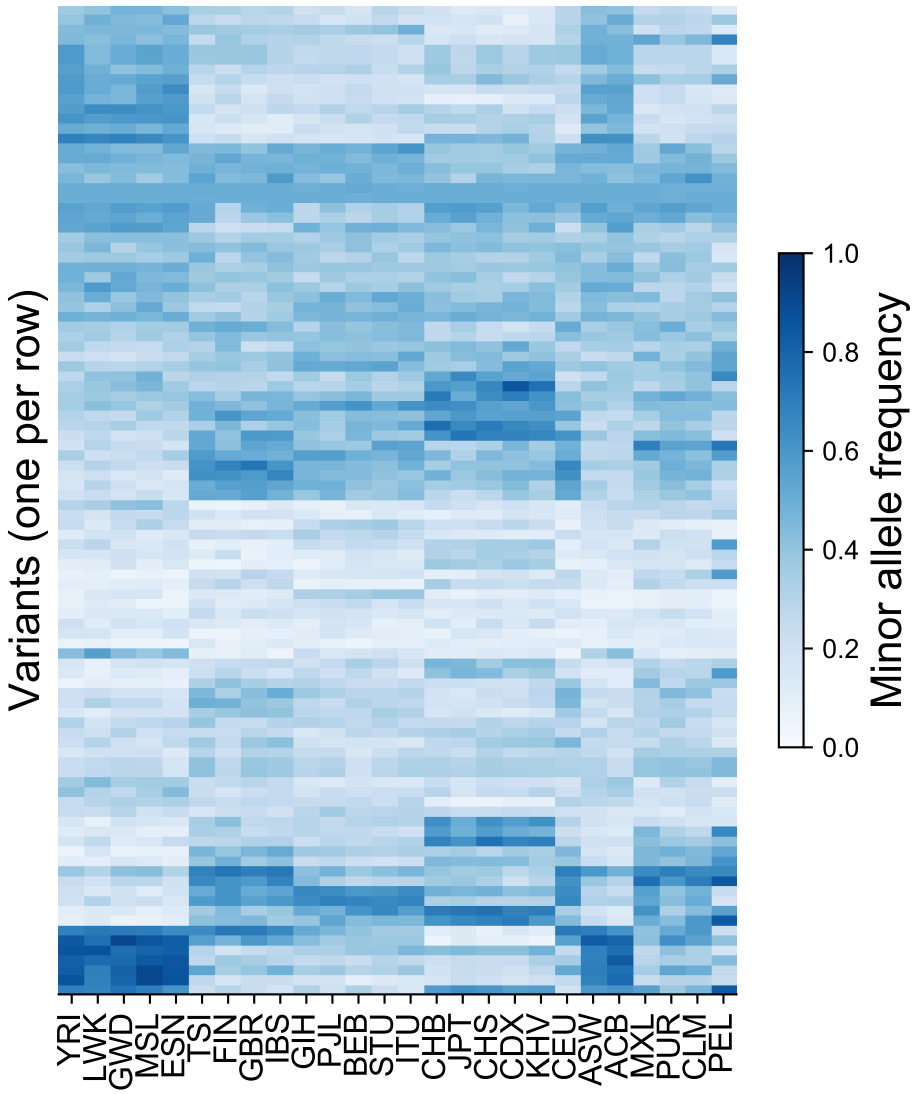

Figure 2.12: The minor allele frequencies of 300 variants in each of the 26 original population labels in the 1KGP. The variants were chosen at random from among those on Chromosome 22 that have MAF $> 5\%$ in all 26 populations. For example, the top row represents an allele that has higher frequency in several African and admixed American populations. Variants are ordered based on hierarchical clustering on the Euclidean distance between minor allele frequency profiles across all populations.

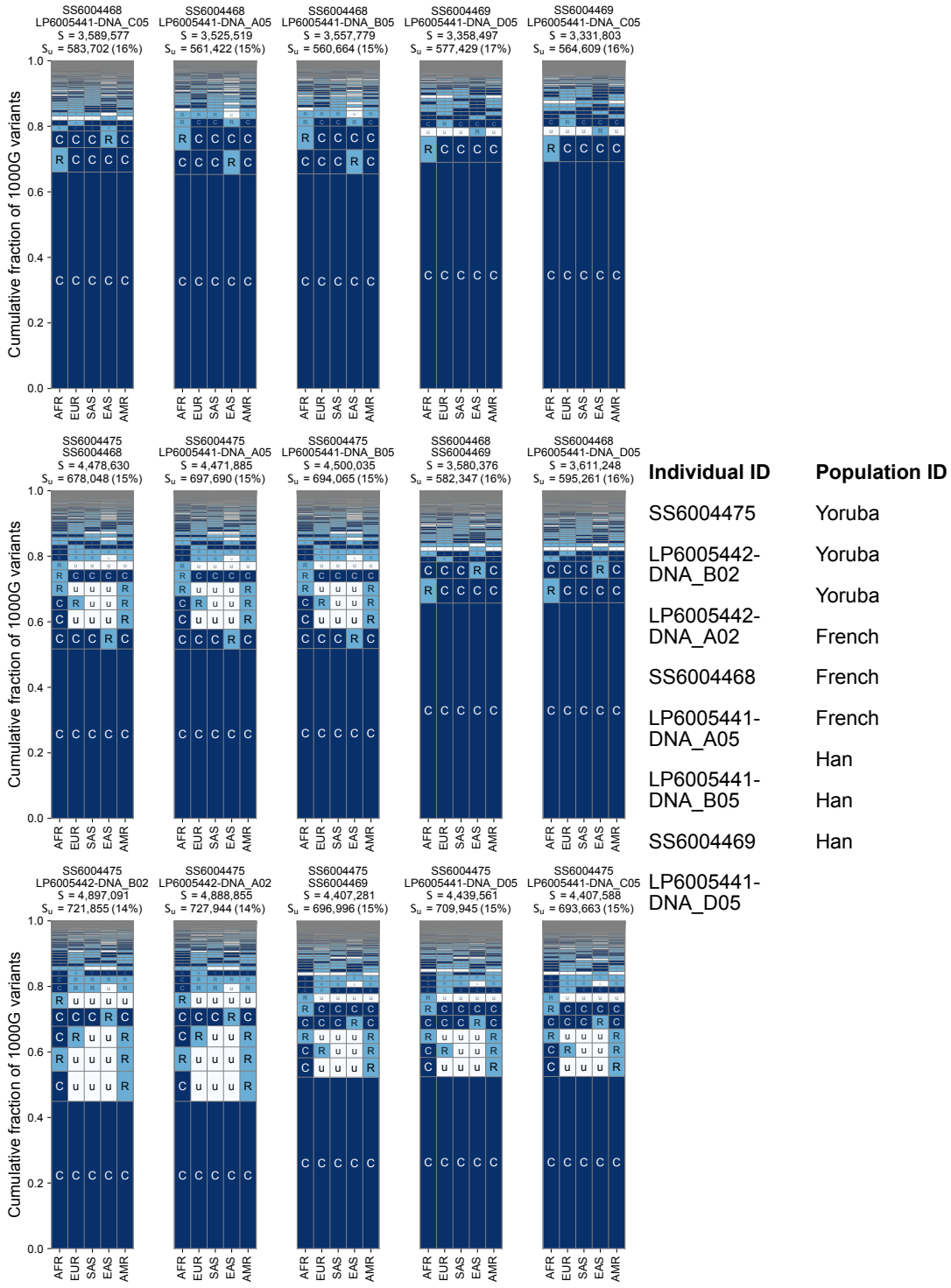

Figure 2.13: Additional examples of geographic distribution codes for pairwise variants from pairs of sampled individuals in the SGDP (Mallick et al., 2016).

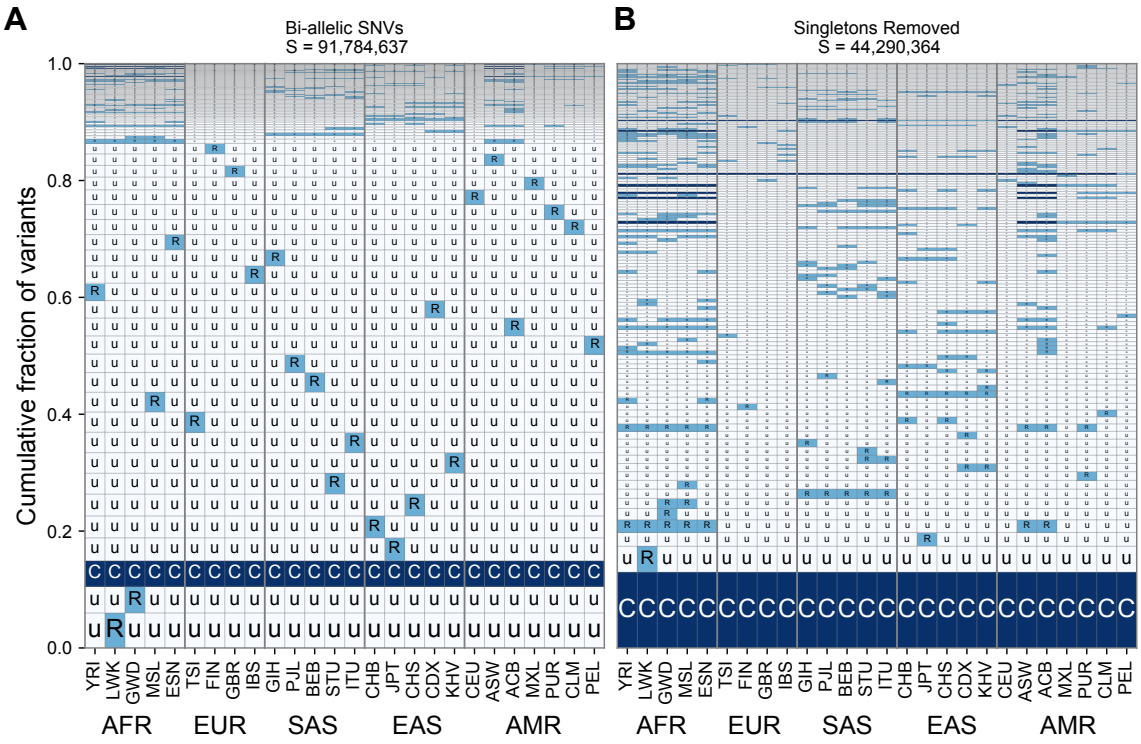

Figure 2.14: The geographic distribution of variants across all 26 populations in the 1KGP both with singletons included **(A)** and removed **(B)** (Auton et al., 2015). Regional groupings are provided on the bottom to reflect our choices for population groupings throughout the main text.

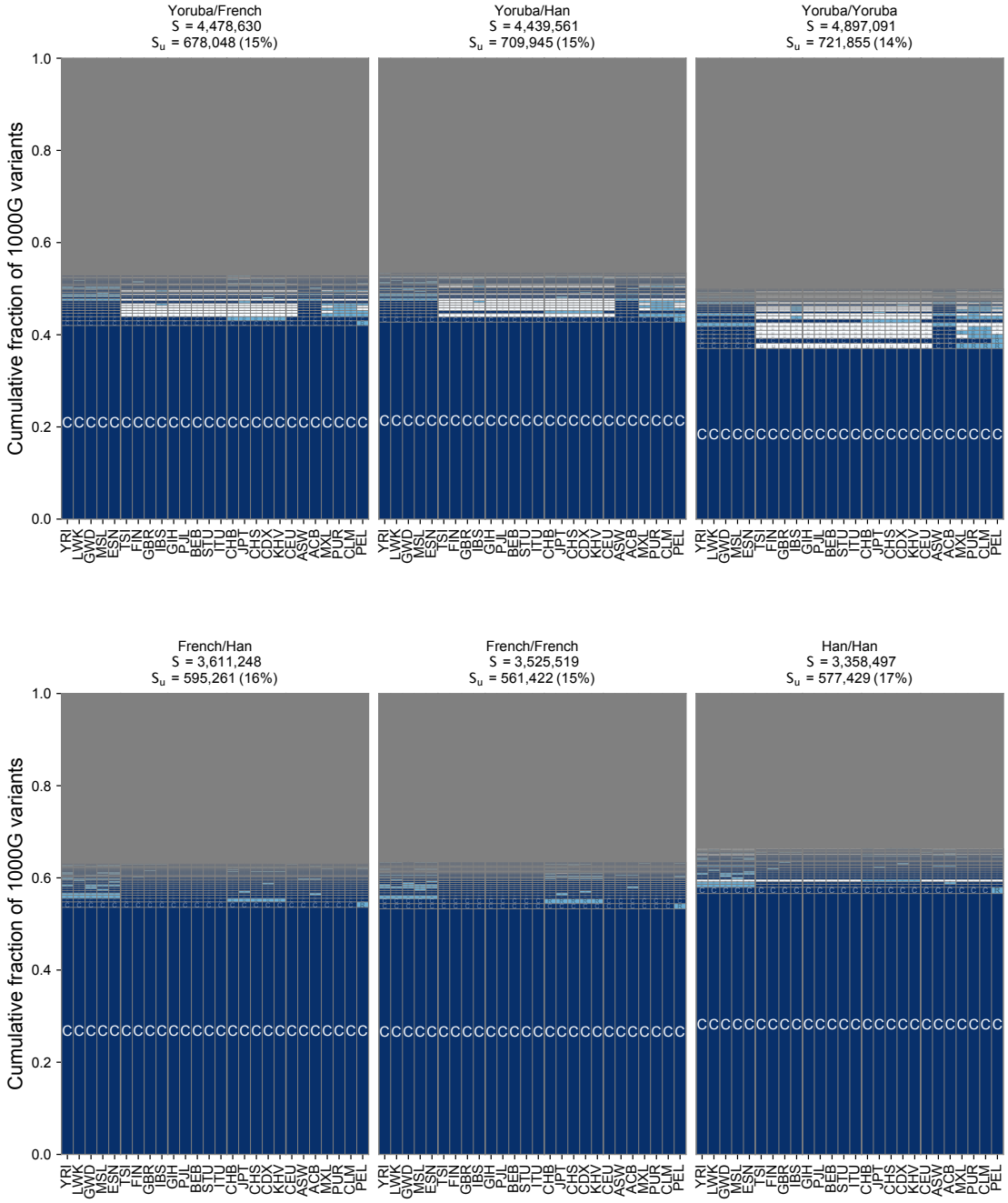

Figure 2.15: The geographic distribution of pairwise SNVs across pairs of individuals from the Simons Genome Diversity Project using the full set of 26 populations from the 1KGP.

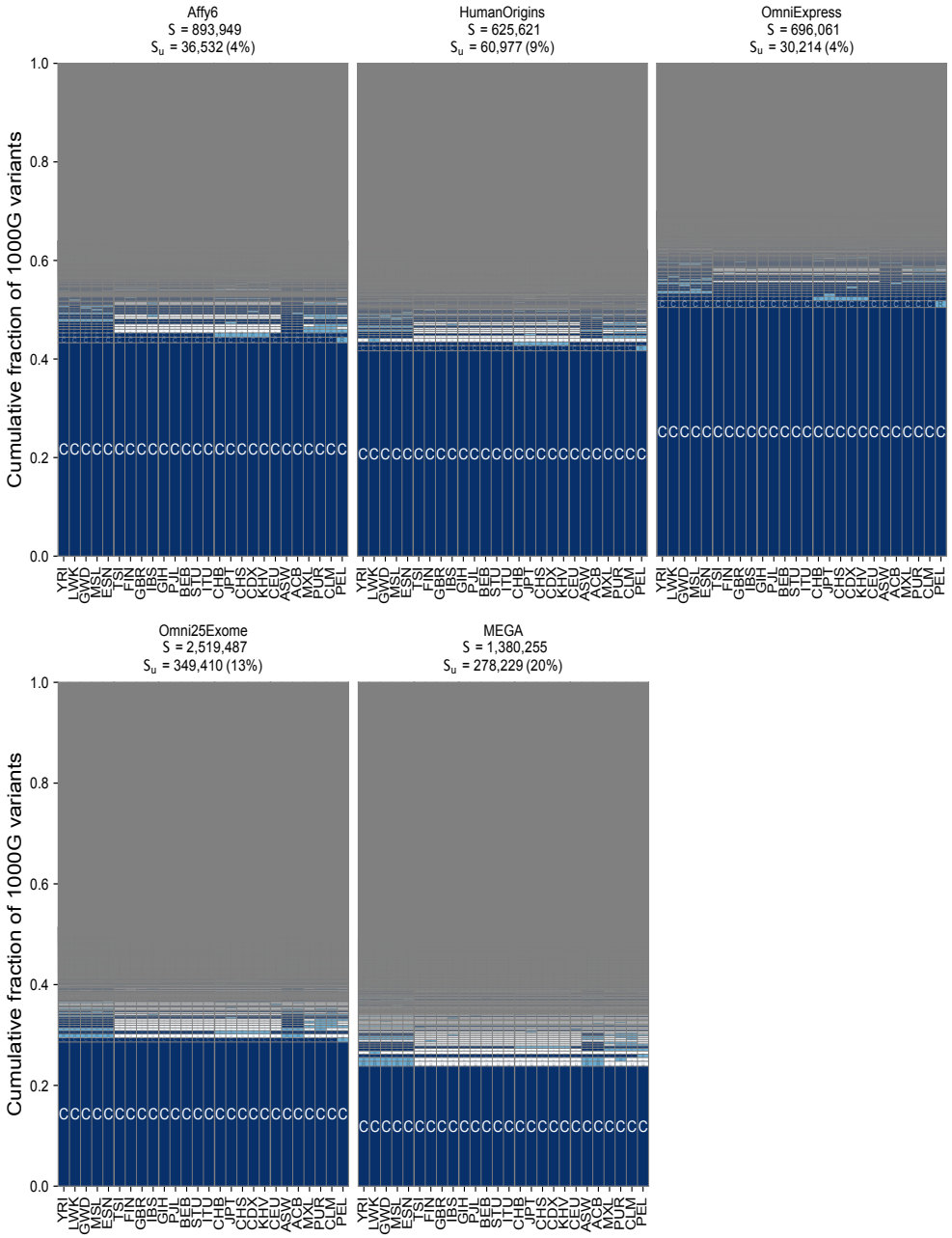

Figure 2.16: The geographic distribution of SNVs on genotyping arrays using the full set of 26 populations from the 1KGP. Note that the distribution is heavily dominated by rare categories that are filtered in our visualization scheme.

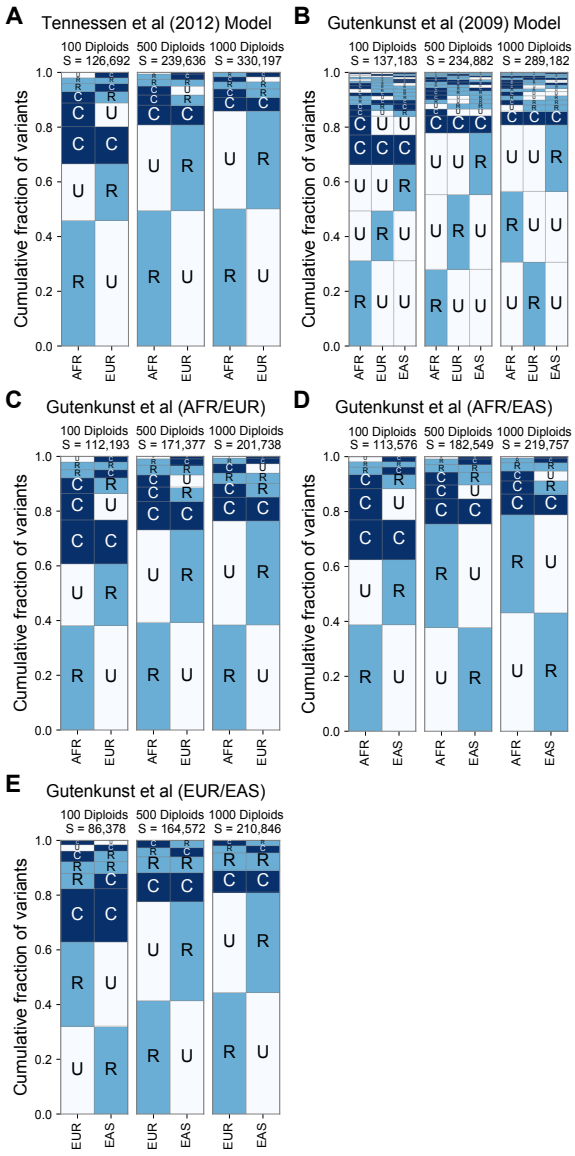

Figure 2.17: **A:** Gutenkunst et al. (2009) **B-E:** Tennessen et al. (2012). For each model, we used `stdpopsim` to simulate 10 replicate loci equivalent to five percent of Chromosome 22 (Adrion et al., 2020). The panels with N = 500 diploid samples per population most closely match the sampling within the 1KGP ($N_{AFR} = 504, N_{EUR} = 503, N_{EAS} = 504$). Both models replicate the qualitative prevalence of the "localized rare" ('RU') and "globally common" ('CC') patterns that we see in the 1KGP data. With higher sample sizes we find an increased proportion of localized rare ('RU') patterns, due to increased detection power. Panels **C**–**E** show specific pairwise comparisons of populations in the model of Gutenkunst et al. (2009) to compare against the two-population model of Tennessen et al. (2012). The prevalence of localized rare and globally common patterns is reproduced across all comparisons, as is the dependence on sample size.

CHAPTER 3

PROPERTIES OF TWO-LOCUS GENEALOGIES AND LINKAGE DISEQUILIBRIUM IN TEMPORALLY STRUCTURED SAMPLES

3.1 Abstract[1]

The rise of ancient DNA studies in population genetics has been revolutionary, revealing insights into demographic history and recent positive selection. However, most studies to date have ignored the non-random association of genetic variants on haplotypes (i.e., linkage disequilibrium, LD). Basic properties of LD in samples with different sampling times are still not well understood. Here, we derive several results for summary statistics of haplotypic variation under a serial sampling model: 1) The expected number of pairwise differences between time-staggered samples ($\pi_{\Delta t}$) and the correlation of this statistic between two loci, in models with and without strict population continuity; 2) The expected value for the time-staggered analog of the LD-metric σ_D^2, which can be interpreted as a measure of haplotypic similarity between a modern and ancient sample; and 3) The expected switch rate in a haplotype copying model, which has implications for how to improve genotype imputation and phasing with ancient samples and modern reference panels. Overall, these results provide a characterization of how haplotype patterns are affected by the time-gap between sampling, recombination rate, and population size. We expect these results will help guide thinking and analysis of haplotype data from ancient and modern samples.

1. Citation for chapter: Properties of Two-Locus Genealogies and Linkage Disequilibrium in Temporally Structured Samples. Arjun Biddanda, Matthias Steinrücken, John Novembre. *In Preparation*

3.2 Introduction

Multi-locus properties of genetic variation have been useful for studying evolutionary processes and maximizing the information extracted from population genetic data. Patterns of multi-locus variation are shaped by mutation and recombination events, generating novel combinations of alleles on chromosomes (i.e., haplotypes). A frequently used summary of haplotype patterns is the covariance in allelic state at two (or more) loci, known as linkage disequilibrium (LD) (R. C. Lewontin and Kenichi Kojima, 1960; Hill and Robertson, 1968; Slatkin, 2008). The decay of LD as a function of the distance between genetic variants plays an important role in dating evolutionary events (e.g. Moorjani et al., 2016), determining the accuracy of complex trait prediction (e.g Vilhjálmsson et al., 2015) and moderating the power to map trait associated loci (Spencer et al., 2009; Wray, 2005).

Several approaches have been impactful for modeling variation at multiple loci. One is through the lens of coalescent theory (Kingman, 1982; Hudson, 1985). With multiple linked loci, the coalescent process involves both recombination (splitting) and coalescence (joining) of ancestral lineages, which means that there can be a different number of lineages at each locus at a given point in time (Hudson, 1985; Simonsen and Churchill, 1997; Richard Durrett, 2002).

Based on the two-locus coalescent models, Hudson (2001) developed a composite likelihood approach to estimating fine-scale recombination rates in early sequencing datasets. This initial attempt has paved the way for subsequent methods to estimate fine-scale recombination rates in humans, accomodating increasing model complexity (McVean et al., 2004; Auton and McVean, 2007; Kamm et al., 2016). Additionally, McVean (2002) showed that an approximation to the r^2 metric of LD, σ_d^2, is related to the correlation in coalescent times between two loci (Hill and Robertson, 1968; Hudson, 1985). The generality of the relationship of σ_d^2 to two-locus coalescent times is key for intuiting the impact of demographic history and sampling design on expected patterns of LD (McVean, 2002; Wakeley and Lessard, 2003).

Both of these historical examples show the utility of theoretical developments of two-locus coalescent models

A second major modeling framework has been haplotype-copying models, such as the Li & Stephen's (LS) model (Li and Stephens, 2003). Haplotype-copying models provide a computationally efficient approximation to compute likelihoods for observed haplotype data generated with recombination (Fearnhead and Donnelly, 2001). The haplotype copying model is also the backbone of many computational tasks in the analysis of population-genomic data, such as genotype imputation(e.g. Howie et al., 2009), computational phasing (Loh et al., 2013, 2016), and ancestry inference (Price et al., 2009; Lawson et al., 2012).

In an increasing number of settings, samples are not all taken from the same time point. This is particularly motivated in settings such as experimental evolution and the growing study of ancient DNA (aDNA) (Slatkin and Racimo, 2016; Skoglund and Mathieson, 2018). For single locus data, genealogical models have been developed to quantify the impact of ancient samples on population genetic statistics such as the expected site-frequency spectrum or the number of variants private to an ancient sample (Rodrigo and Felsenstein, 1999; Forsberg et al., 2005). Recent work has also shown the impact of sample age on expected values of F_{ST}(Ortega-Del Vecchyo and Slatkin, 2018), which can also be interpreted in genealogical terms (Slatkin, 1991). In contrast to the single-locus, for multi-locus genealogical models, the impact of time-separation on patterns of linked variation has not been fully explored.

Here we characterize the impact of temporal sampling on patterns of haplotype variation from a genealogical perspective. Analogous approaches for time-stratified samples in a coalescent framework have generally not been developed for the case of two or more recombining loci. One exception is the approach of (Dialdestoro et al., 2016) that uses importance sampling over the space of latent ancestral recombination graphs when calculating the likelihood of observed sequence data for haplotypes at multiple time-points. Our work here contrasts to that of (Dialdestoro et al., 2016) in that we obtain analytic solutions for the two-locus

scenarios considered and derive results on the effectiveness of a haplotype-copying model. The work presented here is complementary to previous work by Terhorst et al. (2015) who modeled how allelic configurations change over time for multiple loci through a Gaussian approximation to the Wright-Fisher model. The focus of that work was on the frequency trajectories of alleles, rather than the pairwise statistics and haplotype copying model properties that we focus on here.

We first show how time-stratified sampling affects the joint properties of genealogies at two loci, demonstrating that the time gap between a pair of samples has an impact on the rate at which the correlation of tree statistics (tree height and length) decays as recombination distance increases. We also provide a detailed analysis of the Li & Stephen's haplotype copying model with samples of different ages, in particular when the test haplotype is from a different time-point than the haplotype panel. Overall, our results show the effect of time-stratified sampling on expected patterns of haplotypic variation, and their implications for the further development of population genetic methods.

3.3 Results

3.3.1 Two-locus genealogical properties under serial sampling

To understand the impact of serial sampling on patterns of haplotype variation, we first investigated joint genealogical properties at two loci with serial sampling. We use a continuous time Markov process which models the evolution of ancestral lineages in a two-locus system (Hudson, 1983, 1990; Simonsen and Churchill, 1997) (Figure 3.24). This model is a continuous time Markov chain with an absorbing state where *both* haplotypes have coalesced at *both* loci.

Previous analyses with this two-locus ancestral process treat all of the sampled haplotypes as contemporaneous, which is one step removed from our interest in time-stratified sampling.

With the time gap in sampling, there are two natural phases in the ancestral process: (1) the time between the present and when the ancient haplotype is sampled ($t < t_a$), i.e., when only the lineage of the extant haplotype must be traced, and (2) the time when both haplotypes (modern and ancient) are evolving through the full state-space of the process ($t \geq t_a$).

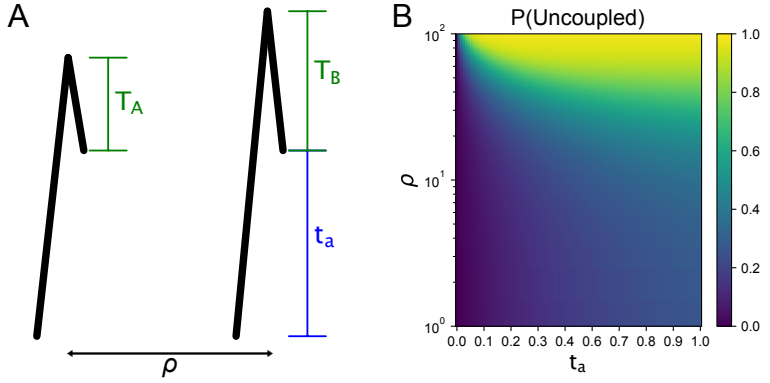

Figure 3.1: **A.**Schematic of genealogies at two loci separated by a population-scaled recombination distance ρ ($\rho = 4N_e r$). The parameter t_a represents the sampling time of the haplotype (measured in coalescent units, i.e., scaled by $2N_e$). The random variables T_A and T_B are the additional time to coalescence at locus $A\&B$, after t_a. **B.** The probability of modern haplotype being "uncoupled" at the time of ancient sampling as a function of t_a and ρ. In this setting, "uncoupled" means that the ancestral lineages at locus A and B are not on the same haplotype, enhancing the probability of different T_A and T_B occurring at each locus.

Within this two-phase model, we derived expressions for the covariance in the T_{MRCA} and total branch length L at two loci (A and B) separated by a population-scaled recombination distance, $\rho = 4N_e r$. We derive expressions for these quantities in the case of one haplotype at present and one ancient haplotype at a known sampling time in the past (see Appendix 3.9.1). We intentionally focus our theoretical results on the case of two-samples at two-loci because it represents the simplest case of serial sampling across multiple loci, is analytically tractable compared with higher sample sizes (Richard Durrett, 2002), and can sufficiently provide insight on expected patterns within data (McVean, 2002).

A key aspect of the model is the effect of recombination within the first phase of the process, when only the modern lineage is evolving backwards in time ($t < t_a$). During this

phase the process has only two states, "uncoupled" and "coupled". By "coupled" we mean that the ancestral lineages are evolving independently at each locus, rather than a single joint. We derive the probability that the ancestral lineages at both loci are uncoupled at $t = t_a$, because this determines the starting probabilities for the phase when both haplotypes are in the process ($t \geq t_a$). Specifically, we obtain the time-dependent probability of being in the uncoupled state by exponentiation of the 2×2 rate matrix for the reduced state-space of the ancestral process during this epoch (Figure 3.1B).

$$\begin{aligned} \mathbf{Q} &= \begin{bmatrix} -\frac{\rho}{2} & \frac{\rho}{2} \\ 1 & -1 \end{bmatrix} \\ \mathbb{P}_{t_a}(\textit{uncoupled}) &= \left(e^{\mathbf{Q}t_a}\right)_{0,1} \\ &= \frac{\rho(1 - e^{-t_a(\frac{\rho}{2}+1)})}{\rho + 2} \end{aligned} \tag{3.1}$$

Intuitively, we see that for large time-separations or high-recombination rates it is likely that the modern haplotype is in the uncoupled state by the time the process begins to include the ancient haplotype. To combine our results for the first phase ($t < t_a$) with the results of the process with both lineages ($t \geq t_a$), we use known results regarding the two-locus ancestral process for two haplotypes, as the remaining evolution is similar to the two-locus process with two contemporary haplotypes, with the evolution in the first phase setting the initial state probabilities for the Markov chain (Simonsen and Churchill, 1997; Richard Durrett, 2002; McVean, 2002). Using this approach, in the next two sub-sections we derive properties of observable quantities from time-staggered haplotype data.

3.3.2 Correlation in pairwise differences

The joint numbers of pairwise differences at two recombining loci under serial sampling is a basic feature of time-sampled sequence data. To understand it in greater detail, we focus on

a model of two non-recombining loci (loci A and B), with recombination occurring at a rate ρ between them. Within each single locus, we model the number of pairwise differences as a Poisson process with rate $\frac{\theta}{2}$, where $\theta = 4N_e\mu L$, where μ is the per-basepair per-generation mutation rate, L is the size of the locus (in basepairs), and N_e is the effective population size using an infinite-sites model assumption. The correlation in the number of pairwise differences at locus A and B is related to the correlation in the total branch length between the loci (Wakeley and Lessard, 2003; Hobolth et al., 2019).

Applying the two-phase approach described above (also see 3.5 for details), we find the correlation in pairwise differences between two loci can be written as:

$$Corr(\pi_A, \pi_B) = \frac{1}{1 + \frac{2+t_a}{2\theta}} Corr(L_A, L_B) \tag{3.2}$$

Where:

$$\begin{aligned} Corr(L_A, L_B) &= \mathbb{E}[T_A T_B] - 1 \\ &= \frac{\rho(1 - e^{-t_a(\frac{\rho}{2}+1)})}{\rho + 2} \frac{\rho^2 + 13\rho + 24}{\rho^2 + 13\rho + 18} + \left(1 - \frac{\rho(1 - e^{-t_a(\frac{\rho}{2}+1)})}{\rho + 2}\right) \frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18} \end{aligned}$$

$Corr(L_A, L_B)$ is the correlation in total branch length, and $\mathbb{E}[T_A T_B]$ (derived in Appendix 1 Eqn. 3.4) is the joint expectation of the time to coalescence *after* both the ancient and modern lineage are allowed to coalesce with one another ($t \geq t_a$). The other expressions are derived in previous results for the two locus, two-haplotype ancestral process (see Appendix 3.9.1) (Simonsen and Churchill, 1997; Richard Durrett, 2002).

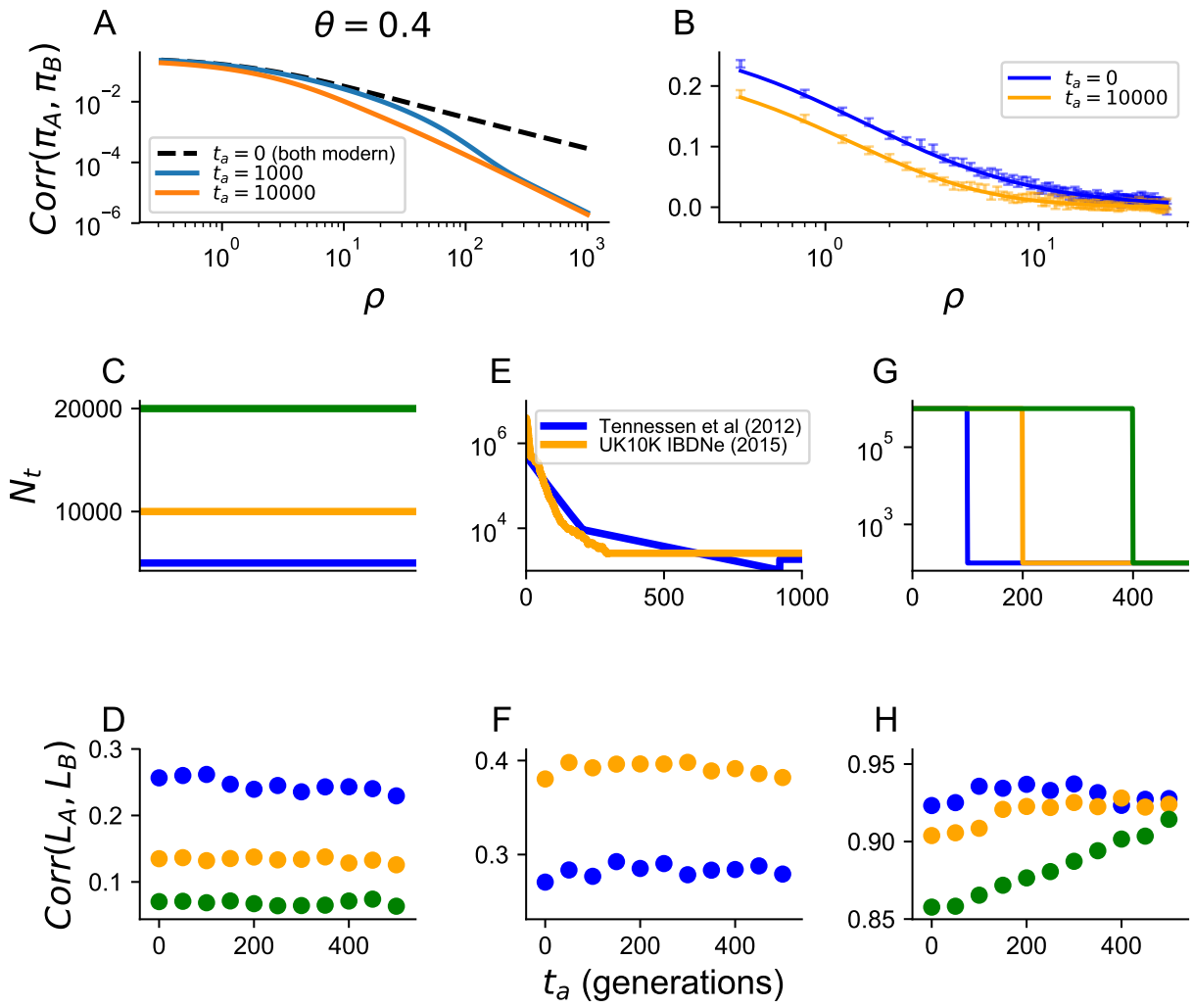

Figure 3.2: **(A,B)**. Correlation in pairwise differences in a model of constant population size. Simulated values of parameters are $N_e = 10^4, r = 10^{-8}, \mu = 10^{-8}$. The length of an independent locus is 1 kb ($\theta = 0.4$), and we tested across two timepoint ($t_a = 0, 10^4$) for the ancient sample. **(C,D)** Estimation of the correlation in branch length at two loci for different population sizes (N_t) for a constant recombination rate. We note that as t_a increases, the correlation in total branch length tends to decrease due to increased opportunity for recombination between the modern and ancient sample. **(E,F)** Simulated values of the correlation in total branch length as a function of sample age in models of recent population growth from Tennessen et al. (2012); Browning and Browning (2015). **(G,H)** The impact of rapid population growth on the correlation in branch length at two loci. We particularly note that the scale of the correlation in branch length is substantially higher than in all of the other models considered, due to the restriction on the coalescent times. In all simulations performed, (**D,F,H**) we held the recombination rate at $r = 10^{-4}$ and simulated 50000 replicate simulations before calculating $\widehat{Corr}(L_A, L_B)$.

The correlation in pairwise differences has two qualitative forms of dependence on the age of the ancient sample t_a. The first dependency is that as t_a increases, the value for the correlation in pairwise differences when $\rho = 0$ (intercept) is smaller than in the case of $t_a = 0$. This is due to the constant-term in (3.2) that decreases as t_a increases and is not dependent on ρ. The second effect is that the *rate* of decay to $Corr(S_A, S_B) \to 0$ is affected by t_a, due to the fact that increasing t_a also increases the probability of the modern haplotype becoming uncoupled (Figure 3.2A). As t_a becomes very large, the decay rate is $\mathcal{O}(\rho^{-2})$ as opposed to $\mathcal{O}(\rho^{-1})$ when $t_a = 0$ (see Appendix 3.9.1 for a derivation). This is primarily because of the additional time (t_a) that the recombination process has to break apart the shared genealogical history at each locus.

Complex demography and population divergence

One important deviation from the assumptions of the theoretical model is variation in population size through time. To explore this, we simulated haplotype data under two models of European population history which features recent exponential population growth and a more ancient bottleneck (Tennessen et al., 2012; Browning and Browning, 2015) . For relatively recent t_a, we expect the relatively high population size in the recent past to favor uncoupling in the first phase, and bottleneck in the past to encourage coalescence and thereby increasing correlation. When comparing our constant-sized theory with ancient haplotypes to these simulations, using appropriately fit estimates of N_e for parameters (see section 3.9.1), we find that our theory under constant population size is unable to capture the qualitative features of the decay in the correlation of pairwise differences under the realized model. Most notably there is an underestimation of the correlation in segregating sites at lower recombination rates relative to the simulated values (Figure 3.7C). We find that holding the recombination rate constant, the effect of growth followed by a population bottleneck increases the correlation in shared branch length relative to the case with constant

population size (Figure 3.2D,F) (Tennessen et al., 2012; Browning and Browning, 2015). In extreme cases of population growth (Figure 3.2H) the correlation in total branch length is quite high, suggesting the ratio of the recombination rate and the coalescence rate at the time of ancient sampling is an important parameter governing the correlation in pairwise differences.

We additionally investigated how population divergence can affect the correlation in pairwise differences at different recombination distances. Under a simple model of population divergence without post-split gene flow, *both* the ancient and modern haplotypes can become uncoupled prior to any possibility of inter-haplotype coalescence (Appendix 3.9.1) thus lowering the overall correlation in pairwise diversity. In simulations we confirm that the decay in the correlation in number of pairwise differences increases as a function of the divergence time (t_{div}) as well as the sampling time (t_a) (Figure 3.8).

Estimating sampling time from the correlation in pairwise differences

While our main goal is to gain understanding of the effect of serial sampling on joint genealogical properties, here we explored the potential for parameters inference. Assuming that the recombination rate and mutation rate are known quantities, we investigated whether one can estimate two parameters: the sampling time t_a and the effective population size N_e. We employed a least-squares estimation procedure of our theoretical expectation for the correlation in pairwise differences against the values estimated from simulations (Figure 3.9). When fitting our procedure, we discretized recombination distance into automatically determined bins in the range of $[10^{-5}, 10^{-3}]$ Morgans using the `histogram` function in `numpy`(Harris et al., 2020).

We find that in the constant population size case N_e is estimated to within the correct order of magnitude and the estimates of t_a monotonically increases with t_a, showing that the correlation function does contain some information about the sample age. However, under

the constant population size, we find that there is upward bias in estimates of N_e and low accuracy in estimation of t_a (for 2 of 4 time points the true value is not contained within 2 standard deviations of the estimated values). The estimation performs increasingly poorly when data is simulated under a model of recent European growth. The estimated sampling time become severe underestimates (e.g. for a sampled haplotype 1000 generations ago, we estimate a $\hat{t}_a = 0$).

Applications to ancient whole genome sequencing data

Finally, we explored the correlation in segregating sites statistics in modern and ancient human whole-genome sequencing data. We restricted ourselves to high-quality whole genome sequencing data to avoid ascertainment biases and more accurately estimate segregating sites for small windows (1 kilobase) in the genome. We masked centromeres and low-mappability regions to avoid biases that can artificially pollute the number of segregating sites (Auton et al., 2015). Specifically we chose two samples at different ages. The first sample we chose is an approximately seven thousand year old sample from modern-day Germany associated with Linear Ban Keramic culture and thus designated the Stuttgart *LBK* sample or simply the *LBK* sample (Lazaridis et al., 2014). The second sample is an approximately forty-five thousand year-old sample from Western Siberia, called *Ust-Ishim* (Fu et al., 2014). We chose these samples to represent an order of magnitude difference in the sampling time-scale (thousands vs. tens-of-thousands years). For a modern comparison, we calculated the correlation in segregating sites using 108 CEU haplotypes from the 1000 Genomes Project (Auton et al., 2015) and taking the average correlation (see Table 3.1 for additional choices of modern focal individuals)

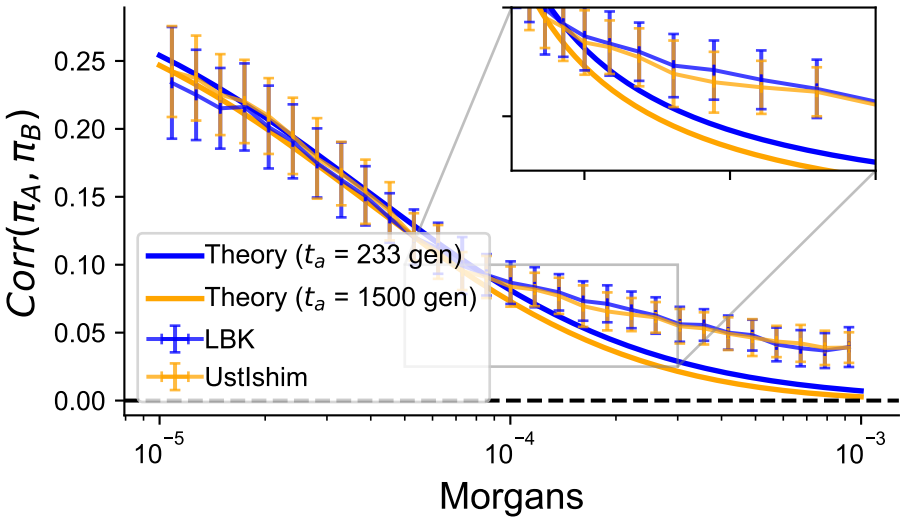

Figure 3.3: Comparison of the correlation in pairwise differences between empirical modern and ancient data. When computing the theoretical curves, we used $N_e = 10^4$ and a mutation rate $\mu = 1.2 \times 10^{-8}$ per basepair-per-generation. When computing the number of generations in the past on which to base our theory we used a generation time of 30 years per generation (Fenner, 2005). We also used 30 log-spaced bins over the range $r = (10^{-5}, 10^{-3})$ to calculate Monte-Carlo estimates of the correlation in pairwise differences.

We find that the correlation in pairwise differences as a function of recombination distance for *Ust-Ishim* is not significantly different from *LBK* (Fig. 3.3, Binomial Test, $p = 0.361$). The variance in the observed correlation in pairwise differences per unit of recombination distance is larger than the difference in the theoretical means (solid lines, Fig. 3.3). Qualitatively, we find that there is a lack of fit with the constant population sized theory at longer recombination length scales. We speculate that this may be due to the effects of population bottlenecks or recent growth in the history of non-African populations leading to elevated levels (Reich et al., 2001; Kamm et al., 2016; Ragsdale and Gravel, 2019) of linkage disequilibrium at this scale and is not well-captured by our theoretical model of constant population size (Figure 3.10).

3.3.3 Expectations of linkage disequilibrium with time-stratified sampling

The joint genealogical history is also related to the co-variation in allelic state at two loci, or linkage disequilibrium (LD). Specifically, the r^2 metric of linkage disequilibrium which is defined as $r^2 = \frac{D^2}{p(1-p)q(1-q)}$, where p, q are the frequency of the derived allele at locus A and B and $D = x_{11} - pq$, and x_{11} is the frequency of the haplotype with both derived alleles. One approximation made is to treat $\mathbb{E}[r^2] \approx \frac{\mathbb{E}[D^2]}{\mathbb{E}[p(1-p)q(1-q)]} = \sigma_d^2$ (McVean, 2002). This approximation is approximating the expectation of a ratio, with the ratio of expectations.

Specifically, the term σ_d^2 is a ratio of terms computed from the covariance in coalescent times of a sample of haplotypes (McVean, 2002; Richard Durrett, 2002). However, there are cancellations in both the numerator and denominator of σ_d^2 that allow one to compute it using expected properties from just pairs of haplotype configurations and avoid considering the sample in its entirety. One can think about this as the expected correlation in the allelic states of markers a particular recombination distance apart. Indeed the realized value of $\hat{r}^2$ can be calculated as the squared Pearson correlation coefficient of genotypes in a collection of sampled chromosomes (VanLiere and Rosenberg, 2008; Rogers and Huff, 2009).

Similar to previous characterizations of σ_d^2, we can theorize how much this metric may change under time-staggered sampling. Using our approach in the section above of a staggered ancestral process for time-staggered sampling to compute the joint expectations of coalescent times (see Appendix 3.9.2 for full derivation), we arrive at an expression for σ_d^2 with a time-staggering of t_a.

$$\begin{aligned} \gamma &= \frac{\rho(1 - e^{-t_a(\frac{\rho}{2}+1)})}{\rho + 2} \\ \sigma_d^2 &= \frac{\mathbb{E}[D^2]}{\mathbb{E}[p(1-p)q(1-q)]} \\ &= \frac{\frac{\rho+10}{\rho^2+13\rho+18}}{(1-\gamma)\left(\frac{\rho^2+13\rho+22}{\rho^2+13\rho+18}\right) + \gamma\left(\frac{\rho^2+13\rho+24}{\rho^2+13\rho+18}\right)} \end{aligned} \tag{3.3}$$

There are two observations from this result that shows the qualitative behavior of how temporal sampling affects σ_d^2. The first observation is that the numerator of the expression is constant with respect to changes in t_a, that is $Var(D)$ or $\mathbb{E}[D^2]$ does not change mathematically. This can be explained by the fact that time-stratified sampling shifts each individual expectation to be a weighted mixture of the contemporary expectations where the weights cancel out in the numerator (see Appendix 2 for a detailed derivation).

The second observation is that the denominator is a mixture of two components (one where the time-separation is not large enough for the modern lineages to have coalesced, and the other where the ancestral lineages at both loci are separated). For example, by setting $t = 0$ we obtain the expression for σ_d^2 with two contemporary samples in a constant population (Appendix 2) (McVean, 2002; Richard Durrett, 2002).

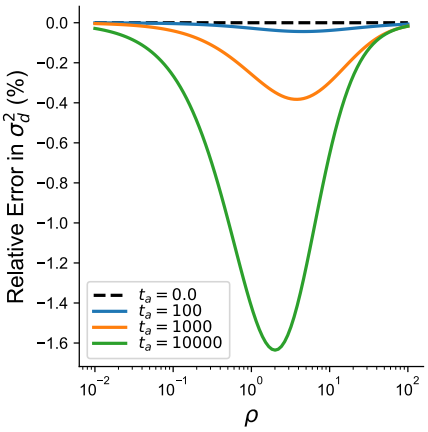

Figure 3.4: Relative error between the time-stratified approximation to σ_d^2 and using contemporary samples. The relative error is measured as $\frac{\sigma_d^2(t_a)-\sigma_d^2(0)}{\sigma_d^2(0)}$, which measures the fractional difference of the expectation of σ_d^2 under serial sampling as opposed to haplotypes sampled from the present.

We find that temporally stratified sampling does not affect σ_d^2 substantially (relative error $< 2\%$ for all values of ρ). This is also apparent as $t \to \infty$ (Appendix 3.9.2, Figure 3.11). The statistic $\sigma_d^2(t)$ can be interpreted as an approximation to the expectation of the

squared correlation coefficient between genotypes of samples t_a coalescent units apart. To corroborate these theoretical results, we also conducted coalescent simulations in a model of constant population size, with equal numbers of ancient and modern haplotypes ($n = 100$ ancient and $n = 100$ modern haplotypes). We calculated $\mathbb{E}[r^2]$ using Monte-Carlo sampling across 500 replicate simulations and find similar qualitative patterns regarding the decay in r^2 as we predict from our analysis of σ_d^2 with time separation above (Figure 3.12).

3.3.4 The impact of serial sampling in haplotype-copying models

Haplotype copying models are models used to capture patterns of similarity among the haplotypes in a reference sample and a test haplotype (Li and Stephens, 2003; Lawson et al., 2012; Lunter, 2019). The underlying statistical model for haplotype-copying models is also known as the Li-Stephens model(Li and Stephens, 2003), which is an approximation to the full genealogy of a set of sequences subject to recombination, the ancestral recombination graph (ARG). We use a slight modification of the original haplotype copying model throughout our analysis in this section (Li and Stephens, 2003; Lawson et al., 2012) (details in Appendix 3.9.4). Our modifications makes the assumption that the recombination map is known and parameterizes the jump rate in terms of this known map, identical to the model of (Lawson et al., 2012).

With ancient samples, inference under the haplotype copying model would typically be performed using a modern reference panel. This time-separation provides an opportunity for recombination events to occur among the modern reference haplotypes before the ancient lineage is able to coalesce with any individuals from the modern panel. This is substantiated by our earlier calculations (Equation 3.1) on the increased probability for the modern ancestral lineage to become uncoupled moving farther back in time (Figure 3.1).

Throughout our results for haplotype copying models, we focus on the estimated haplotype copying jump rate ($\hat{\lambda}$) as a function of the sampling time. The key intuition is that this

jump rate is inversely related to the expected copying-tract length in Morgans ($\mathbb{E}[L] \approx \frac{1}{\hat{\lambda}}$), which is a summary statistic of the spatial scale of information contained by linkage disequilibrium in the modern panel that is informative about linked variation in the ancient test haplotype (Li and Stephens, 2003; Loh et al., 2013, 2016).

The joint impact of population demography and sample age on the haplotype-copying jump rate

Our first question is whether the transition rate under the haplotype-copying model increases or decreases as a function of the time-separation between the modern panel and our ancient test haplotype. Subsequently, we are interested on how the underlying population demographic history can modulate this effect to create positive or negative effects on the estimated jump-rate vs. samplimg time ($\hat{\lambda}$ vs. t_a). We calculate the maximum-likelihood estimator of the copying jump-rate $\hat{\lambda}$ in coalescent simulations (see 3.9.4 for details on the model) to address these particular questions (Kelleher et al., 2016). We observe that similar to our theoretical arguments based on two-loci, the transition rate increases as a function of the age of the test haplotype in the case of coalescent simulations with constant N_e (Figure 3.5A). This increase appears to be approximately linear as a function of the sample age in generations for the case of constant population size (Figure 3.5D).

There are two primary effects that we observe for the increase in the transition rate as a function of the age of the test haplotype. The first effect is that the time-separation allows for recombination events to break apart the modern haplotypes before the ancient lineage can coalesce with a member of the modern panel, which makes the possible copying tracts shorter in length. This is also illustrated by the simulation experiment showing that the number of topological changes also increases as a function of the age of the ancient haplotype, due to recombination events breaking up the topological structure of the local genealogy and disrupting the genealogical nearest neighbors (Figure 3.5A).

The second effect is that the reference panel actually is able to coalesce with itself, making the effective copying panel size smaller moving farther back in time. This effect of coalescence within the reference panel is not considered within the original haplotype copying model (Li and Stephens, 2003) When quantifying this effect using the time to first coalescence between an ancient sample and a modern member of the haplotype panel we find that for sufficiently ancient samples the within-panel coalescent process cannot be ignored (see Appendix 3.9.4).

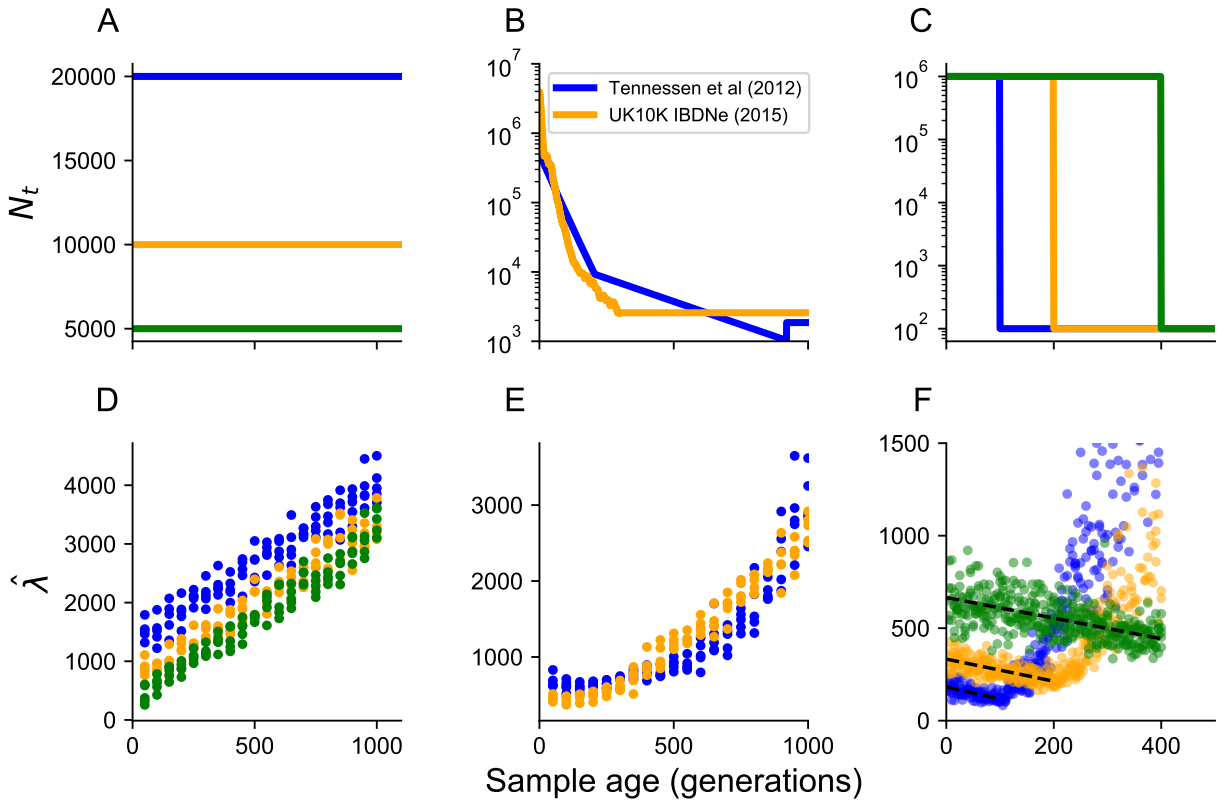

Figure 3.5: Estimation of haplotype copying jump-rate vs sampling time ($\hat{\lambda}$ vs. t_a) in various models of population demographic history. Under constant population size **(A,C)** we find that there is an apparent linear increase in the jump-rate as a function of sampling time. The absolute value of the jump-rate increases with population size as well. All simulations were conducted using chromosomes 40 megabases in length, recombination rate of 10^{-8} per basepair per generation, and a mutation rate of 10^{-8} per basepair per generation. Every modern panel consisted of $K = 100$ haplotypes.

Similar to the correlation in pairwise differences, we explore the effect of previously

inferred models of human population growth on the relationship between the haplotype-copying jump rate and sampling time (Browning and Browning, 2015; Tennessen et al., 2012). We observe across both models (Figure 3.5B,E) that there is an initial decrease in the copying jump-rate as a function of sampling time before a more rapid increase moving back into the past. We interpret this effect as recent population growth initially slowing the increase in the jump rate due to the large population size and limited coalescence events and subsequently increasing due to the bottleneck and long-term smaller population size moving farther in the past (Tennessen et al., 2012; Browning and Browning, 2015).

The observation of apparent suppression in the increase of the jump-rate due to models of population growth (Figure 3.5B,E) motivates the question of whether there are demographic scenarios that can *decrease* the jump-rate as a function of the sampling time. We start with the hypothetical case of an intense bottleneck at time T_{bot} in the past, which requires that *all* ancestral lineages at time T_{bot} instantaneously coalesce at the onset of the bottleneck. The haplotype-copying jump-rate, $\hat{\lambda}$, is directly related to the expected time of coalescence between an ancient haplotype and a member of the modern panel, because recombination events that occur on these branches can initiate copying-switch events (Paul et al., 2011; Steinrücken et al., 2013; Li and Stephens, 2003). In the scenario of an instantaneous bottleneck, we expect that the branch length subtending only the ancient sample will decrease as a function of the sampling time (until the time of the bottleneck) and this is reflected by simulations as well (Figure 3.20).

To address this effect, we simulate a model of instant population bottleneck starting at T_{bot} up to 400 generations ago and constituting a $\times 10^4$ increase in the population size (Figure 3.5C&F). We find that these models of instantaneous growth show a *decrease* in the jump-rate with sampling time, reflective of a strong conditioning on the coalescent time (Figure 3.20). When we reduce the magnitude of growth to be lower, with a $\times 10^2$ increase, we find that negative and positive regression slopes are possible, due to the lower bottleneck

strength and the genealogy behaving less star-like. (Figure 3.21)

3.3.5 Haplotype-copying jump-rates in human ancient DNA

To bridge our simulation experiments on the dependence of the jump-rate with sampling time to applications in human ancient DNA datasets, we applied our jump rate estimation to a collection of 1159 ancient samples typed at $\sim$ 1.24 million markers (see 3.5.2 and 3.5). We use only male X chromosomes to avoid potential errors introduced by statistical haplotype phasing, and used male X chromosomes from Auton et al. (2015) as the modern reference panel.

To avoid the potential effects of population structure (Figure 3.17) confounding the impact of serial sampling on the jump-rate estimation estimation, we focus primarily on samples $<$ 1500 kilometers from an assumed location for Central European (CEU) individuals (although see Figure 3.18 for alternative panel structure). For estimating the jump-rate, we use 49 CEU male X chromosomes to reflect a local modern panel.

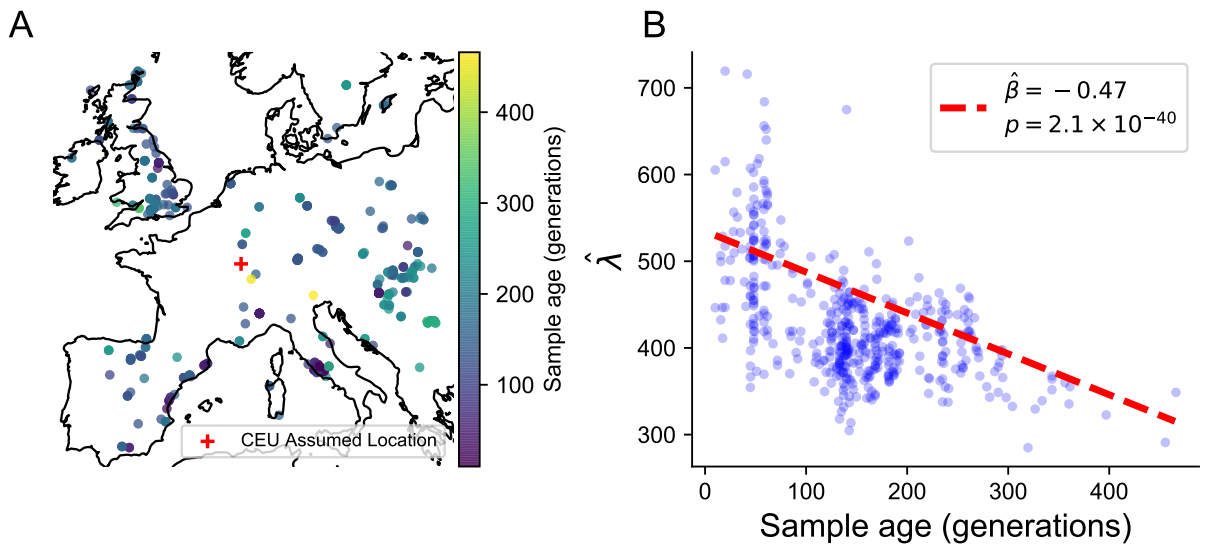

Figure 3.6: **(A)** Map of samples < 1500 kilometers from hypothetical location of central Europe (our proxy for CEU individuals from Auton et al. (2015)). Color represents sample age in generations, assuming 30 years per generation (Fenner, 2005). **(B)** Decrease in estimated haplotype copying jump-rate $\hat{\lambda}$ (per Morgan) as a function of the sample age in generations when estimated from male X chromosomes in panel. Regression was fit using ordinary least-squares with additional terms for latitude, longitude, and their interaction.

Surprisingly, we find that the estimated jump rate *decreases* as a function of sampling time ($p = 2.1 \times 10^{-40}$), contrary to our results in the constant population size case. We interpret this as largely the consequence of a population bottleneck followed by recent rapid growth (Reppell et al., 2014; Tennessen et al., 2012; Browning and Browning, 2015). In addition to our simpler simulations of instantaneous growth and simplified growth above (Figure 3.5C,F), we corroborate this decrease in the jump-rate as a function of time due to recent growth by simulating data under two models exhibiting such rapid human population growth (Figure 3.23, Figure 3.22) (Tennessen et al., 2012; Browning and Browning, 2015). We replicate the temporal sampling structure in the real dataset and use a realistic sex-averaged recombination map for the X chromosome (Kong et al., 2010). We find that we are able to replicate a qualitative decrease in the jump-rate as a function of sampling time across both assumed demographies, similar to that from the real data (Figure 3.6).

The two inferred demographic models differ in several key ways that are important to consider when understanding their effects on the haplotype-copying jump rate inference. The first difference is that the model of Tennessen et al. (2012) is constructed from the sample-frequency spectrum, where the model from Browning and Browning (2015) is estimated using inferred identity-by-descent (IBD) segments (e.g. Zhou et al., 2020). IBD segments by definition contain information about recent demographic history in their length, and are a closer summary statistic to actual haplotype-copying tracts (Palamara et al., 2012; Lawson et al., 2012; Browning and Browning, 2015). A second difference between these two models is the scale and extent of population growth in the recent past (Figure 3.22). Both of the simulations do not capture the temporal extent of the decrease in jump-rate with sample age to ~ 400 generations. However, we do not necessarily expect the models proposed here to reflect the empirical data to a sufficient degree, since the models are unable to capture features such as ancestral population structure that could be responsible for such differences (Kamm et al., 2016; Lazaridis et al., 2014). The empirically observed negative relationship between sample age and the haplotype-copying jump-rate within European samples has important implications on the scale at which genotype imputation and phasing may be applicable for ancient samples and we expand upon these implications in the discussion below.

3.4 Discussion

In this article, we have theoretically investigated the effects of serial sampling in the context of two models: the two-locus ancestral process and the haplotype copying model. We primarily focused on these models (1) because they provide intuition for the expected patterns of linked variation and (2) they are used frequently to model modern haplotype data.

The genealogical properties at two loci are related to expected patterns of mutations on the observed haplotypes. We find that with larger time-separation between samples, the

correlation in the total branch length, and consequently the correlation in segregating sites, decreases faster between two loci ($\mathcal{O}(\rho^{-2}$ vs. $\mathcal{O}(\rho^{-1})$. Intuitively, this is because there is additional marginal branch length on which a recombination event can occur ($1 + t_a$ vs. 1 in expectation) that can disrupt this correlation.

Surprisingly, our two locus results show that serial sampling has negligible effects on σ_d^2, an approximation to the r^2 metric of linkage disequilibrium. In particular the metric σ_d^2, shows $< 2\%$ relative deviation (Figure 3.11) from the expectation with modern sampling (McVean, 2002). We find that the numerator, or the variance in the quantity D is mathematically equivalent to that under modern sampling (see Appendix 3.9.2) and that the difference is entirely due to the denominator. The denominator represents the probability of drawing two haplotypes, one modern and one ancient, that differ at both loci.

One limitation of the theory that we have developed here is that it is derived under an assumption of constant population size. Evidence from simulations (Figure 3.10) suggests that accounting for variable demographic history would provide expectations more in line with results from realized human data sets (Kamm et al., 2016). One potentially promising approach to account for non-constant demography may be to leverage recently developed two-locus "phase-type" theory (Hobolth et al., 2019) which may allow for incorporating demographic history, while retaining the serial sampling aspects of the work presented here. We expect such theoretical developments to be a fruitful avenue for further exploration of multi-locus properties in the coalescent setting.

There are additional statistics related to the genealogical history at two loci that we did not explore within this manuscript that are relevant to the goals of this work. One statistic is "haplotype homozygosity", which is the probability of selecting two identical haplotypes in a sample (Sabatti and Risch, 2002; Fry et al., 2006). One reason this statistic not easily approachable with the framework employed here is that it does not exhibit a similar cancellation of terms as σ_d^2 (McVean, 2002). To obtain these quantities, in principle

one can calculate them using Monte-Carlo estimators from two-locus simulations but this is beyond the scope of the work presented here. Another statistic is the "correlation in zygosity" (Lynch et al., 2014), which is a measure of the deviation of the frequency of pairs of loci with mixed zygosities from random assortment. Our results on the correlation in coalescent times between loci could also be used to calculate the expectations for the correlation in zygosity between two haplotypes sampled at different times (see Eq 8 in (Lynch et al., 2014)). As the derivation for the correlation in zygosity is well defined, we acknowledge that it is a related statistic to those that we display here but did not explore its properties under time-stratified sampling in this manuscript.

While haplotype-based models have been applied to datasets consisting of modern and ancient DNA to understand fine-scale population structure(Martiniano et al., 2017), we have here theoretically characterized the effect that time-stratification can have on parameter inference within the copying model. Our finding that the copying jump rate increases as a function of the age of the sample in populations of constant size can also be thought of in a coalescent interpretation of greater branch length on which recombination events can occur (Jewett et al., 2012). This effect is not limited to the time-stratified case, and can be observed when we have test haplotypes from diverged populations relative to the reference as well (Smith et al., 2018; Jewett et al., 2012).

Many methods have been developed in the context of haplotype copying models, from imputation and phasing(Howie et al., 2009), estimation of recombination rates(Li and Stephens, 2003), to fine-scale ancestry estimation(Lawson et al., 2012). Our theoretical results leave important considerations for each of these application domains with serially-sampled data. For imputation and phasing, the increase in the copying jump rate as a function of time under constant population sizes implies that the extent of linkage disequilibrium will generally be lower in relation to the first coalescent time with a member of the modern panel, and will lower the copying accuracy at long scales (Appendix 3.9.3, (Jewett et al., 2012)). For

samples that are sufficiently old, there is a diminishing benefit for generating larger reference panels as well (Appendix 3.9.3), which results in improvements in imputation and phasing for modern samples due to recent relatedness and sharing of rare variants (Jewett et al., 2012; McCarthy et al., 2016).

Our exploration of the impact of population demography (particularly population growth) and our empirical analysis of male X chromosome paints a more optimistic picture for the analysis of human ancient DNA using the haplotype-copying model. We find that there is a substantial attenuation of the increase in the haplotype-copying jump-rate ($\hat{\lambda}$) under scenarios of recent growth, and even potential decreases in the case of instant population growth (Figure 3.5). Together with our empirical result of the jump rate decreasing as a function of time across male X chromosomes in ancient European samples (Figure 3.6), we would hypothesize that with the caveat of no noise in the data we may be able to impute data with reasonable accuracy due to the specific effect of demographic history. Indeed Gamba et al. (2014) performed an empirical test of imputation accuracy by down-sampling variants on two high-coverage ancient genomes (5,070 - 5,310 and 830-980 years old). Using the haplotype panel from Auton et al. (2015), Gamba et al. (2014) found that the more recent sample had a larger number of variants imputed with ¿ 99% accuracy (80% of sites vs. 78% of sites for the older sample). Martiniano et al. (2017) also found a similar empirical range of imputation accuracy for down-sampled high-coverage genomes, but the relationship between sample age and imputation accuracy is not as clear (see Figure S6 in (Martiniano et al., 2017)). These results, in line with our empirical results on the male X chromosome, suggest that imputation efforts for ancient DNA samples in Europe (up to the past $\sim$ 400 generations ($\sim$ 12,000 years) may be quite accurate, and aid in exploring temporal population structure and the evolution of complex traits.

Our empirical analysis of human ancient DNA data does have some caveats and implications for applications of the haplotype copying model. In our efforts here, we have not

attempted to model genotyping error and low-coverage, which are both common in the analysis of ancient DNA (Dabney et al., 2013). Methods using haplotype-copying HMMs with emission probabilities dealing with low-coverage sequencing data (e.g. Rubinacci et al., 2020) are more applicable to account for this sparsity in ancient DNA analysis. However, we filter to samples at sufficiently high coverage, and we also do not find any significant effects of coverage on the qualitative result that the jump-rate decreases as a function of time (not shown)

Another caveat is that we have focused the majority of our analysis on the aDNA record in western Eurasia. This is largely due to the wide temporal range and the absolute number of samples (Figure 3.16, Olalde and Posth (2020)); however it does carry some differences in interpretation with other regions of the world. One particularly large difference is the analysis of ancient DNA from Africa, where populations have not undergone the Out-of-Africa bottleneck and maintained more stable effective population sizes and higher levels of diversity (Auton et al., 2015; Gutenkunst et al., 2009; Vicente and Schlebusch, 2020). We observe when applying our same analysis of jump-rate estimation using the full set of male X chromosomes from Auton et al. (2015) that the samples from Africa (Figure 3.19) have a substantially higher jump rate, indicating either a larger genealogical distance from our modern panel or that the effect of sample age in this region is substantially greater due to the larger long-term effective population size. We do note that this is fairly weak evidence due to difficulties in obtaining ancient DNA from climates in Africa and its low-representation in the current aDNA record. As ancient DNA technology improves and global sampling becomes less centered on western Eurasia, it will be intriguing to re-analyze this relationship between the jump-rate due to sample age across multiple regions with varied demographic histories.

With the abundance of ancient DNA data being generated across an increasingly wide array of organisms, statistical and theoretical advances will need to similarly account for

this new dimension in the data. Here we have highlighted the impact of serial sampling for two related models, the two-locus coalescent with recombination and the haplotype copying model. We expect that our theoretical treatment of these models will serve to inform advances in statistical population genetic methods that account for serially sampled data to maximize their utility for inference.

3.5 Materials and methods

3.5.1 Quality control of publicly available 1240K Human Ancient DNA Dataset

The human ancient DNA data that we used was typed at a set of 1,233,013 sites in the human genome. Genotypes are drawn using psuedo-haploid sampling based on the available reads. We filter the data subject to several criteria based on available metadata (see 3.5.2).

1. Must be a sample that is determined to be a male
2. Samples must not have a significant amount of modern DNA contamination (e.g. "PASS" contamination checks)
3. Samples must have ≥ 1 non-missing variants per 25 kilobases on average across the X chromosome. Following this filter, we retain samples with a median autosomal coverage (based on the metadata) of 1.54×.

From these analyses we then have a total set of 1159 samples on which we performed our estimation of the haplotype-copying jump-rate. Note that not all of these samples are used for our primary figures as they are not within the spatial location we have chosen, but we have conducted estimation of the jump-rate using multiple reference panels as well (see 3.5.2).

3.5.2 Web Resources

- Publicly-available compiled Human aDNA datasets
- 1000 Genomes X Chromosome Data
- Recombination Maps

3.6 Acknowledgements

We would like to thank all members of the Novembre, Steinücken, and Berg Labs for thoughtful feedback on this work. Particular thanks to Harald Ringbauer and Joe Marcus for detailed discussions and comments on this work. We additionally thank Dr.Sharon Browning for sharing the estimated demography for the UK10K based on IBD segments from their paper. Arjun Biddanda was supported by NIH T32 GM07197. This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

3.7 Supplementary Figures

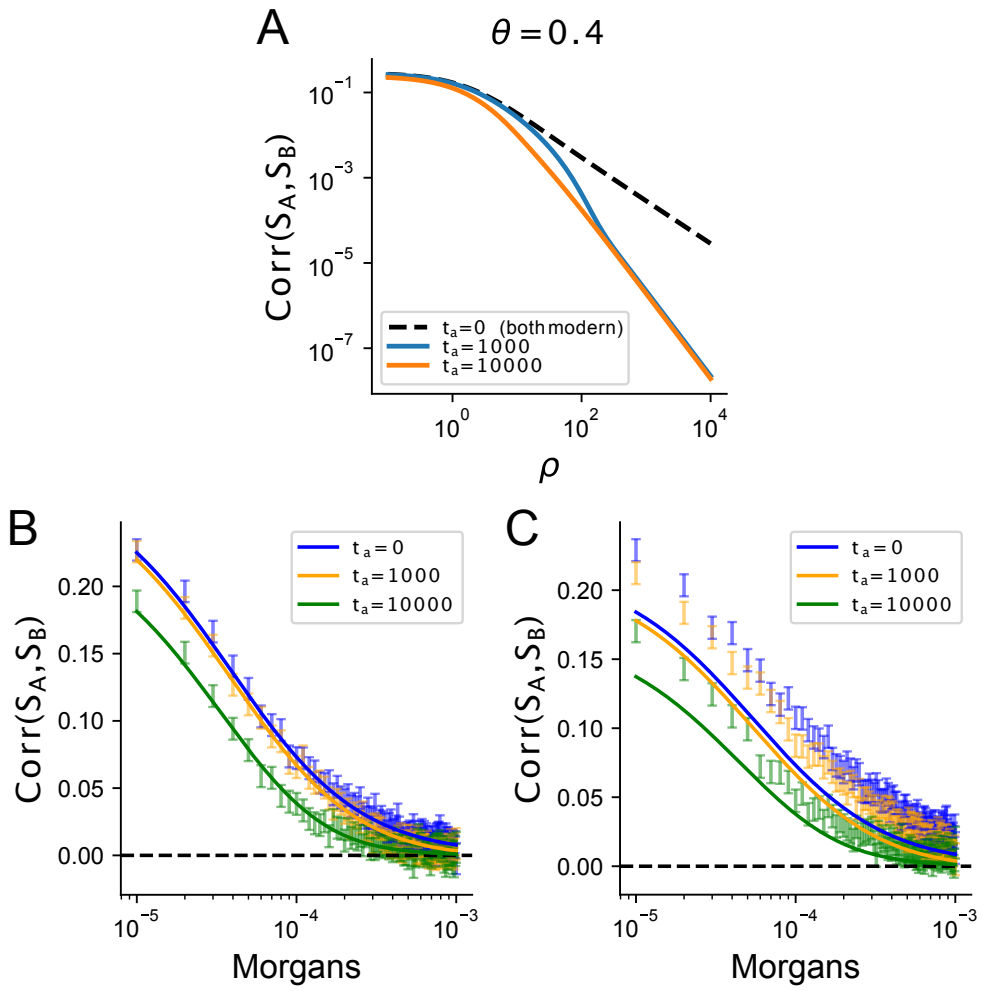

Figure 3.7: **A,B** Correlation in segregating sites in a model of constant population size. Simulated values of parameters are under $N_e = 10^4, r = 10^{-8}, \mu = 10^{-8}$. The length of an independent locus is 1 kb, and we tested across three times for the ancient sample. **C.** When estimating the correlation in segregating sites under a model of recent human population growth from (Tennessen et al., 2012) (only using European samples), our constant-sized theory underestimates the correlation in segregating sites. Monte-Carlo simulations in **(B,C)** were conducted by simulating 20 chromosomes of 20 Mb and calculating the correlation in segregating sites within 1 kb windows separated by a given recombination distance. The underlying mutation and recombination parameters for these simulations are $N_e = 10^4, r = 10^{-8}, \mu = 10^{-8}$.

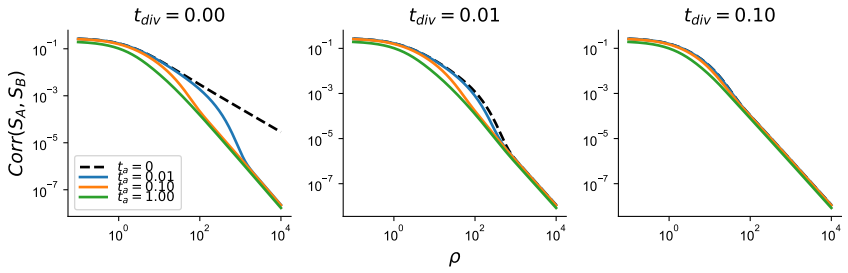

Figure 3.8: Correlation in pairwise differences for samples under a model with divergence at time t_{div} in the past. Locus length simulated is 1 kb with the following parameters $N_e = 10^4, r = 10^{-8}, \mu = 10^{-8}$. ($\theta = 4 \times 10^{-1}$).

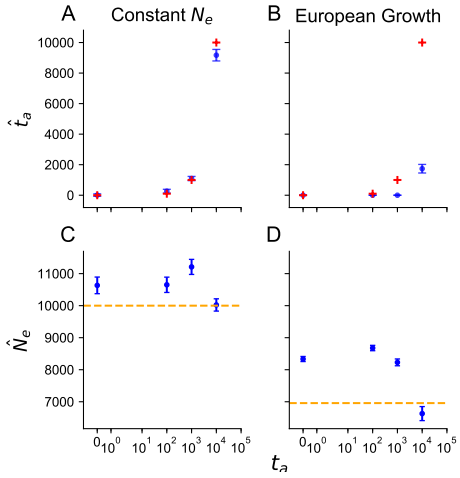

Figure 3.9: Estimation of sample age t_a (top row) and N_e (bottom row) in two demographic histories (columns). Estimation was performed by minimization of least-squares between the empirical correlation in pairwise differences (in 1 kb windows) and the theoretical correlation using `scipy.optimize`. Windows of recombination distance are binned using the `numpy.histogram` function with automatic binning for recombination distance in the range $r \in (10^{-5}, 5 \times 10^{-3})$ Morgans. Standard errors are estimated via a bootstrap (leave one chromosome out). Simulations were run under a constant population size of 10^4 and a demographic history from (Tennessen et al., 2012). We simulated 20 replicate chromosomes 20 megabases in length, with mutation rates $\mu = 10^{-8}$ and a recombination rate of 10^{-8} as well.

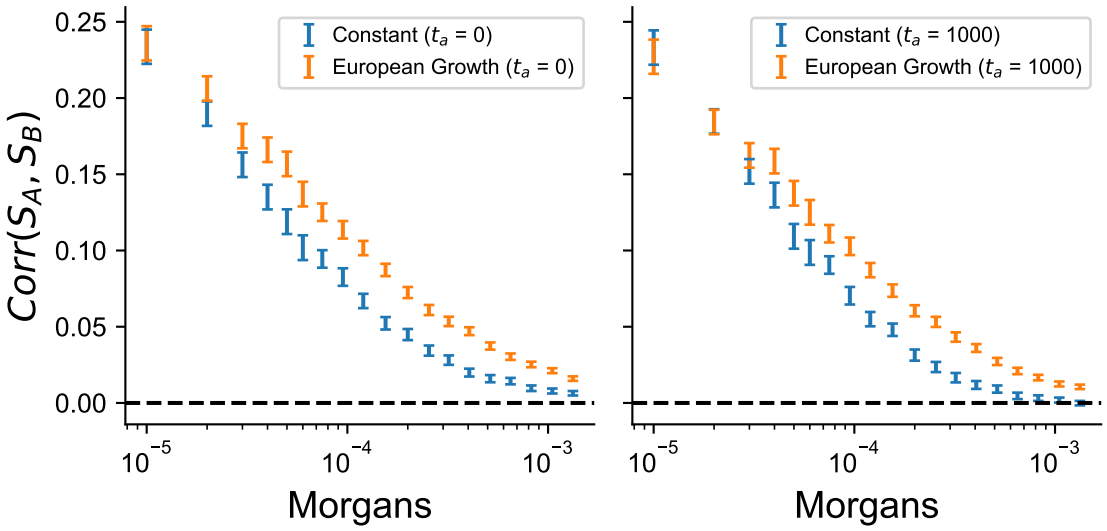

Figure 3.10: The "European Growth" model is the model of Tennessen et al. (2012). The locus size considered here is one kilobase, and the mutation rate is set to be $\mu = 10^{-8}$. Windows of recombination distance are binned using the `numpy.histogram` function with automatic binning for recombination distance in the range $r \in (10^{-5}, 5 \times 10^{-3})$ Morgans. In both cases for temporal sampling, the more realistic demographic model increases the length scale over which the correlation extends.

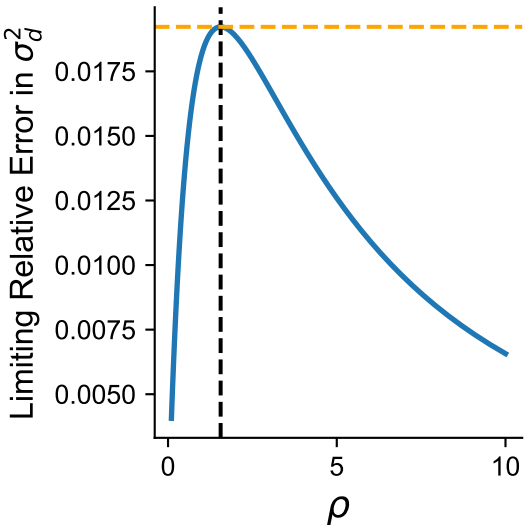

Figure 3.11: Limiting results on the maximum relative error in when the sampling time of the ancient haplotype is $t_a \to \infty$ relative to $t_a = 0$. The orange dashed line corresponds to the theoretical maximum relative error in this case (see Appendix 2 for derivation).

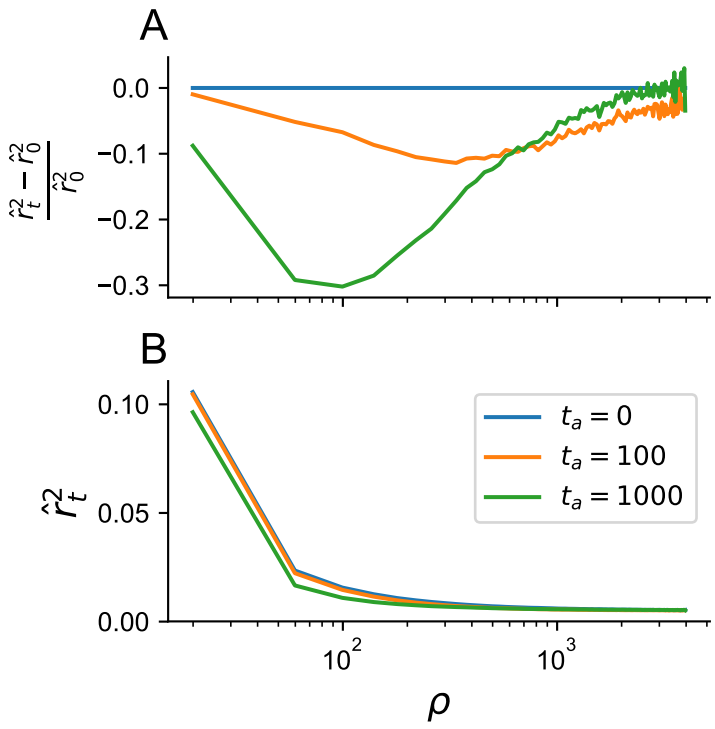

Figure 3.12: Empirical calculation of $\hat{r}^2$ in simulated haplotype data in 1 Mb chunks with 100 ancient haplotypes and 100 modern haplotypes across 500 replicate simulations. Means across the 500 simulations are plotted for ease of visualization.

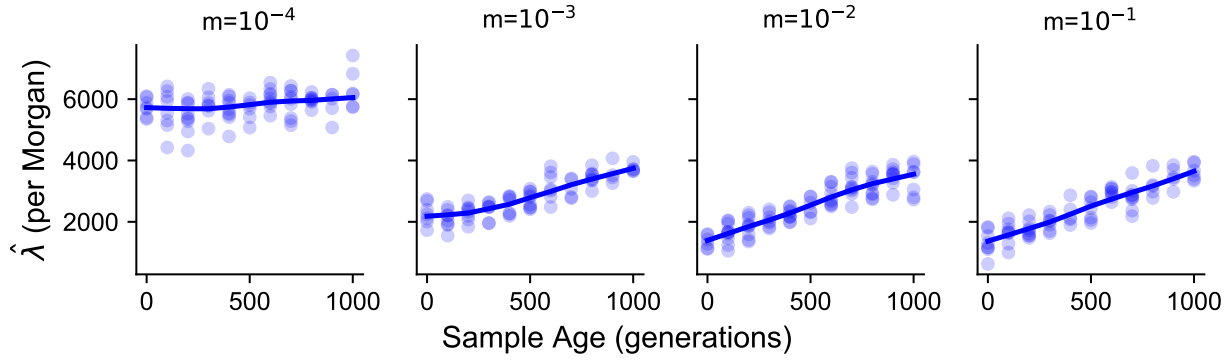

Figure 3.13: Estimation of copying rates in a model of two demes with the modern panel ($n = 100$) in one population, and the ancient individual is in a second deme with a constant migration rate of m per generation between the two. As migration rates grow lower, we expect that the T_{mrca} will also increase between the ancient sample and a member of the panel and we similarly observe an increase in the haplotype copying model jump rate.

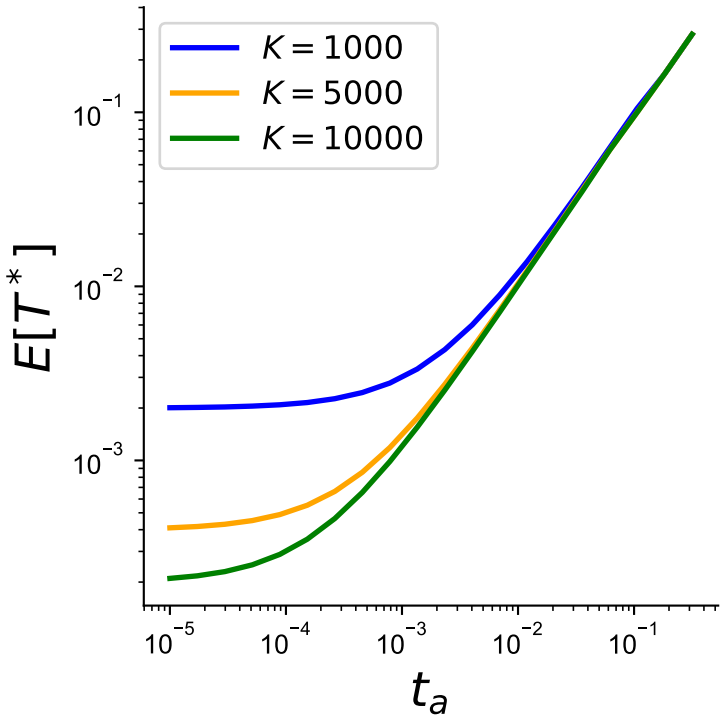

Figure 3.14: Expected time to first coalescent event involving an ancient haplotype with lineages ancestral to modern panel. This is the expected external branch length as a function of the sample age t_a in coalescent units in a model of constant population size. Note that past a certain time-sampling, we expect all of the external branch lengths to be quite similar since the expected number of lineages ancestral to the panel will be approximately equal. Here we use Griffith's approximation (Griffiths, 1984) to the expected to the number of lineages left at time t_a. As an approximation, with an $N_e = 10^4$, a sample with $t_a = 10^{-3}$ is 10 generations old($\sim$ 280 years old (Fenner, 2005)), and there is little difference between $K = 5000$ and $K = 10000$.

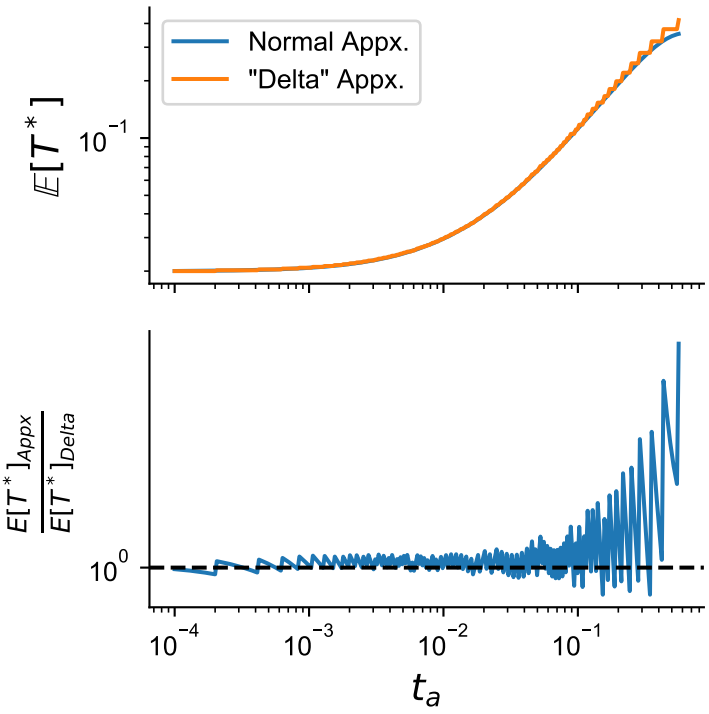

Figure 3.15: Measuring the relative error between using the expectation of the number of lineages (Jewett and Rosenberg, 2014) as opposed to using the asymptotic normal distribution from (Chen and Chen, 2013) when calculating the expected time to first coalescence for an ancient haplotype. The haplotype panel size used here is $K = 100$ haplotypes. We observe that for relatively short time-scales we have little relative error in our approximation (< 0.05), but that at larger time-scales on the order of $\sim N_e/10$ we begin to observe a sizeable discrepancy between the approximations.

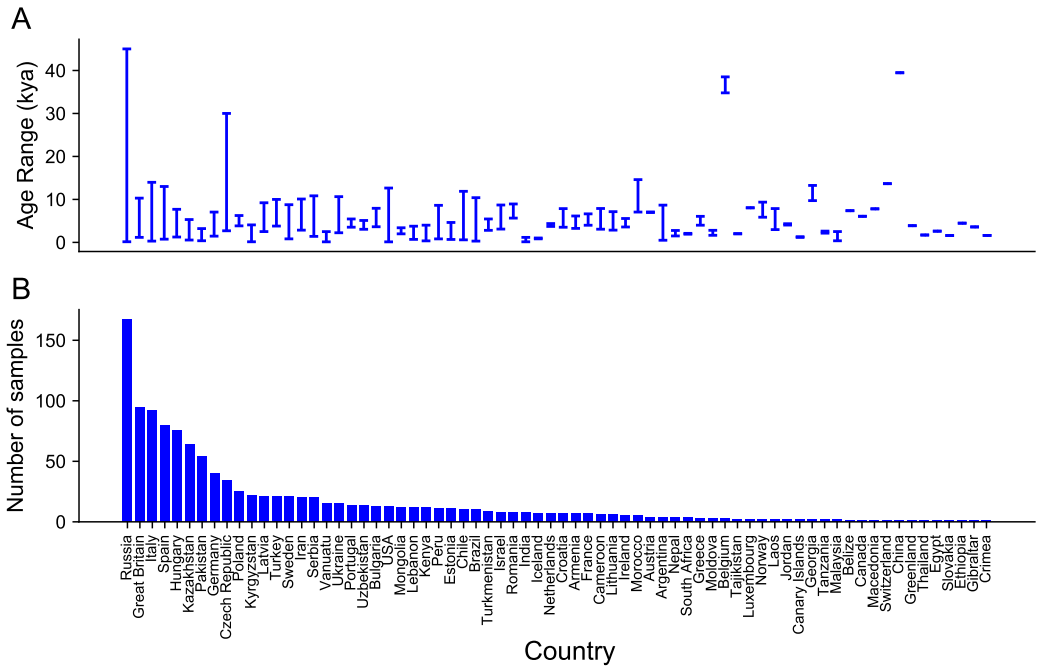

Figure 3.16: **(A)** Age range of samples from a single country in thousands of years (kya) **(B)** Samples per country that pass quality control filters (see 3.5).

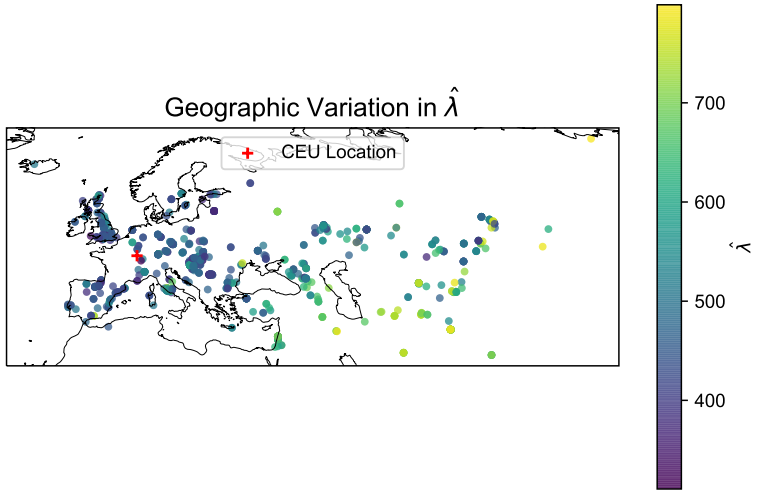

Figure 3.17: Estimation of $\hat{\lambda}$ under the haplotype-copying model using 49 CEU male X chromosomes as a modern haplotype panel and its magnitude across geographic space. Qualitatively, we note that there is an increase in the haplotype copying jump rate with increasing distance from the hypothetical location for the CEU individuals.

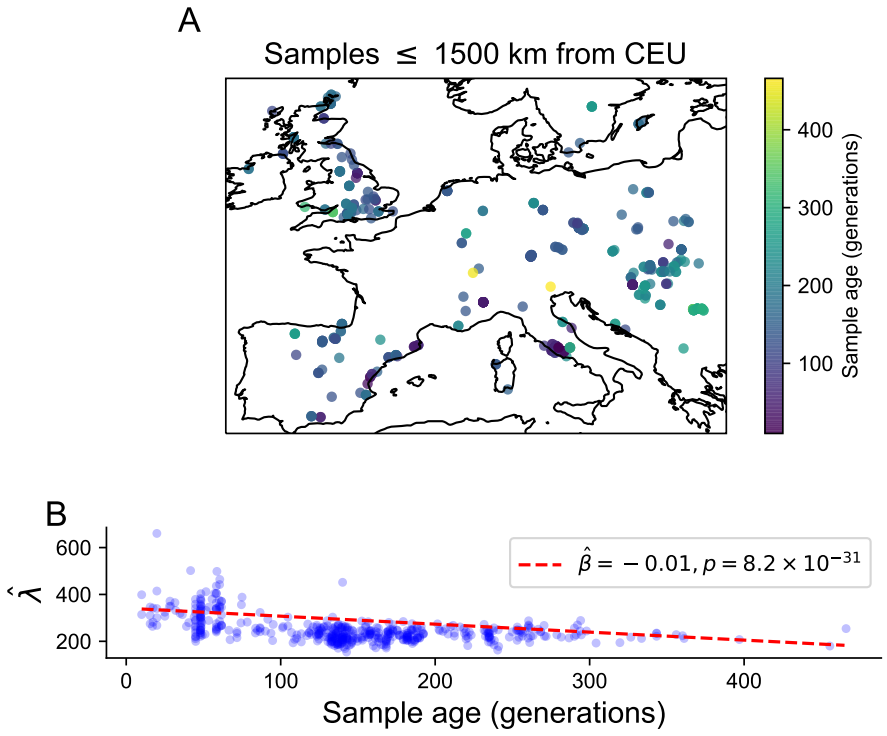

Figure 3.18: **(A)** Map of samples $<$ 1500 kilometers from hypothetical location of central Europe **(B)** Decrease in estimated $\hat{\lambda}$ as a function of sample age in generations when estimated from male X chromosomes using all male X chromosomes from samples in the `EUR` regional grouping from (Auton et al., 2015). Multiple linear regression line was fit with additional terms for latitude, longitude, and the interaction term of latitude $\times$ longitude as covariates.

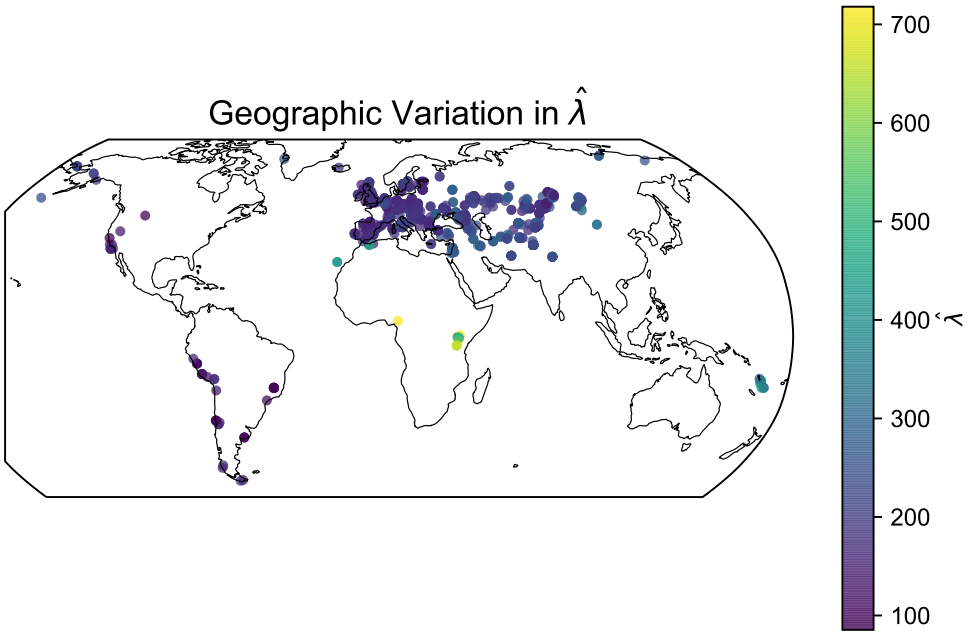

Figure 3.19: Estimation of haplotype copying jump-rate (per Morgan) on the X-chromosome for all samples with > 1 variants per 25 kb.

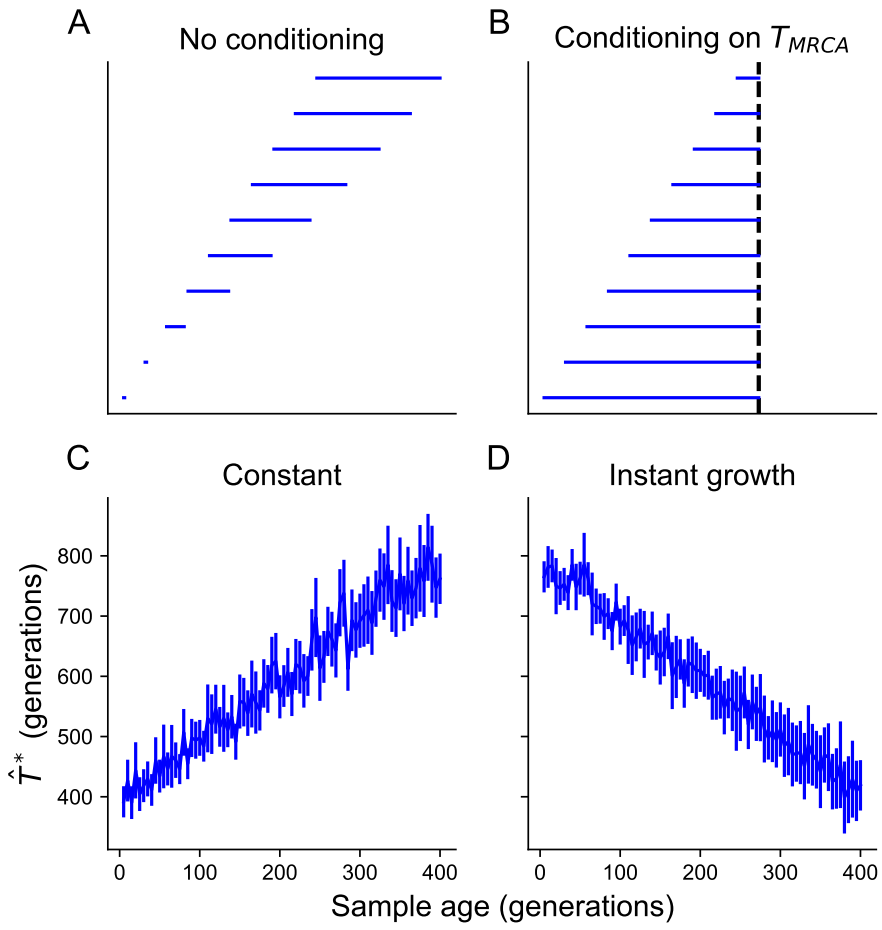

Figure 3.20: **(A),(B)** Visual depiction of impact of conditioning on T_{mrca} and its impact on coalescent intervals, implicating less overall branch length on which a recombination can initiate a haplotype switching event when conditioning on the T_{mrca}. **(C)** Time till first coalescence with a lineage ancestral to a member of a modern panel ($K = 100$) in a constant population size scenario of $N_e = 10^4$ and **(D)** in a simulation with instantaneous growth at 400 generations from a population size of 10^4 to 10^6. In both simulations, ancient haplotypes are sampled every 5 generations, are conducted using 5000 replicates and error bars represent 2 standard deviations from the mean time to first coalescence.

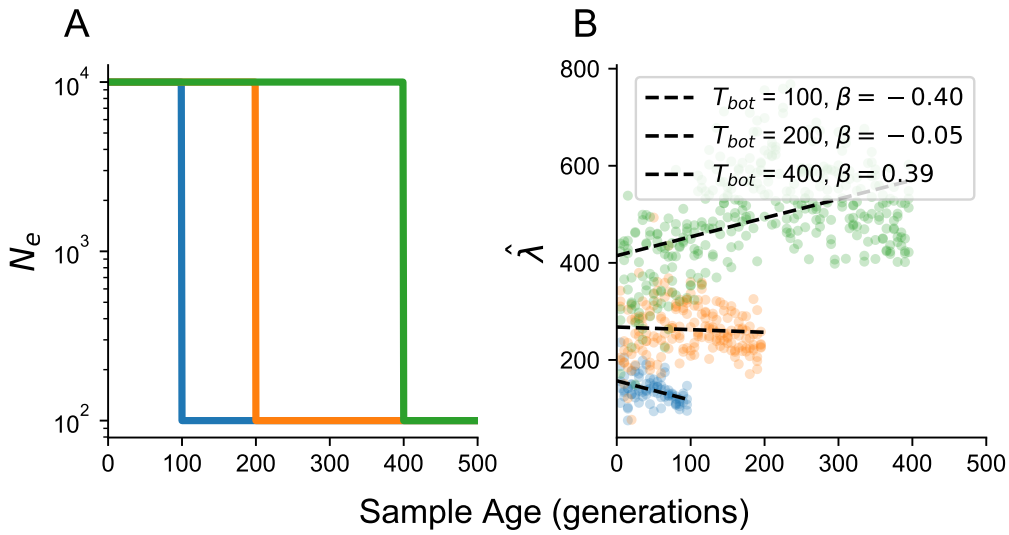

Figure 3.21: **(A)** Model of instant population growth occurring at 100 (blue) 200 (orange) and 400 (green) generations in the past. **(B)** Estimates of linear regression for estimated jump-rate $\hat{\lambda}$ as a function of time (only using time-points prior to the bottleneck). Note that in this case we observe both postive and negative slopes regarding the relationship between $\hat{\lambda}$ and sample age.

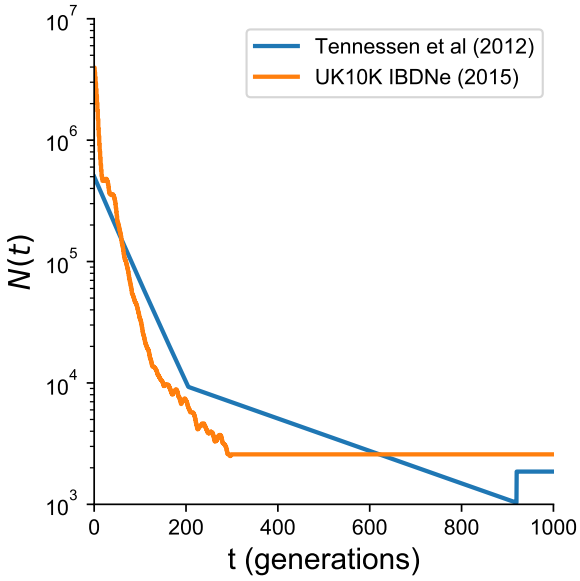

Figure 3.22: Demographic history estimated by Tennessen et al. (2012) and Browning and Browning (2015). Note that we have simply truncated the history from Browning and Browning (2015) at 300 generations to extend into the past with a constant population size.

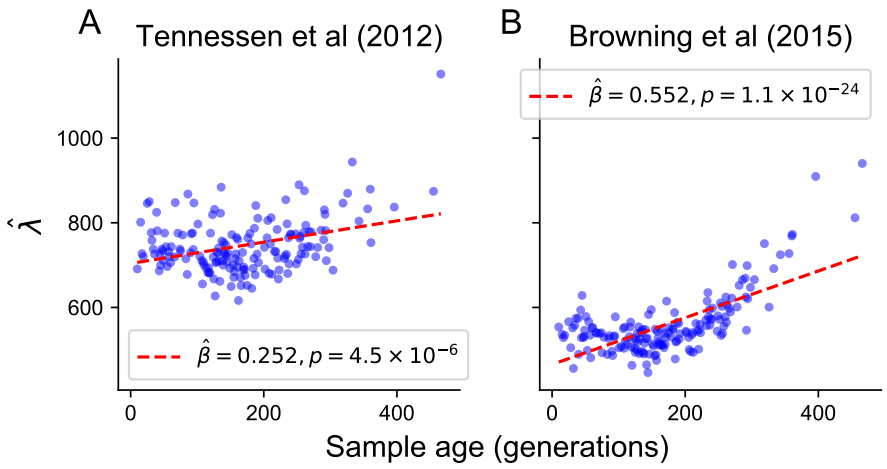

Figure 3.23: Estimation of maximum-likelihood haplotype copying jump-rate as a function of sampling time under **(A)** the Tennessen et al. (2012) and **(B)** Browning and Browning (2015) models of demographic history shown in Figure 3.22. In both simulations data were simulated using the full length of the X chromosome and recombination according to the map of Kong et al. (2010). Timepoints simulated are the same as all samples < 1500 kilometers from the assumed location of CEU (Figure 3.6).

3.8 Supplementary Tables

Sample ID	Population	p-value (Binomial Test)
NA06984	CEU	0.0161
NA06985	CEU	0.5847
NA06986	CEU	0.0987
NA18525	CHB	0.8555
NA18526	CHB	0.5847
NA18528	CHB	0.3616
NA20845	GIH	0.2005
NA20846	GIH	0.2005
NA20847	GIH	0.5847
NA18486	YRI	0.3616
NA18488	YRI	0.0428
NA18489	YRI	0.2005

Table 3.1: The impact of different modern references from the 1000 Genomes Project (Auton et al., 2015) on the difference in the correlation in pairwise differences between *LBK* (Lazaridis et al., 2014) and *Ust-Ishim* (Fu et al., 2014). From our theory, we expect there to be a difference but we find little signal for this in the data (Figure 3.3)

3.9 Appendices

3.9.1 Appendix 1: Mathematical details of the two-locus model

Appendix 1a : The two-locus model with population continuity and serial sampling

We first begin with a model of constant population size and where we sample one haplotype from the present and one haplotype at time t_a ago (in coalescent units). The population is assumed to be constant in size with population scaled recombination rate $\rho = 4N_e r$. Since we have two-samples from different time-points, we have two phases of the process : (1) where only the modern lineage can evolve at two loci (t, t_a) and when both haplotypes are available to coalesce and recombine with one another $(t \geq t_a)$.

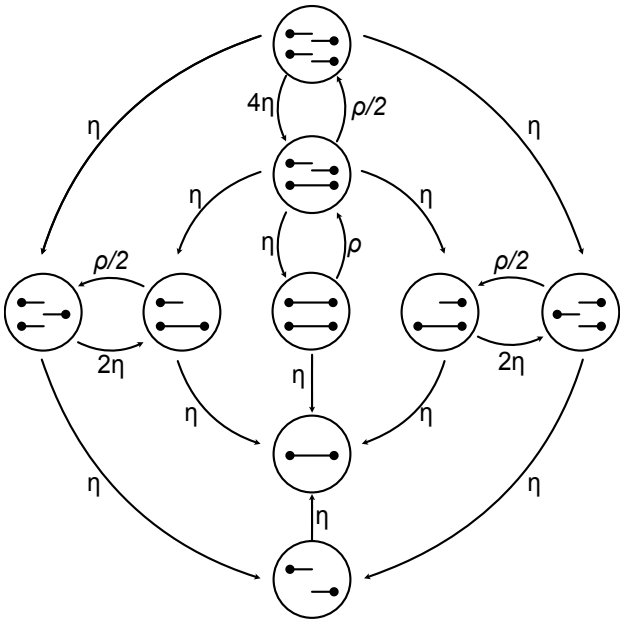

Figure 3.24: Markov chain model for the ancestral process at two loci from Simonsen and Churchill(Simonsen and Churchill, 1997). In all settings for two modern haplotypes we assume that we start from the state in the middle (state "0") in all applications, which means that all sampled haplotypes are joint. The parameter η represents the coalescent rate and the parameter ρ represents the recombination rate (all in the coalescent scale). Figure adapted from (Hobolth and Jensen, 2014).

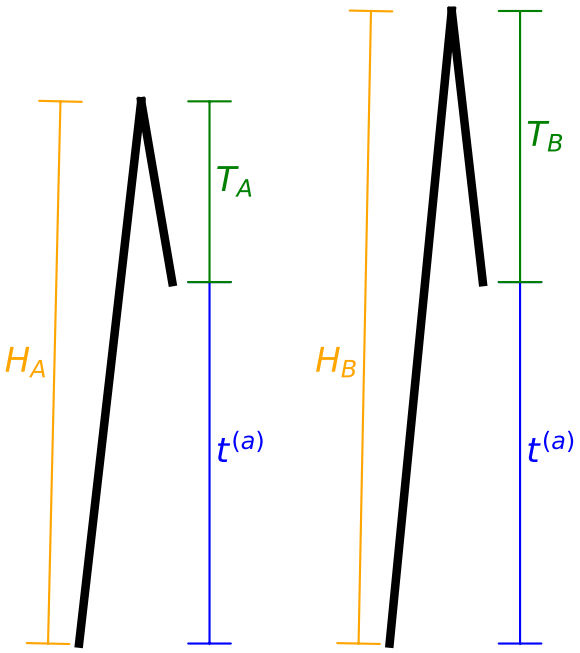

Figure 3.25: Description of variables in the two-locus case. H is the total tree height, T is the coalescent time of the ancient and modern lineage, and t_a is the sampling time of the ancient lineage (in coalescent units). Here subscripts A, B are used to denote the two loci separated by scaled recombination distance ρ.

First we start with the single-locus expectations and variances for tree height and total branch length, omitting the subscripts for the loci when considering marginal quantities:

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{V}ar[T] = 1 \\
H &= T + t_a \\
\mathbb{E}[H] &= \mathbb{E}[T + t_a] \\
&= 1 + t_a \\
\mathbb{V}ar[H] &= \mathbb{V}ar[T + t_a] \qquad = 1 \\
\mathbb{E}[L] &= \mathbb{E}[2H - t_a] \\
&= 2(1 + t_a) - t_a \\
&= 2 + t_a \\
\mathbb{V}ar[L] &= \mathbb{V}ar[2H - t_a] \\
&= 4\mathbb{V}ar[H] = 4
\end{aligned}$$

In order to calculate the covariance $\mathbb{C}ov(L_A, L_B)$, we require joint statistics for the trees:

$$\begin{aligned}
\mathbb{E}[L_A L_B] &= \mathbb{E}[(2H_A - t_a)(2H_B - t_a)] \\
&= \mathbb{E}\left[4H_A H_B - 2t_a H_A - 2t_a H_B + t_a^2\right] \\
&= 4\mathbb{E}[H_A H_B] - 4t_a \mathbb{E}[H_A] + t_a^2 \\
&= 4\mathbb{E}[H_A H_B] - 4t_a(1 + t_a) + t_a^2 \\
\mathbb{E}[H_A H_B] &= \mathbb{E}[(T_A + t_a)(T_B + t_a)] \\
&= \mathbb{E}[T_A T_B] + 2t_a \mathbb{E}[T_A] + t_a^2 \\
&= \mathbb{E}[T_A T_B] + 2t_a + t_a^2
\end{aligned}$$

Therefore, we can compute $\mathbb{C}ov(L_A, L_B)$ by computing $\mathbb{E}[T_A T_B]$. We solve this using an "staggered" version of the Simonsen-Churchill Model (Simonsen and Churchill, 1997;

Hobolth and Jensen, 2014). In the phase where $t < t_a$, we have to consider this as a two state continuous time Markov process with the rate matrix:

$$
\begin{aligned}
Q &= \begin{bmatrix} -\frac{\rho}{2} & \frac{\rho}{2} \\ 1 & -1 \end{bmatrix} \\
\mathbb{P}_{t_a}(x = (1,1,1)) &= \left(e^{\mathbf{Q}t_a}\right)_{0,1} \\
&= \frac{\rho(1 - e^{-t(\frac{\rho}{2}+1)})}{\rho + 2} \\
\mathbb{P}_{t_a}(x = (2,0,0)) &= 1 - \mathbb{P}_{t_a}(x = (1,1,1))
\end{aligned}
$$

The state $x = (1,1,1)$ represents that there is 1 lineage ancestral to both locus A and B, only one lineage ancestral to locus A, and only one lineage ancestral to locus B. This also corresponds to our "uncoupled state" in the main text. The two states in the Markov process with a single haplotype represent either "coupled" (e.g. $(2,0,0)$ in the notation of (Hobolth and Jensen, 2014; Simonsen and Churchill, 1997)) or "uncoupled". We have overloaded the notation for the triplets leading to particular states above to remain consistent with the case of two haplotypes, which is the case when the ancient haplotype enters the process. Returning to our motivation to compute $\mathbb{E}[T_A T_B]$, or the joint expectation of the time to coalescence from when both haplotypes are in the process.

$$
\begin{aligned}
\mathbb{E}_{(2,0,0)}[T_A T_B] &= \frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18} \\
\mathbb{E}_{(1,1,1)}[T_A T_B] &= \frac{\rho^2 + 13\rho + 24}{\rho^2 + 13\rho + 18} \\
\mathbb{E}[T_A T_B] &= \mathbb{P}_{t_a}(x = (2,0,0))\mathbb{E}_{(2,0,0)}[T_A T_B] + \mathbb{P}_{t_a}(x = (1,1,1))\mathbb{E}_{(1,1,1)}[T_A T_B] \\
&= \left(1 - \frac{\rho(1 - e^{-t(\frac{\rho}{2}+1)})}{\rho + 2}\right) \frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18} + \frac{\rho(1 - e^{-t(\frac{\rho}{2}+1)})}{\rho + 2} \frac{\rho^2 + 13\rho + 24}{\rho^2 + 13\rho + 18}
\end{aligned} \tag{3.4}
$$

The first two expressions can be found directly in (Richard Durrett, 2002). The last

expression represents a weighting of the expectations from different starting states in the two-locus ancestral process, where the weight corresponds to the probabilities that the modern haplotype is uncoupled at the time the ancient haplotype is sampled, t_a. From this we can compute the covariance and correlation in the total branch length.

$$\begin{aligned}
\mathbb{C}ov(L_A, L_B) &= \mathbb{E}[L_A L_B] - \mathbb{E}[L_A]\mathbb{E}[L_B] \\
&= \left(4\mathbb{E}[H_A H_B] - 4t_a(1+t_a) + t_a^2\right) - (2+t_a)^2 \\
&= 4\mathbb{E}[H_A H_B] - 4t_a(1+t_a) + t_a^2 - (4 + 4t_a + t_a^2) \\
&= 4\left(\mathbb{E}[T_A T_B] + t_a + t_a^2 - t_a(1+t_a) - 1\right) \\
&= 4\left(\mathbb{E}[T_A T_B] - 1\right) \\
\mathbb{C}orr(L_A, L_B) &= \frac{\mathbb{C}ov(L_A, L_B)}{\sqrt{Var(L_A)Var(L_B)}} \\
&= \mathbb{E}[T_A T_B] - 1
\end{aligned}$$

If we take the limits of t as 0 and ∞, we can exhibit the asymptotic behavior of $Corr(L_A, L_B)$ in terms of ρ.

$$\begin{aligned}
\frac{2}{\rho+2} < \mathbb{P}(x = (2,0,0)|t) < 1 \\
Corr(L_A, L_B) &= \mathbb{E}[T_A T_B] - 1 \\
Corr(L_A, L_B)|t = 0 &= \frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18} - 1 \\
&= \frac{\rho + 18}{\rho^2 + 13\rho + 18} \\
Corr(L_A, L_B)|t \to \infty &= \frac{2}{\rho+2}\frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18} + \frac{\rho}{\rho+2}\frac{\rho^2 + 13\rho + 24}{\rho^2 + 13\rho + 18} - 1 \\
&= \frac{8\rho + 36}{\rho^3 + 15\rho^2 + 44\rho + 36}
\end{aligned}$$

This derivation serves to highlight the rate of decay in the branch length correlation as a function of the sampling time from $\mathcal{O}(\rho^{-1})$ to $\mathcal{O}(\rho^{-2})$ that we also observe in our main

results.

To relate the correlation in total branch length to potential correlations in the number of segregating sites, we can utilize the following identities in the case of the population-scaled mutation rate (θ) being the same at locus A and locus B (Hobolth et al., 2019):

$$
\begin{aligned}
S_A|L_A &\sim \text{Pois}\left(\frac{\theta}{2}L_A\right) \\
S_B|L_B &\sim \text{Pois}\left(\frac{\theta}{2}L_B\right) \\
\mathbb{E}[S_A] &= \mathbb{E}[S_B] = \mathbb{E}[\mathbb{E}[S_A|L_A]] = \frac{\theta}{2}\mathbb{E}[L_A] \\
Var(S_A) &= \mathbb{E}[Var(S_A|L_A)] + Var(\mathbb{E}[S_A|L_A]) \\
&= \frac{\theta}{2}\mathbb{E}[L_A] + \left(\frac{\theta}{2}\right)^2 Var(L_A) \\
\mathbb{E}[S_A S_B] &= \mathbb{E}[\mathbb{E}[S_A S_B|L_A L_B]] = \frac{\theta^2}{4}\mathbb{E}[L_A L_B] \\
\mathbb{C}ov(S_A, S_B) &= \mathbb{E}[S_A S_B] - \mathbb{E}[S_A]\mathbb{E}[S_B] = \frac{\theta^2}{4}Cov(L_A, L_B) \\
\mathbb{C}orr(S_A, S_B) &= \frac{Cov(S_A, S_B)}{\sqrt{Var(S_A)Var(S_B)}} \\
&= \frac{\frac{\theta^2}{4}Cov(L_A, L_B)}{\sqrt{\left(\frac{\theta}{2}\mathbb{E}[L_A] + \frac{\theta^2}{4}Var(L_A)\right)^2}} \\
&= \frac{\mathbb{C}ov(L_A, L_B)}{\frac{2}{\theta}\mathbb{E}[L_A] + Var(L_A)} \\
&= \frac{4(\mathbb{E}[T_A T_B] - 1)}{\frac{2}{\theta}(2 + t_a) + 4} \\
&= \frac{1}{1 + \frac{2+t_a}{2\theta}}(\mathbb{E}[T_A T_B] - 1) \\
\mathbb{C}orr(S_A, S_B) &= \frac{1}{1 + \frac{2+t_a}{2\theta}}\mathbb{C}orr(L_A, L_B)
\end{aligned}
\tag{3.5}
$$

For the case where we have $n = 2$ haplotypes, this corresponds to the correlation in pairwise diversity between two loci or $Corr(\pi_A, \pi_B)$. We note that this expression holds in all of the cases explored further in this appendix and do not derive it specifically for each

specific case (not shown)

Appendix 1b : Two-Locus Model with Population Divergence

In this section we assume that there has been a divergence between the populations containing the ancient lineage as well as the modern lineage at the coalescent scale (t_{div}). Following arguments from appendix 1c, we can break the ancestral process into three phases: (1) when the modern lineage is the only one evolving backwards (2) when the ancient lineage and the modern lineage are both evolving *but are not able* to coalescent with one another and (3) when both lineages are in the ancestral population and can coalesce with one another.

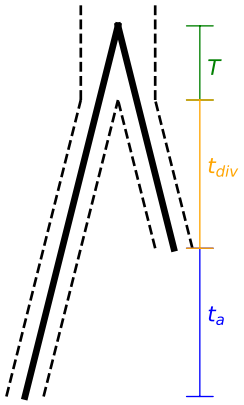

Figure 3.26: H is the total tree height, T is the coalescent time of the ancient and modern lineage, t_{div} is the divergence time, and t_a is the sampling time of the ancient lineage (in coalescent units).

The model with population divergence has an additional parameter in the model, t_{div}, the divergence time of the two populations. If we first show the marginal tree properties under the divergence model:

$$
\begin{aligned}
\mathbb{E}[T] &= Var[T] = 1 \\
\mathbb{E}[H] &= \mathbb{E}[T + t_a + t_{div}] \\
&= 1 + t_a + t_{div} \\
Var[H] &= Var[T + t_a + t_{div}] \\
&= 1 \\
\mathbb{E}[L] &= \mathbb{E}[2H - t_a] \\
&= 2\mathbb{E}[H] - t_a \\
&= 2(1 + t_a + t_{div}) - t_a \\
&= 2 + t_a + 2t_{div} \\
Var[L] &= Var[2H - t_a] \\
&= 4Var[H] = 4
\end{aligned}
$$

We can now calculate certain moments of the joint distribution of genealogical properties such as tree height and tree length under this model with divergence, where the two-locus process behaves independently for time t_a and t_{div}. We will start with derivations for the joint tree-height and tree length:

$$
\begin{aligned}
\mathbb{E}[H_A H_B] &= \mathbb{E}[(T_A + t_a + t_{div})(T_B + t_a + t_{div})] \\
&= \mathbb{E}[T_A T_B] + 2t_{div} + t_{div}^2 + 2t_a + 2t_{div}t_a + t_a^2 \\
&= \mathbb{E}[T_A T_B] + 2t_{div} + 2t_a + (t_a + t_{div})^2 \\
\mathbb{E}[L_A L_B] &= \mathbb{E}[(2H_A - t_a)(2H_B - t_a)] \\
&= 4\mathbb{E}[H_A H_B] - 4t_a + t_a^2
\end{aligned}
$$

Again it is apparent that we have to solve for the joint expectation of $\mathbb{E}[T_A T_B]$, but in this model we need to consider that we have two independent processes in each population. We begin the derivation below:

$$\begin{aligned}
\mathbb{P}(x=(0,0,2)|t_a,t_{div}) &= \mathbb{P}(x_1=(0,0,1)|t_a+t_{div})\mathbb{P}(x_2=(0,0,1)|t_{div}) \\
&= \frac{\rho e^{-(t_a+t_{div})(\rho+1)}+1}{\rho+1}\frac{\rho e^{-t_{div}(\rho+1)}+1}{\rho+1} \\
\mathbb{P}(x=(2,2,0)|t_a,t_{div}) &= \mathbb{P}(x_1=(1,1,0)|t_a+t_{div})\mathbb{P}(x_2=(1,1,0)|t_{div}) \\
&= \frac{\rho(1-e^{-(t_a+t_{div})(\rho+1)})}{\rho+1}\frac{\rho(1-e^{-t_{div}(\rho+1)})}{\rho+1} \\
\mathbb{P}(x=(1,1,1)|t_a,t_{div}) &= \mathbb{P}(x_1=(0,0,1)|t_a+t_{div})\mathbb{P}(x_2=(1,1,0)|t_{div}) \\
&\quad + \mathbb{P}(x_1=(1,1,0)|t_a+t_{div})\mathbb{P}(x_2=(0,0,1)|t_{div}) \\
&= \left(\frac{\rho e^{-(t_a+t_{div})(\rho+1)}+1}{\rho+1}\frac{\rho(1-e^{-t_{div}(\rho+1)})}{\rho+1}\right) \\
&\quad + \left(\frac{\rho(1-e^{-(t_a+t_{div})(\rho+1)})}{\rho+1}\frac{\rho e^{-t_{div}(\rho+1)}+1}{\rho+1}\right)
\end{aligned}$$

From these probabilities we can calculate the expectation of the joint coalescent times conditional on being in a particular state at time $t_a + t_{div}$.

$$\mathbb{E}[T_A T_B] = \sum_{x\in\{(1,1,1),(0,0,2),(2,2,0)\}} \mathbb{P}(x=x|t_a,t_{div})\mathbb{E}_x[T_A T_B]$$

Where each of $\mathbb{E}_x[T_A T_B]$ are defined using previously derived results under the Simonsen-Churchill model for a constant-sized population (Simonsen and Churchill, 1997), We can see that this is different from the model before (where the $x = (2, 2, 0)$ state was not possible). Although if we set $t_{div} = 0$, then this corresponds directly to the model without divergence as above. While the underlying mathematical results are slightly more involved - they do provide insights on how divergence can affect joint coalescent times that are in line with qualitative intuitions.

We can now compute joint statistics (e.g. covariance and correlation) of the tree height and total tree length at each of the loci as well following common formulas such as that for the correlation in total branch length:

$$Corr(L_A, L_B) = \mathbb{E}[T_A T_B] - 1$$

Estimation of N_e for Tennessen et al. Growth Model

In order to estimate an effective N_e to compare our constant population size theory for two-loci with simulations under varying demographic history, we took a Monte-Carlo estimation approach based on comparing the theoretical amount of time for pair-wise coalescence to known expectations under constant population size:

$$\begin{aligned} T_j &\sim \text{Exponential}\left(\frac{\binom{j}{2}}{N_e}\right) \\ \mathbb{E}[T_2] &= \frac{N_e}{2(2-1)} = \frac{N_e}{2} \\ \hat{N}_e &= 2\hat{T}_2 \end{aligned}$$

To generate a Monte-Carlo estimate of $\hat{T}_2$ under the model of (Tennessen et al., 2012), we used `msprime` (Kelleher et al., 2016) to simulate 10^4 replicates of genealogical trees for two modern samples and calculated $\hat{T}_2$ as the empirical mean of the pairwise T_{mrca} and computed $\hat{N}_e$ according to the above formula. We find through our simulations $\hat{N}_e = 6958$ under the model of (Tennessen et al., 2012) and use that when comparing our values for the correlation in segregating sites in the main text.

Parameter Estimation for t_a and N_e from the correlation in segregating sites

Since we have analytic formulas for the correlation in segregating sites in a constant-sized population, here we investigate whether we can perform parameter inference on the sample age in generations t_a, and the effective population size N_e from data. We have:

$$
\begin{aligned}
t_a &= \frac{\tau_a}{2N_e} \\
\rho &= 4N_e r \\
\theta &= 4N_e \mu \\
Corr(\pi_A, \pi_B) &= \frac{1}{1 + \frac{2+t_a}{2\theta}} Corr(L_A, L_B) \\
\gamma &= \frac{\rho(1 - e^{-t_a(\frac{\rho}{2}+1)})}{\rho + 2} \\
&= \frac{1}{1 + \frac{2+t_a}{2\theta}} \left[(1-\gamma) \frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18} + \gamma \frac{\rho^2 + 13\rho + 24}{\rho^2 + 13\rho + 18} \right]
\end{aligned}
$$

Assuming we know both r and μ, using previous estimates of the mutation rate and the recombination map, we can subsequently estimate N_e and t_a. These are independent sources of information on the value of N_e, outside of the compound parameter $t = \frac{t_a}{2N_e}$. This is because there are terms in the derivation that only involve one of the two parameters (ρ contains information on N_e that is not coupled with τ_a).

We attempted to estimate $\hat{\tau}_a, \hat{N}_e$ by minimizing the loss function:

$$
(\hat{t}_a, \hat{N}_e) = \underset{t_a, N_e}{\arg\min} \sum_{r \in \mathbf{R}} \left(\widehat{Corr}(\pi_A, \pi_B | r) - Corr(\pi_A, \pi_B | r) \right)^2 \tag{3.6}
$$

Where $\widehat{Corr}$ is the empirical correlation measured in the data. Here $\mathbf{R}$ is a set of recombination distances at which the squared distance in evaluated (measured in Morgans, or per-generation recombination probability). The empirical correlation is measured in a recombination window by Monte-Carlo sampling of pairs of windows of a certain size (we used 1 kb for our analysis) at a particular recombination distance.

3.9.2 Appendix 2: Theoretical expectations of r^2 with temporally structured samples

If we recall the genealogical approximation to r^2 (McVean, 2002):

$$\begin{aligned}\sigma_d^2 &= \frac{\mathbb{E}[D^2]}{\mathbb{E}[p(1-p)q(1-q)]} \\ &= \frac{\mathbb{E}[T_A T_B|x=(2,0,0)] - 2\mathbb{E}[T_A T_B|x=(1,1,1)] + \mathbb{E}[T_A T_B|x=(0,2,2)]}{\mathbb{E}[T_A T_B|x=(0,2,2)]}\end{aligned} \tag{3.7}$$

Where $\mathbb{E}[T_A T_B|x]$ is the joint expectation of the time to coalescence at loci A, B given a particular starting state in the two-locus Markov Chain (Simonsen and Churchill, 1997) (Figure 3.24). However here, we adopt the notation similar to (McVean, 2002) where $T_A T_B|x = (2,0,0)$ is denoted as $t_{x(ij)}t_{y(ij)}$, signifying that the ancestral lineages are the same at both loci. We will use this notation throughout this derivation. Given the marginal pairwise coalescent times $\mathbb{E}[t_{x(..)}] = 1$, we can write the approximation σ_d^2 using the covariance in coalescent times (which is the more commonly seen result):

$$\sigma_d^2 = \frac{\mathbb{C}ov(t_{x(ij)}, t_{y(ij)}) - 2\mathbb{C}ov(t_{ij}, t_{ik}) + \mathbb{C}ov(t_{ij,t_{kl}})}{\mathbb{C}ov(t_{ij}, t_{kl}) + \mathbb{E}[t^2]} \tag{3.8}$$

However this can also be written in terms of the joint expectations of coalescent times with a particular starting scenario:

$$\sigma_d^2 = \frac{\mathbb{E}[t_{x(ij)}t_{x(ij)}] - 2\mathbb{E}[t_{x(ij)}t_{y(ik)}] + \mathbb{E}[t_{x(ij)}t_{y(kl)}]}{\mathbb{E}[t_{ij}t_{kl}]} \tag{3.9}$$

Under the Simonsen-Churchill model(Simonsen and Churchill, 1997), these expectations can be analytically derived (Richard Durrett, 2002) for a population-scaled recombination rate of $\rho = 4N_e r$ with both samples at the present:

$$\mathbb{E}_0[t_{x(ij)}t_{y(ij)}] = \frac{\rho^2 + 14\rho + 36}{\rho^2 + 13\rho + 18}$$
$$\mathbb{E}_0[t_{x(ij)}t_{y(ik)}] = \frac{\rho^2 + 13\rho + 24}{\rho^2 + 13\rho + 18}$$
$$\mathbb{E}_0[t_{x(ij)}t_{y(kl)}] = \frac{\rho^2 + 13\rho + 22}{\rho^2 + 13\rho + 18}$$

Here we ask about the form of each expectation when haplotypes are sampled from different time-points, and its effect on our expectation of linkage disequilibrium.

Bringing these two scenarios together with some changing of the variables in the case of one ancient and one modern haplotype we have:

$$\gamma = \frac{\rho(1 - e^{-t_a(\frac{\rho}{2}+1)})}{\rho + 2}$$
$$\mathbb{E}_{t_a}[t_{x(ij)}t_{y(ik)}] = (1-\gamma)\mathbb{E}_0[t_{x(ij)}t_{y(ik)}] + \frac{\gamma}{2}\mathbb{E}_0[t_{x(ij)}t_{y(ij)}] + \frac{\gamma}{2}\mathbb{E}_0[t_{x(ij)}t_{y(kl)}]$$

The term γ is the probability of a coupled haplotype at time 0 being in the uncoupled state at time t_a. It is the exact same term used in our calculation of the correlation in total branch length in Appendix 1.

Re-defining our previous results with the notation:

$$\mathbb{E}[t_{x(ij)}t_{y(ij)}] = (1-\gamma)\mathbb{E}_0[t_{x(ij)}t_{y(ij)}] + \gamma\mathbb{E}_0[t_{x(ij)}t_{y(ik)}]$$
$$\mathbb{E}[t_{x(ij)}t_{y(kl)}] = (1-\gamma)\mathbb{E}_0[t_{x(ij)}t_{y(kl)}] + \gamma\mathbb{E}_0[t_{x(ij)}t_{y(ik)}]$$

From the above derivations we can calculate the numerator of σ_d^2, $\mathbb{E}[D^2]$:

$$\begin{aligned}\mathbb{E}[D^2] &= \mathbb{E}_{t_a}[t_{x(ij)}t_{y(ij)}] - 2\mathbb{E}_{t_a}[t_{x(ij)}t_{y(ik)}] + \mathbb{E}_{t_a}[t_{x(ij)}t_{y(kl)}] \\ &= \mathbb{E}_0[t_{x(ij)}t_{y(ij)}] - 2\mathbb{E}_0[t_{x(ij)}t_{y(ik)}] + \mathbb{E}_0[t_{x(ij)}t_{y(kl)}\end{aligned}$$

Which is exactly the same as in the constant population size and non-temporal sampling

case! However, the denominator will be different.

$$\begin{aligned}
\sigma_d^2 &= \frac{\mathbb{E}[D^2]}{\mathbb{E}[p(1-p)q(1-q)]} \\
&= \frac{\mathbb{E}_{t_a}[t_{x(ij)}t_{y(ij)}] - 2\mathbb{E}_{t_a}[t_{x(ij)}t_{y(ik)}] + \mathbb{E}_{t_a}[t_{x(ij)}t_{y(kl)}]}{\mathbb{E}_{t_a}[t_{x(ij)}t_{y(kl)}]} \\
&= \frac{\mathbb{E}_0[t_{x(ij)}t_{y(ij)}] - 2\mathbb{E}_0[t_{x(ij)}t_{y(ik)}] + \mathbb{E}_0[t_{x(ij)}t_{y(kl)}]}{(1-\gamma)\mathbb{E}_0[t_{x(ij)}t_{y(kl)}] + \gamma\mathbb{E}_0[t_{x(ij)}t_{y(ik)}]} \\
\sigma_d^2 &= \frac{\frac{\rho+10}{\rho^2+13\rho+18}}{(1-\gamma)\left(\frac{\rho^2+13\rho+22}{\rho^2+13\rho+18}\right) + \gamma\left(\frac{\rho^2+13\rho+24}{\rho^2+13\rho+18}\right)}
\end{aligned} \tag{3.10}$$

It is useful to think about the bounds on the quantity $\gamma = \frac{\rho(1-e^{-t(\frac{\rho}{2}+1)})}{\rho+2}$ in the limits of $t_a \to 0$ and $t_a \to \infty$. When $t_a = 0$, we have $\gamma = 0$ and when $t_a \to \infty$, we have $\gamma = \frac{\rho}{\rho+2}$. With these limiting results, we can similarly calculate the impact of $t_a \to \infty$ on the relative difference (δ) between $\sigma_d^2(\infty)$ and $\sigma_d^2(0)$.

$$\begin{aligned}
\lim_{t\to+\infty} \sigma_d^2 &= \frac{\frac{\rho+10}{\rho^2+13\rho+18}}{\frac{2}{\rho+2}\left(\frac{\rho^2+13\rho+22}{\rho^2+13\rho+18}\right) + \frac{\rho}{\rho+2}\left(\frac{\rho^2+13\rho+24}{\rho^2+13\rho+18}\right)} \\
&= \frac{\rho^2+12\rho+20}{\rho^3+15\rho^2+50\rho+44} \\
\lim_{t\to 0} \sigma_d^2 &= \frac{10+\rho}{22+13\rho+\rho^2}
\end{aligned}$$

The relative difference between these is:

$$\begin{aligned}
\delta &= \frac{\sigma_d^2(0) - \sigma^2(d)(\infty)}{\sigma_d^2(0)} \\
&= \frac{2\rho(\rho+10)}{(\rho^2+13\rho+22)(\rho^3+15\rho^2+50\rho+44)} \frac{\rho^2+13\rho+22}{\rho+10} \\
&= \frac{2\rho}{\rho^3+15\rho^2+50\rho+44}
\end{aligned}$$

We find that the maximum relative difference δ_{max} is achieved at $\hat{\rho} = \frac{3\sqrt{33}-11}{4}$, and at this value $\delta_{max} = 8\frac{3\sqrt{33}-11}{105\sqrt{33}+1991}$. This suggests that the impact of extreme temporal

sampling is really only observable at intermediate ranges of $\hat{\rho} \approx 1.558$, so with an $N_e = 10^4$, this would be at $\approx 3 \times 10^{-3}$ centimorgans which is at a very short recombinational scale.

3.9.3 Appendix 3: Expected time to first coalescent for ancient samples

Here we are interested in the scenario of a single ancient haplotype sampled at time $t_a > 0$ coalescing with a reference panel of size K haplotypes sampled at the present. We define the random variable T^* as the time of a coalescent event involving the ancient haplotype and a lineage ancestral to the modern reference panel after the time that the ancient haplotype is sampled (t_a). The expectation of this quantity can be written as:

$$\begin{aligned}\mathbb{E}_{t_a,K}[T^*] &= \mathbb{E}[\mathbb{E}[T^*|A_K(t_a)]] \\ &= \mathbb{E}\left[\mathbb{E}\left[\sum_{j=2}^{A_K(t_a)+1}\mathbb{P}(I_j)\sum_{i=A_K(t_a)+1}^{j}T_i\right]|A_K(t_a)\right]\end{aligned}$$

The random variable $A_K(t_a)$ is the number of lineages ancestral to the modern reference panel at time t_a. Where $\mathbb{P}(I_j)$ is the probability that the j^{th} coalescent event involves the ancient lineage, and T_i is the i^{th} inter-coalescent time. In a constant population size model $T_i \sim Exp\left(\frac{i(i-1)}{2}\right)$.

Starting at time t_a with n_t lineages, we can calculate the probability that the j^{th} coalescent event involves the ancient lineage as follows:

$$\begin{aligned}\mathbb{P}(I_j) &= \left(1-\frac{\binom{j-1}{2}}{\binom{j}{2}}\right)\prod_{k=A_n(t_a)}^{j+1}\frac{\binom{k-1}{2}}{\binom{k}{2}} \\ &= \left(1-\frac{j-2}{j}\right)\prod_{k=A_n(t_a)}^{j+1}\frac{k-2}{k} \\ &= \frac{2}{j}\prod_{k=A_n(t_a)}^{j+1}\left(1-\frac{2}{k}\right)\end{aligned}$$

In a constant population size model, we have $\mathbb{E}[T_j] = \frac{2}{j(j-1)}$, and thus we can the expected time till the first coalescence involving the ancient lineage (T^*) is:

$$\mathbb{E}[T^*|A_K(t_a)] = \mathbb{E}\left[\sum_{j=2}^{A_K(t_a)+1} \mathbb{P}(I_j) \sum_{i=A_K(t_a)+1}^{j} T_i\right]$$
$$= \sum_{j=A_K(t_a)+1}^{2} \left[\frac{2}{j} \prod_{k=A_K(t_a)+1}^{j+1} \left(1-\frac{2}{k}\right) \sum_{i=A_K(t_a)+1}^{j} \frac{2}{i(i-1)}\right]$$

In our previous expressions we conditioned on the number of lineages ancestral to the modern panel at the time the ancient lineage was sampled ($A_K(t_a)$). This quantity is a random variable and has a corresponding probability distribution $\mathbb{P}(A_K(t_a) = a)$, where $a \in 1...K$ and K is the size of the modern reference panel. Integrating over the distribution of $A_K(t_a)$ yields:

$$\mathbb{E}_{t_a,K}[T^*] = \mathbb{E}[\mathbb{E}[T^*|A_K(t_a)]]$$
$$= \sum_{a=K}^{1} \mathbb{P}(A_K(t_a) = a) \left[\sum_{j=a+1}^{2} \left[\frac{2}{j} \prod_{k=a+1}^{j+1} \left(1-\frac{2}{k}\right) \sum_{i=a+1}^{j} \frac{2}{i(i-1)}\right]\right] \tag{3.11}$$

The probability distribution $\mathbb{P}(A_K(t) = a)$ involves a number of alternating sums and leads rapidly to numerical error as the sample size gets large (see Eq 15 in (Chen and Chen, 2013)). To alleviate this issue we choose one possible approximation, to approximate $\mathbb{P}(A_K(t) = a)$ as $\delta(A_K(t) = \lceil \mathbb{E}[A_K(t)] \rceil)$. Rather than calculate the probability distribution of $A_K(t)$ across states $K...1$, we will approximate it with its expectation $\mathbb{E}[A_K(t)]$. One reasonable approximation for $\mathbb{E}[A_n(t)]$ is found in (Griffiths, 1984):

$$\mathbb{E}[A_K(t_a)] \approx \frac{K}{K + (1-K)e^{-t_a}}$$

Other approximations for this expectation exist and are explored in greater detail in (Jewett and Rosenberg, 2014). We choose the above approximation largely for computational convenience as it does not involve any summation, has a simple form, and is comparably

accurate when compared to other approximations (Jewett and Rosenberg, 2014).

We also explored approximating $\mathbb{P}(A_K(t) = a)$ using a normal distribution, such as those explored by (Chen and Chen, 2013; Griffiths, 1984). We define the mean and variance of this asymptotic distribution using the following results (see Eq 17 and eq 18 in (Chen and Chen, 2013))

$$\begin{aligned} \mu(t) &= \frac{2\eta}{t}, \text{and} \\ \sigma^2(t) &= \frac{2\eta}{t}\frac{(\eta+\beta)^2}{\beta^2}\left(1 + \frac{\eta}{\eta+\beta} - \frac{\eta}{\alpha} - \frac{\eta}{\alpha+\beta}\right) \\ \beta &= \frac{1}{2}t, \\ \alpha &= \frac{1}{2}Kt, \text{and} \\ \eta &= \frac{\alpha\beta}{\alpha(e^\beta - 1) + \beta e^\beta} \end{aligned} \tag{3.12}$$

Since this asymptotic normal distribution is continuous, to approximate $\mathbb{E}[T^*]$ we will approximate it as:

$$\mathbb{E}[T^*] = \sum_{a \in \mathbf{D}} \mathbb{P}(A_K(t_a) = a)\mathbb{E}[T^* | A_K(t_a) = a]$$

The approach here is to weight particular values of $A_K(t)$ according to the asymptotic normal distribution ($\mathbb{P}(A_K(t_a) = a)$) and sum accordingly where $\mathbf{D} = \{\mathbb{E}[A_K(t)] \pm d\}$. In our explorations we set $d = 30$, although we find that in practice this does not change the broad-scale results except for very deep times (Jewett and Rosenberg, 2014). We term the approximation of setting $A_K(t) = \mathbb{E}[A_K(t)]$ as the "Delta" approximation and the normal approximation as the "Normal" approximation.

We find that the expected time till an ancient lineage has its first coalescent event with a member of the modern panel shows interesting behavior as being approximately constant at first and then rising rapidly as a function of age. This rapid increase can be attributable

to the fewer number of lineages ancestral to the modern panel at the time of the ancient lineages sampling, which makes the inter-coalescent intervals longer (Figure 3.14).

More importantly for understanding the context within the haplotype copying model, we can use this expectation to learn about the effect of panel size to be able to copy haplotypes over a particular length scale. For large panel sizes and short sampling times, we can see that the time for the ancient haplotype to coalesce with the panel is quite short and therefore we expect the haplotype copying rate to be fairly small (leading to longer shared blocks). This is the key intuition behind long-range phasing methods that take advantage of recent relatedness within a sample.

As an illustration of this approach we can see that for a sample t_a of 10^{-2} or 10^{-1} there is no appreciable difference in the expectation of the time to the first coalescent event between having a reference panel of size $K = 1000$ and $K = 10000$, suggesting that the reference panel will coalesce quite quickly and larger reference panels are of limited utility for older samples in terms of long stretches of recent relatedness. We find numerically that this is $t_a \approx 0.01$ for the case of $K = 5000$ and $K = 10000$, it is likely that for larger sample sizes this inflection point will occur more rapidly (Figure 3.14).

We also observe that the expected external branch length rises exponentially as a function of the sample age (linear in log-log space) and therefore we expect that there should be an increase in the transition rate of the haplotype copying model. This will by definition create a larger number of transitions and exponentially shrink the average haplotype copying tract.

3.9.4 Appendix 4: The haplotype copying model

We want to explore properties of the generalized Li & Stephens (LS) model. We refer the reader to the original paper (Li and Stephens, 2003) for the full details of the model, but here we detail modifications made to the transition and emission probabilities to generalize the model. The model we describe here is very similar to that described in (Lawson et al., 2012), which uses genetic map distances similar to the way that we do here. Starting with the transition probability between hidden states, X_l, where X_l represents the haplotype in the panel that the test haplotype copies off of at site l. :

$$\mathbb{P}(X_l = x' | X_{l-1} = x) = \begin{cases} e^{-\lambda g_l} + \frac{1}{K}(1 - e^{-\lambda g_l}) & , \text{if} x' = x \\ \frac{1}{K}(1 - e^{-\lambda g_l}) & , \text{else} \end{cases} \tag{3.13}$$

where g_l is the *genetic distance* between markers $l-1$ and l (in Morgans), K is the size of the haplotype panel, and λ is the "jump rate" or rate at which the model moves between the haplotype copying states.

The emission probability distribution of the test haplotype at site l, h_l, can be similarly parameterized using an error parameter ϵ.

$$\mathbb{P}(h_l = a' | X_l = a) = \begin{cases} \epsilon & , a' \neq a \\ (1 - \epsilon) & , a' = a \end{cases} \tag{3.14}$$

Since we treat jumps between the hidden states as a Poisson process with rate λg_l, the scaling factor λ informs us about the mean copying tract length. We note that the genetic map positions are determined *a priori* using a pre-defined genetic map, and λ acts as a scaling parameter here. By comparison, in (Li and Stephens, 2003), the authors use $\lambda = \frac{\rho}{K}$, where $\rho = 4N_e r$ is the population-scaled recombination rate. We use this alternative parameterization because our aim is not to infer recombination rates and we treat the

recombination map as known.

This general model defines the likelihood of our test haplotype on two parameters (λ, ϵ). Using this likelihood we can employ expectation-maximization (EM) or numerical optimization to learn the maximum-likelihood estimates for each parameter. For the simulation results within the main text, we have chosen to use numerical optimization as these are fairly small regions and the likelihood function can be computed efficiently using the Forward algorithm (Rabiner, 1989).

When considering an ancient test haplotype, we consider the effects that the sampling time should have on the two underlying parameters λ and ϵ. First we start with ϵ, as this will be influenced by two features of the data as a function of the samples age t_a : (1) variants that are private to the ancient haplotype and (2) variants that are private to the modern haplotype panel. Both features are dependent on the total external branch length subtending either only the ancient sample (1) or exclusively the modern panel (2). Previous explorations into time-stratified coalescent models have shown that as the maximum sampling time increases, the external branch length leading to only ancient samples also increases (Forsberg et al., 2005). By this argument, we should expect the parameter ϵ to increase as a function of sampling time t_a as there will be a higher proportion of private mutations for the ancient sample. This effect is consistently shown in our simulations as well (not shown).

CHAPTER 4

POPULATION GENOMIC HISTORY OF THE KODAVA: A TEST OF PROPOSED ORIGINS

4.1 Abstract

The Kodava are a population group from south India whose history prior to the 15^{th} century remains largely unknown. With recent surveys of genetic variation from India, contextualizing the Kodava within modern Indian genetic diversity is of interest to the Kodava community. Their oral traditions describe a history with more recent ancestry from central Asian populations, reflecting their unique cultural customs and linguistic differences with neighboring populations. This has led to two primary hypotheses regarding their origins: one that they are more closely related to neighboring populations and developed cultural customs locally (the "local origin hypothesis"), and another that they have substantial ancestry from a more distant founding with larger genetic contributions from western Eurasia ("non-local origin hypothesis").

To address these hypotheses, we generated new data from $n = 119$ individuals of Kodava ancestry living in south India and the United States, as well as $n = 66$ individuals from neighboring populations to the Kodava in south India. We merged these data with existing population genetic data from Eurasia and conducted genome-wide analyses based on principal components analysis, ADMIXTURE, and allele-frequency differentiation to assess patterns of ancestry contributions.

We find that the genetic ancestry patterns in the Kodava individuals are similar to neighboring south Indian populations, suggesting evidence for the "local origin hypothesis", though we cannot exclude versions of the non-local origins hypothesis where a modest fraction of ancestry is from non-local sources. This study provides a better understanding of the genetic diversity of the Kodava population in the context of south Indian genetic variation,

broadening our understanding of genetic diversity in this region of the world.

4.2 Introduction

India is home to more than 1.3 billion people, consisting of many small endogamous groups with a large degree of genetic differentiation between them (Majumder, 2010; Mastana, 2014). Many of these groups can be modeled as a mixture of ancestral north Indian (ANI) and ancestral south Indian (ASI) ancestry, forming a cline of ancestry in modern-day Indian groups correlating strongly with geographic and linguistic structure (Reich et al., 2009; Narasimhan et al., 2019) (although see Basu et al. (2016) for a deeper description of Austro-asiatic and Tibeto-burman ancestry clusters in Indian populations). Several Indian populations also exhibit strong founder effects, leading to elevated levels of linkage disequilibrium and reduced heterozygosity, with a predicted increases in recessive disease risk (Peltonen et al., 2000; Moorjani et al., 2013; Nakatsuka et al., 2017). Taken at a broad-scale, many Indian populations can be modeled as falling on a gradient of ANI to ASI genetic ancestry followed by endogamy within smaller population groups.

The Kodavas are a putative founder population in south India. The Kodava live in Kodagu, a district in south Karnataka state (Census of India, 2011). The earliest documented evidence of Kodavas in Kodagu come from stone inscriptions approximately 800 - 900 CE, with depictions of their clan-based social structure, called *okkas*(Kamat, 1993; Kushalappa, 2013). While they are no longer the majority inhabitants of Kodagu, they are thought to be among the original settlers of the region (Kushalappa, 2013).

Kodavas show several unique cultural and linguistic differences from neighboring populations in the surrounding region. Kodavas speak *Kodava-thak*, a language in the "South Dravidian" language family (Emeneau, 1967; Krishnamurti, 1985). However, there are structural differences to *Kodava-thak* that differentiate it from neighboring Dravidian languages. For example, Balakrishnan (1976) notes that while the majority of Dravidian languages

typically have five long and short vowels, *Kodava-thak* has an additional short and long vowel. Culturally, Kodavas have unique traditional dress (Ganapathy, 1967; Kamat, 1993) and religious practices that differentiate them from neighboring populations. For example, in contrast to neighboring populations, Hindu priests do not preside over important events such as marriage and funerals in their culture (Ganapathy, 1967). These cultural differences have played a central role in establishing Kodava identity within southern Indian society.

The cultural differences between the Kodava and neighboring populations have also led to hypotheses on Kodava origins. These hypotheses suggest potential sources of ancestry for the Kodava prior to their arrival in Kodagu and have been incoporated into their oral history. One hypothesis posits a model of indigenous or "local origin", namely that the Kodava are the descendents of the earliest settlers in the region and are similar in ancestry to present-day neighboring populations (Kamat, 1993). An alternative hypothesis is that the Kodavas are descended from an original ancestral source in central Eurasia, which we call the "non-local origin hypothesis". Proposed sources of non-local origins include the Scythians from the central Asian steppe, Kurdish groups in Iran, and more speculative suggestions that the Kodava population were founded by deserting soldiers from Alexander the Great's army after reaching the Indus river and who migrated south to Kodagu (Balakrishnan, 1976; Ponnappa, 1999). To our knowledge, none of these hypotheses have been tested using genetic data. In this study, we aim to use patterns of genetic variation in Kodava individuals in order to (a) gain a better understanding of the Kodava in the context of south Indian genetic diversity and (b) to assess the genetic evidence for the local or non-local origins hypotheses. Under the non-local origin hypothesis, the key signal we expect is a higher proportion of ancestry in the Kodava coming from western Eurasian sources when compared to neighboring Dravidian populations in south India.

To address these two aims, we generated a dataset of Kodava ancestry individuals and neighboring populations from south India. The Kodava dataset consists of $n = 104$ indi-

viduals of Kodava ancestry sampled in the United States and $n = 15$ Kodava individuals from south India. All of these samples were whole-genome sequencing samples with different coverage between the US-based and India-based subsets (6.07× vs 2.91× median coverage respectively). After quality control 4.6.2, we retained $n = 91$ individuals who have all recently immigrated to the United States and have at least 4 grandparents who lived in Kodagu. We also generated a dataset of $n = 66$ neighboring populations to the Kodava in south India. The populations included in this second dataset are the Bunt ($n = 10$), Kapla ($n = 10$), and Nairs ($n = 46$) and were sequenced to a median coverage of 2.91× (Figure 4.2).

The populations in this second subset are useful for comparing the Kodava with for testing the local and non-local origins hypothesis as they are from the same language family (Dravidian) and are geographically from the same region. We then merged the called genotypes from our two newly sequenced datasets with publicly available Eurasian population genetic data to perform our downstream inference of population structure (Wall et al., 2019; Bergström et al., 2020; Nakatsuka et al., 2017) (Figure 4.1) (See 4.6.4 for additional details on merging and specific datasets).

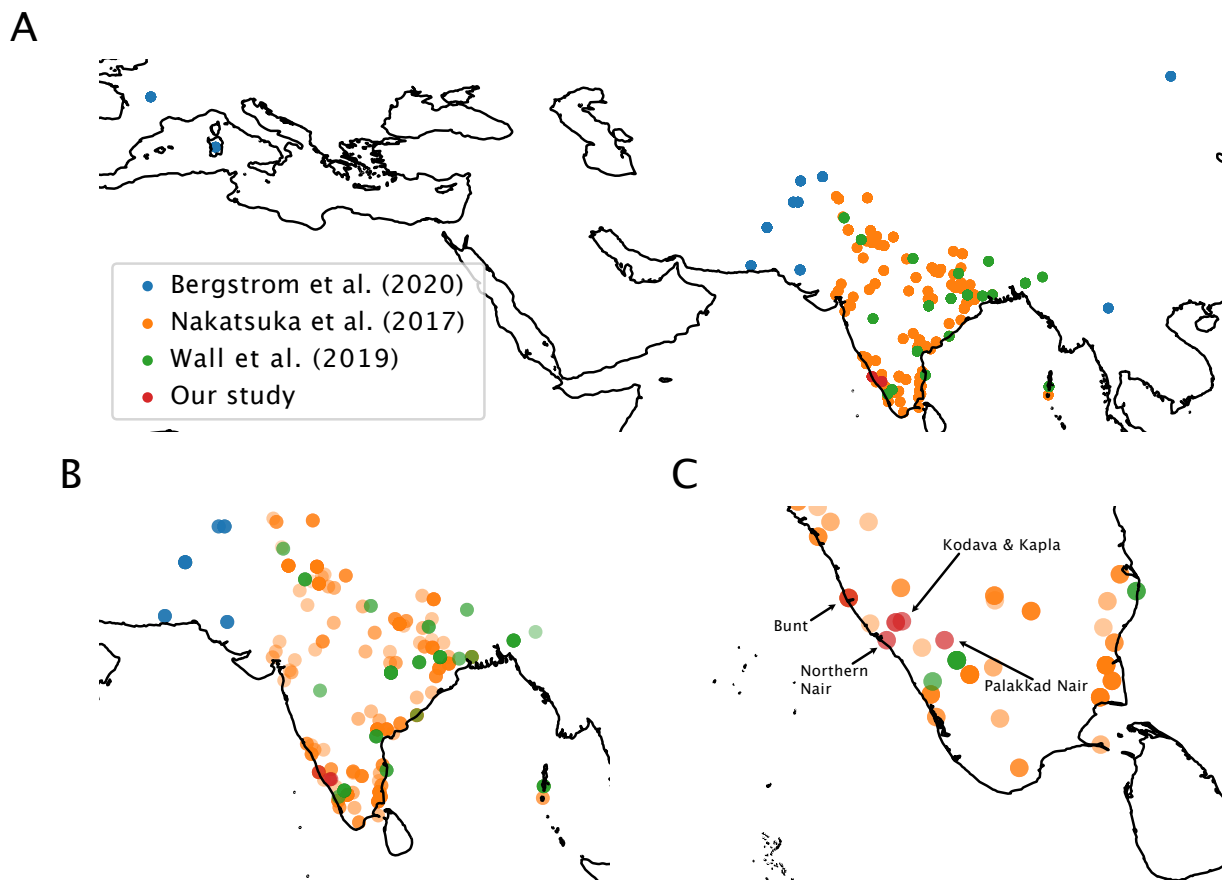

Figure 4.1: Legend entries correspond to data from specific publicly available datasets (Bergström et al., 2020; Nakatsuka et al., 2017; Wall et al., 2019). **(A)** Map of samples used in this study. **(B)** A zoomed-in map of sampling across India. **(C)** A further zoomed-in map of south India reflecting the location of our samples relative to previously sampled diversity in south India. US-based Kodava samples are not shown.

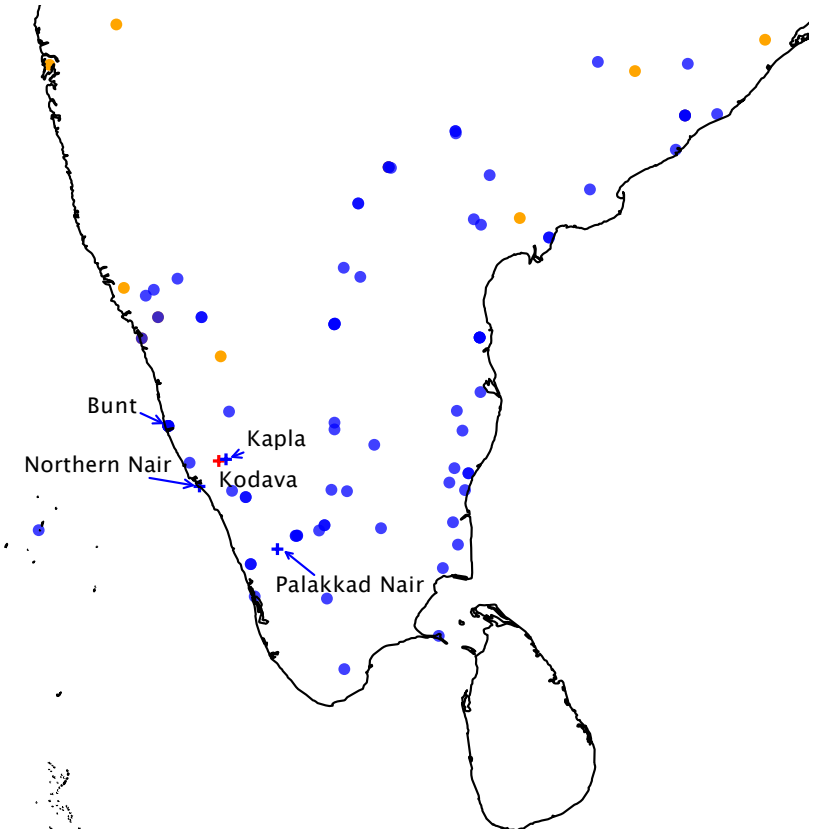

Figure 4.2: Locations of Dravidian-speaking populations in south India (blue dots) used as the set of overall neighboring populations. The geographic location of the Kodava (red cross) and specifically used comparator populations (blue crosses) are also shown. Populations that speak an Indo-european language are shown in orange dots.

4.3 Results

4.3.1 Kodava individuals within the context of Indian genetic diversity

We applied principal components analysis (PCA) to gain a clearer picture of the Kodava in terms of Indian genetic diversity. First, we replicate previous findings in the literature, in that we observe the ANI/ASI cline described by multiple genetic studies in India using PCA (Reich et al., 2009; Moorjani et al., 2013; Nakatsuka et al., 2017) (Figure 4.3A).

Under the non-local hypothesis with higher central/western Eurasian ancestry in the Kodava, we expect that the Kodava (both US-based and Indian cohorts) should appear closer to the ANI end of the ancestry cline when compared to neighboring populations. We find that both the US-based Kodava and south India Kodava samples cluster in PC-space with neighboring populations, such as the Bunts and Nairs (Figure 4.3A,B). Extending our analysis to a larger number of principal components or admixture components (K) does not change our conclusion that the Kodava have similar genetic ancestry to neighboring populations in south India (Figures 4.8).

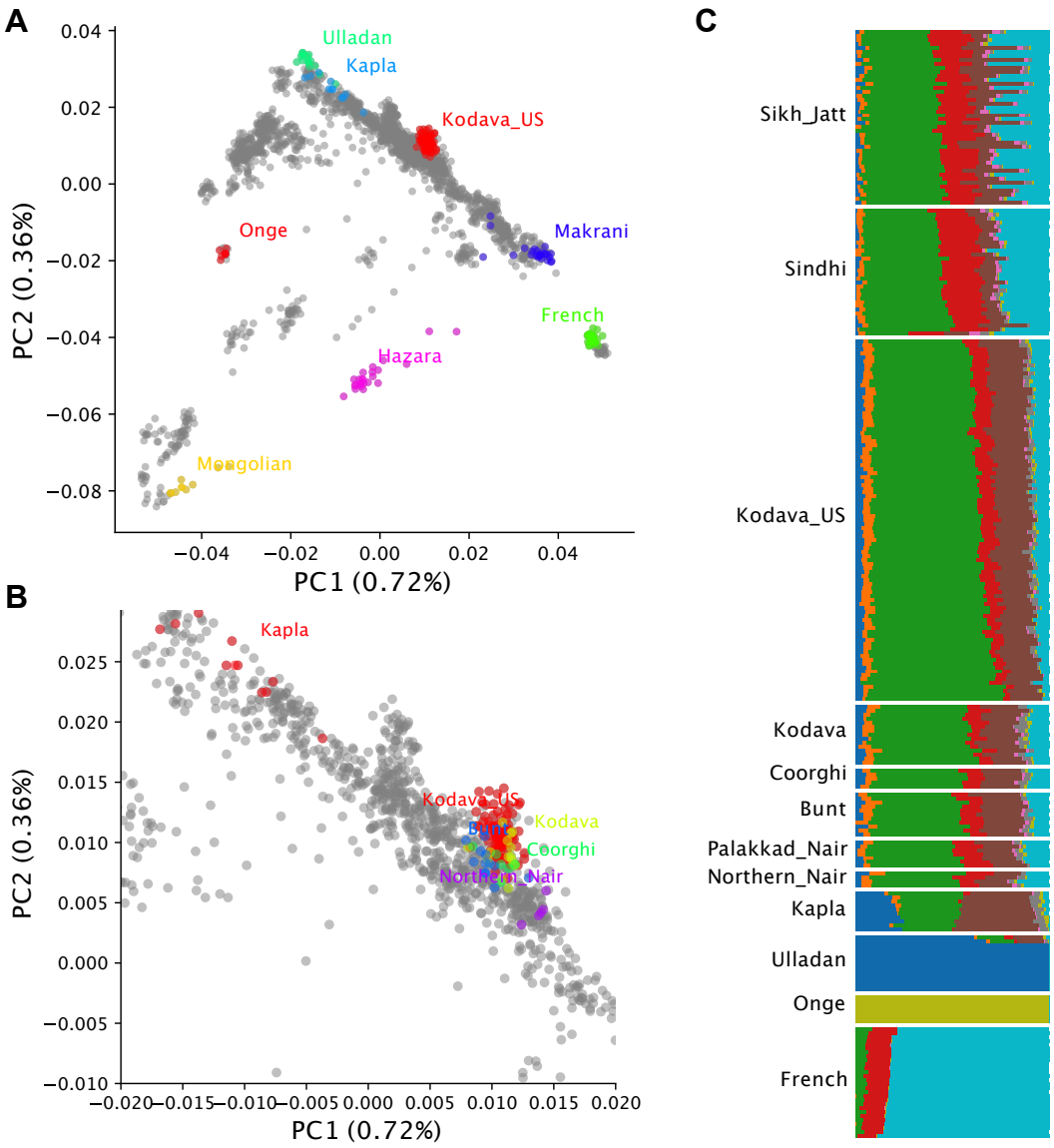

Figure 4.3: **(A)** PCA of the merged dataset with samples from our newly sequenced individuals, and publicly available datasets (Bergström et al., 2020; Wall et al., 2019; Nakatsuka et al., 2017). **(B)** Principal components zoomed into region with novel data generated as a part of this project. **(C)** ADMIXTURE ($K = 9$) results subsetted to new data generated from this project and proxies for western Eurasian ancestry (e.g. French), Ancestral South Indian ancestry (Onge), and a tribal south Indian population (Ulladan). In panel **(C)** we have included the "Sikh_Jatt" and Sindhi populations as well to show two representative populations from north India.

One newly sampled population in our dataset, the Kapla, are a visual outlier in genetic ancestry from the other newly sampled populations (Figure 4.3A,B). The Kapla are a small indigenous population of approximately $\sim$ 160 members living in a single village called *Kaplakeri* in Kodagu (Kushalappa, 2018). The Kapla are closer in PCA to other indigenous populations in south India like the Ulladan (Figure 4.3C). Kushalappa (2013) hypothesized that the Kapla may be more closely related to Siddis, descendants of African slaves brought over by sailors to Karnaktaka approximately 300-500 years ago (Shah et al., 2011; Narang et al., 2011). We find little support for this hypothesis based on comparisons of F_{ST} including Siddi populations from Karnataka (Figure 4.13). For example, $F_{ST} = 0.118$ between the Kapla and the Siddi population from Karnataka is observed to be higher (for contrast, $F_{ST} = 0.078$ between the Kapla and the US-based Kodava population).

4.3.2 No evidence for increased western Eurasian affinity in the Kodava compared to neighboring populations

We are also interested in determining what specific source populations may have contributed more to the Kodava over neighboring populations. While PCA and ADMIXTURE are useful for assessing where the Kodava fall on the ANI/ASI cline, our analysis here is not suited to directly testing whether a specific western/central Eurasian source population (e.g. Greek) have contributed more ancestry to the Kodava over neighboring populations. To address this, we turn to tests based on allele frequency differentiation to test hypotheses of ancestry contributions from specific sources as evidence to resolve the local vs. non-local origins hypotheses for the Kodava.

To statistically test the hypothesis of increased western Eurasian affinity in the Kodava relative to neighboring populations in south India we use f-statistics, which measure the covariance of allele frequency differences between populations. We specifically use the outgroup f_3 statistic.Outgroup f_3 statistics measure the extent of shared drift between two

populations relative to an extant outgroup based on allele frequency covariance between the two tested samples (Patterson et al., 2012; Raghavan et al., 2014) (Figure 4.4A).

Under the null hypothesis (e.g. local origins), the Kodava are genetically similar to neighboring south Indian populations and we expect outgroup f_3-statistics between each neighboring south Indian population and a non-local source from central Eurasia to be equal due to a shared history. To extend our analysis to include more samples as potential sources of central Asian ancestry, we combined our merged data with additional data from a collection of 10,061 samples typed at 597,573 sites on the Affymetrix Human Origins Array (see Section 4.6.4 for more details). Including this additional source of data is important to test for the contribution of different central Asian ancestry sources to the Kodava.

We propose a set of proxies for potential sources of higher western Eurasian or central Asian (Scythian) affinity in the Kodava. We used all sampled populations from south India to focus on relevant comparisons with neighboring and spatially close groups (see Section 4.6.5 for details).

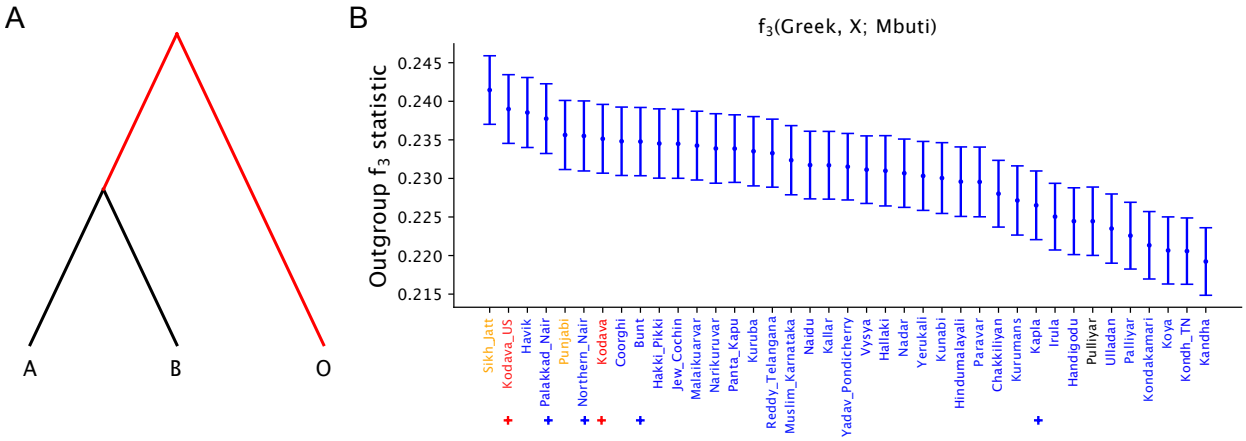

Figure 4.4: **(A)** Outgroup f_3 statistics aim to quantify the extent of shared drive between populations A and B, and provide an estimate of the branch-length shown in red. **(B)** Outgroup f_3 with modern Greek individuals across a subset of Dravidian populations from India (names in blue). The Kodava are denoted in red letters and newly sequenced samples that are used as comparators are shown with a blue cross (similar to Figure 4.2). Error bars represent two standard errors of the mean outgroup f_3 statistic. NOTE: we have included the "Sikh_Jatt" and "Punjabi" populations within the plot as well to reflect the outgroup f_3 value for two populations from northern India and that speak an Indo-European languages (shown in orange), but we do not use these populations in any downstream hypothesis testing.

We find that across all neighboring Dravidian populations, the US-based Kodava cohort exhibits the closest genetic affinity to the Greek population (Figure 4.4B) which is what we use as a proxy for western Eurasian ancestry (see Figures 4.12 for alternative focal populations). We used the Greek population in our main analysis to test the hypothesis of recent Greek ancestry in the Kodava, which is a source of claimed ancestry within Kodava oral history. Specifically, we want to test whether the mean outgroup-f_3 statistic for the Kodava is equal to that of the neighboring populations. For this test we compare with neighboring populations that satisfy both of the following: they are (1) geographically in south India (see 4.6.5 for definition) and (2) speak a Dravidian language.

When testing that the mean of the outgroup f_3 statistic for the US-based Kodava is equal to neighboring population, we fail to reject this null hypothesis ($p = 0.098$; One-way

ANOVA). We can also test for equality of the means of the outgroup f_3 statistic with a specific comparison population (e.g. Bunt). When we compare the Kodava (US) and the Bunt using this simpler pairwise comparison, we also do not reject the null hypothesis of equal outgroup f_3 in both populations ($p = 0.179$; t-test). Although the US-based Kodava cohort appears to share more genetic drift with the Greek population, it is not a statistically significant signal when compared with all other neighboring populations and we cannot reject the null hypothesis of similar western Eurasian ancestry in the Kodava (the "local origins model").

One concern regarding the analysis above is that we may be under-powered to detect an effect of increased central Eurasian ancestry in the Kodava with the outgroup f_3 statistic. To evaluate power we use a simulation framework based on masking genotypes and replacing them with genotypes from a putative source of non-local ancestry (see section 4.6.6 for more details). For each simulated fraction of admixture from an central Eurasian source, we evaluated the power to detect a difference between the outgroup f_3 statistic of the Kodava and the Bunt population. We focus on the pairwise comparison with the Bunt since they are geographically close to the Kodava population and speak a Dravidian language and to keep the hypothesis simpler.

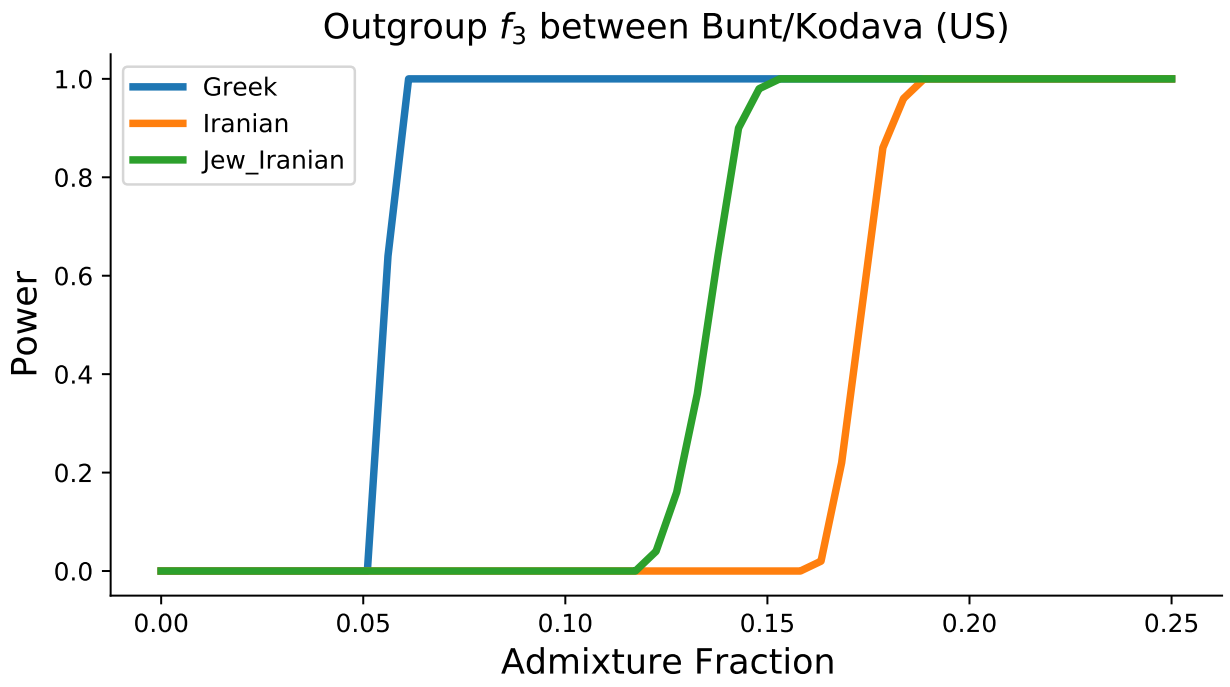

Figure 4.5: Power to reject the null hypothesis that the outgroup f_3 statistics between the Bunt and Kodava (US) are the same using Welch's t-test. Each line corresponds to a particular source of ancestry admixed into the Kodava at a fraction shown on the x-axis and in all situations we use Greek as the comparison population and Mbuti as the outgroup (population B and O in Figure 4.4 respectively). See 4.6.6 for details on simulation performed.

Admixture Source	Age (calibrated date BP)	Fraction of admixture for 80% Power
Anatolia_N	8599.7	0.087
Assyrian	0	0.137
Georgian	0	0.112
Greek	0	0.0612
Iran_C_TepeHissar	4742.6	> 0.25
Iranian	0	0.178
Iranian_Bandari	0	> 0.25
Jew_Iranian	0	0.142
Lebanese_Muslim	0	0.183
Turkish	0	0.132
Ukraine_Neolithic	7382.8	0.132
Ukrainian	0	0.097

Table 4.1: Evaluation of power to detect differences in outgroup f_3 statistics measuring shared drift between the focal population and the modern Greek population, using Mbuti as an outgroup. All comparisons in this table are between the outgroup f_3 statistics with Kodava (US) population and the Bunt population as focal populations. Power is evaluated as the fraction of replicate tests where the p-value for Welch's t-test is < 0.05, and estimated from 50 replicate simulations. We test admixture fractions up to 0.25 and have added the first admixture fraction at which there is $\geq 80\%$ power to reject the null hypothesis. See Figure 4.5 for variation in power as a function of the admixture fraction.

From the power analysis, we find that we cannot detect small amounts of non-local ancestry that differ between the Kodava and the Bunt populations. In the case of direct admixture from the source population (modern Greeks) we require > 0.061 of admixture for 80% power to detect the effect (Table 4.3.2). This is a lower bound on the fraction of admixture nec-

essary to reject the local-origins model using the difference between outgroup f_3 statistics. Furthermore, in the case that the test population (e.g. Greek) and admixture source (e.g. Iranian) are not the same but more distantly related, we find that more admixture is required for 80% power (admixture fraction of 0.178 with Iranian population as admixture source). The fraction of ancestry required for 80% detection power is qualitatively similar when using different potential central Asian test populations (Supplementary File 4.1)

4.4 Discussion

The first goal of this study was to characterize the Kodava in light of modern Indian genetic diversity. We find that the Kodava cluster with neighboring Dravidian populations in south western India, suggesting limited isolation from neighboring populations in the region. This is corroborated with anthropological evidence showing connections between the Bunts, Nair, and Kodava populations (Iyer and Iyer, 1969). In this context, it is perhaps less surprising that we do not observe a statistically significant genetic differentiation between the Kodava and neighboring Dravidian-speaking populations.

We sampled multiple novel populations in this study, the Bunts, Nairs, Kodava, and the Kapla. Among these new populations we find that the Kapla population contains a substantially higher proportion of ASI ancestry than any other group newly sampled as a part of this study, despite being from a similar geographic region. While the Kapla were not a focal point of the current study, we anticipate future studies to investigate their specific population history and status as a population isolate in the region.

Our second goal in this study was to assess the extent of genetic support for the "non-local" versus the "local" origins model for the Kodava population. Based on PCA and ADMIXTURE analyses, we do not find that the Kodava cluster separately from other neighboring dravidian populations in south India (e.g. Bunt). Based on the comparison of outgroup f_3 statistics we do not find a statistically significant signal of non-equal non-local ancestry in

the Kodava relative to the neighboring Dravidian populations ($p > 0.05$; one way ANOVA). The test of the null hypothesis of equal outgroup f_3 statistics between the Kodava and neighboring populations is consistently not statistically significant across numerous source populations including modern Greek, ancient Scythian, and Iranian populations (Figures 4.12).

When evaluating the power to detect non-local ancestry using outgroup f_3 statistics we find that large amounts of non-local ancestry are required to differentiate between the US-based Kodava and Bunt populations. For example, $> 6\%$ admixture is necessary for $\sim 80\%$ power to reject the null hypothesis that the outgroup f_3 statistics are equal using Welch's t-test). While we do not find evidence supporting the non-local origins model for the Kodava based on outgroup f_3 statistics, we do not have power to reject models with more subtle contributions of non-local ancestry into the Kodava. Further analysis using patterns of haplotype-sharing may be better powered to detect signals of more subtle admixture (Lawson et al., 2012; Hellenthal et al., 2014).

A complication regarding the interpretation of local vs. non-local origins for the Kodava is that the first evidence of the establishment of the Kodavas (800 - 900 CE) is during the time which the original ANI/ASI admixture is estimated to have occurred (1900 - 4900 years ago) (Moorjani et al., 2013). Any additional western Eurasian ancestry contributed to the Kodava during that time may be difficult to disentangle from the original mixture event. Moorjani et al. (2013) used signatures of linkage disequilibrium decay to highlight that for several Indo-European and Dravidian-speaking populations, multiple pulses of admixture between different sources of western Eurasian ancestry are required to explain the linkage disequilibrium patterns. Metspalu et al. (2011) also argues for multiple potential sources of western Eurasian ancestry in many Indian populations. However, even under multiple pulses of admixture, the magnitude of ANI ancestry should still be reflected in PCA by the position of the population along the ANI/ASI cline. Based on the PCA in Figure 4.3B, we find that

the Kodava have a similar ANI fraction to neighboring populations such as the Bunts.

Across multiple sources we find that the US-based Kodava cohort has a higher mean genetic affinity to western Eurasian sources (Figure 4.3, 4.12). When testing the null hypothesis that the outgroup f_3 statistic between the US-based Kodava and the Indian Kodava population are equal we fail to reject this null hypothesis ($p = 0.221$; Welch's t-test) when using the Greek population. One hypothesis is that biased migration to the US of Kodava individuals is assortative by the fraction of ANI ancestry and this contributes to the higher mean genetic affinity to western Eurasian sources within our US-based Kodava dataset. There is evidence of social structure and endogamy within Indian populations leading to biases in population genetic inference, and it is possible that such social structure (which may or may not be aligned with ancestry) may impose societal restrictions on recent immigrants to the US (Pemberton et al., 2012). Immigrant communities in the US can be genetically heterogeneous (Dai et al., 2020), and it is interesting future direction to consider the degree that immigrant communities reflect ancestry from their ancestral region.

The genetics of the Kodava can also improve our understanding of complex traits and disease within the population and south India more broadly. South India is underrepresented in terms of studies of complex traits and learning about risk factors specific to the region is central to precision medicine efforts in India (Wall et al., 2019; Popejoy and Fullerton, 2016). The social structure and endogamy within Indian populations also presents a unique opportunity to discover highly penetrant recessive variants with implications for disease risk (Nakatsuka et al., 2017; Narasimhan et al., 2016; Sivasubbu and Scaria, 2019). In this work, we have focused on the intersection of our newly generated whole genome sequencing dataset with previous surveys of polymorphism on genotyping arrays. However, interrogating the full sequencing data will allow for a richer characterization of functional variation in the Kodava population and a clearer picture of disease risk in south India.

In summary, this study characterizes the genetic diversity of the Kodava population in

the broader context of Indian populations and locally within south India. We find that the Kodava are genetically similar to many Dravidian-speaking neighboring populations, although we cannot rule out more subtle signals of non-local admixture. To provide a perspective of our results to the Kodava community, we have also generated a frequently asked questions section as an appendix based on questions asked during the North American Kodava Koota meeting in 2019 (Appendix 4.8.1). We encourage future studies of population isolates in south India to understand the human demographic history of this understudied region in greater detail.

4.5 Acknowledgements

We would like to thank and acknowledge all of the participants from the "Kodava Koota" community who donated their DNA for this project. We also would like to acknowledge the members of the Nair, Kodava, Bunt, and Kapla communities in India for their DNA contribution as well. We also thank Anna Di Rienzo, John Novembre, Matthew Stephens, and Matthias Steinrücken for comments that greatly improved this manuscript.

4.6 Materials and Methods

4.6.1 Cohort description, DNA extraction and library preparation

We included 104 individuals who volunteered from the North American Kodava Society to donate their DNA. Saliva samples were collected using the Oragene DISCOVER OGR-500 kit. DNA was extracted following the manufacturers instructions. In addition, a second extraction was performed using the QIAamp DNA Mini Kit, automated on the QIAcube. DNA extracts were submitted to Novagene (USA) for library preparation and sequencing using the NovaSeq 6000 platform with paired-end 150 bp reads. All participants provided informed consent, with protocols approved by the institutional review board of the University

of Chicago.

4.6.2 Short read alignment and genotype calling pipeline

The "Kodava (US)" dataset includes 104 individuals sequenced to an average sequencing depth of 6.07x. Our pipeline for alignment and variant calling consists of the following steps: (1) trimming reads using AdapterRemoval v2.2.3 with default settings (Schubert et al., 2016), (2) aligning reads to the hg19 human reference genome using `bwa-mem` (Li, 2013), and (3) removing unmapped reads and PCR duplicated reads. We applied two complementary variant calling strategies, one using `samtools mpileup` (Li, 2013) and one using the GATK best practices (Poplin et al., 2017; Van der Auwera et al., 2013).

The overall goal was to retain variants called by both genotyping pipelines to ensure a reduction of false-positive variant calls. In both cases we filter variants with a quality score < 30 and we take the intersection of variants called by each strategy using `bcftools +isecGT` and keep only biallelic autosomal variants for downstream population genetic analyses. This approach is a conservative way to limit the impact of false-positive genotype calls.

4.6.3 Low-coverage South Indian sample processing

We additionally processed 81 samples from south India from 4 populations: Kodava (15 individuals), Bunt (10 individuals), Kapla (10 individuals), and Nairs (46 samples). Additionally, the Nairs population grouping has three separate population groups as well (Pallakad Nair, Palakkad Nair Menon, and Northern Nair populations), based on self-assigned cultural identities and location within Kerala. The median sequencing depth across this set of samples is 2.91x, which is lower than the "Kodava (US)" cohort. To avoid difficulties in calling heterozygotes and missing genotype calls at sites relevant to downstream merges, we used pseudo-haploid calling for this set of low-coverage south Indian samples at positions that are retained across the entire merged dataset (see 4.6.4). Psuedo-haploid calling randomly

samples a single read overlapping a site for an individual. For analyses requiring population allele frequency information (like f-statistics), the use of psuedo haploid genotypes is not expected to bias the allele frequency estimates (Patterson et al., 2012).

4.6.4 Merging with external datasets

To provide additional population genetic context, we merged our low-coverage samples from south India with the "Kodava (US)" cohort and the following datasets:

- 1,662 individuals genotyped at on the Affymetrix Human Origins array from Nakatsuka et al. (2017)
- 929 individuals from the Human Genome Diversity Project (HGDP) (Bergström et al., 2020)
- 1,163 publicly available individuals whole-genome sequencing data from the Genome Asia 100k dataset (Wall et al., 2019)

The dataset from Nakatsuka et al. (2017) is the only dataset where variants are typed using a genotyping array and not whole genome-sequencing, and therefore it is the dataset that decreases the number of biallelic variants the most. However, we have included it in our analyses here because it provides more population genetic context across the Indian subcontinent. Following the merging of data, our dataset consists of 3804 samples typed at 499,158 autosomal bi-allelic single-nucleotide polymorphisms. This is the primary dataset we use for our downstream population genetic analyses. We find that overall this merged dataset has low per-variant missingness (Figure 4.1)

For the analysis concerning f-statistics, we took the merged dataset above and merged it with a set of 10,061 unique individuals (3589 ancient, 6472 modern) typed at 597,573 biallelic single-nucleotide polymorphisms on the Affymetrix Human Origins array publicly available (see Section **??**). We find that this modestly reduces the number of variants,

as Nakatsuka et al. (2017) is also typed on the Human Origins array. After merging and controlling for missingness (plink flags) in the US-based Kodava cohort, we are left with a dataset consisting of 13,036 modern and ancient individuals at 402,987 bi-allelic autosomal variants. This merged dataset is the version used in Section 4.3.2.

4.6.5 Population genetic methods and analyses

Unless otherwise stated, we use the following filters for our population genetic analyses: (1) filter to variants with $< 5\%$ missingness, (2) prune for linkage disequilibrium (`--indep-pairwise 200 25 0.4`), (3) filter to variants with MAF $> 5\%$, and (4) filter out individuals up to 2^{nd} degree relatives using the software `KING` (Manichaikul et al., 2010). For principal components analysis we used `plink` (Chang et al., 2015), with no outlier removal iterations. When running `ADMIXTURE`, we ran 5 independent replicates for each value of $K \in \{6, \cdots, 14\}$ and present our primary results for the value of $K = 9$ since that is the value that minimizes the cross-validation error (Figure 4.8) across all replicates (see Figure 4.11 for multiple values of K displayed)

We use the `qp3pop` and `f4ratio` programs to compute the f-statistics, as part of the `ADMIXTOOLs` software package (Patterson et al., 2012). Since we are primarily concerned with patterns of ancestry in south India (and neighboring populations to the Kodava), we use a geographic bounding box for latitude and longitude: $N20^\circ >$ latitude $> N5^\circ$ and $E70^\circ <$ longitude $< E85^\circ$ (see Figure 4.2 for a visual depiction of this bounding box). This captures many spatially proximal populations, but we restrict our attention to the .

4.6.6 Power simulations for testing local and non-local hypotheses

To evaluate our power to test the non-local vs local hypotheses we used a simulation framework based on the resampling the real data. The power to reject the hypothesis that the outgroup f_3 statistic is equal between the Kodava and a neighboring population provides

bounds on the level of non-local ancestry that is detectable in the Kodava.

First we denote each population (collection of samples at M genotypes) in our power simulations for the outgroup-f_3 statistic:

- O - outgroup population (typically Mbuti in our applications)
- X' - focal population (typically Kodava)
- X - comparator population (e.g. Bunt, Northern Nair)
- B - ancestry source for outgroup f_3 test
- B' - source of ancestry for artificial simulation

For our simulations we conduct the following steps:

1. Choose a fraction β of genotypes to mask in samples from population X'
2. For each masked genotype in the i^{th} individual from population X', resample their genotypes at random coming from the source population samples B'. This step results in a fraction α of the genotypes in X' being sampled from B'.
3. Using this newly synthetically "admixed" set of individuals from population X', then calculate the outgroup f_3 statistic between $f_3(O; X', B)$ and $f_3(O; X, B)$ to estimate the effect of the source admixture on the difference between the statistics.
4. Test the null hypothesis $H_0 : f_3(O; X', B) = f_3(O; X, B)$ using Welch's t-test to account for unequal variances of the two samples.
5. Repeat the above steps for 50 replicates for each values of β to average over sampling variation across the genotypes and calculate the power as the proportion of p-values $< \alpha = 0.05$. Note that this specifically tests against a single other population, not a set of populations.

When $\beta = 0$, this reflects the baseline case where X' is not a synthetically admixed and is exactly the same as comparing the outgroup f_3 statistic between X' and X (e.g. Kodava vs. Bunt). We use a grid of admixture fractions $\beta \in \{0, ..., 0.25\}$ to visually inspect the power to detect a difference in the outgroup f_3 statistic as a result of the simulated admixture fraction

4.5. We additionally provide a supplementary table with the estimates of power to detect a difference in the outgroup-f_3 means with a given set of O, X', X, B, B' (Supplementary File 4.1)

4.7 Supplementary Figures

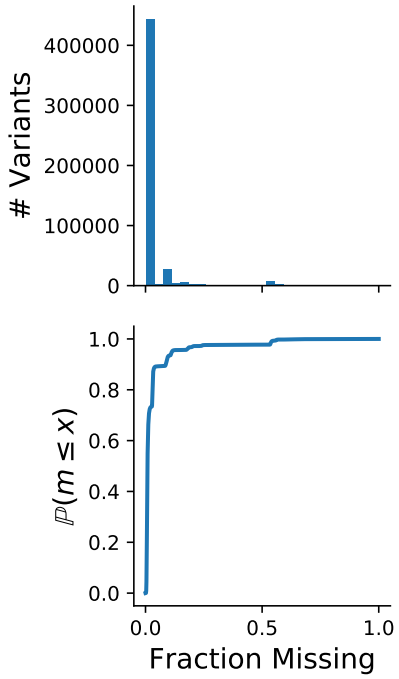

Figure 4.6: **(Top)** Number of variants against the proportion of individuals missing. **(Bottom)** empirical cumulative distribution of sites against the missingness fraction.

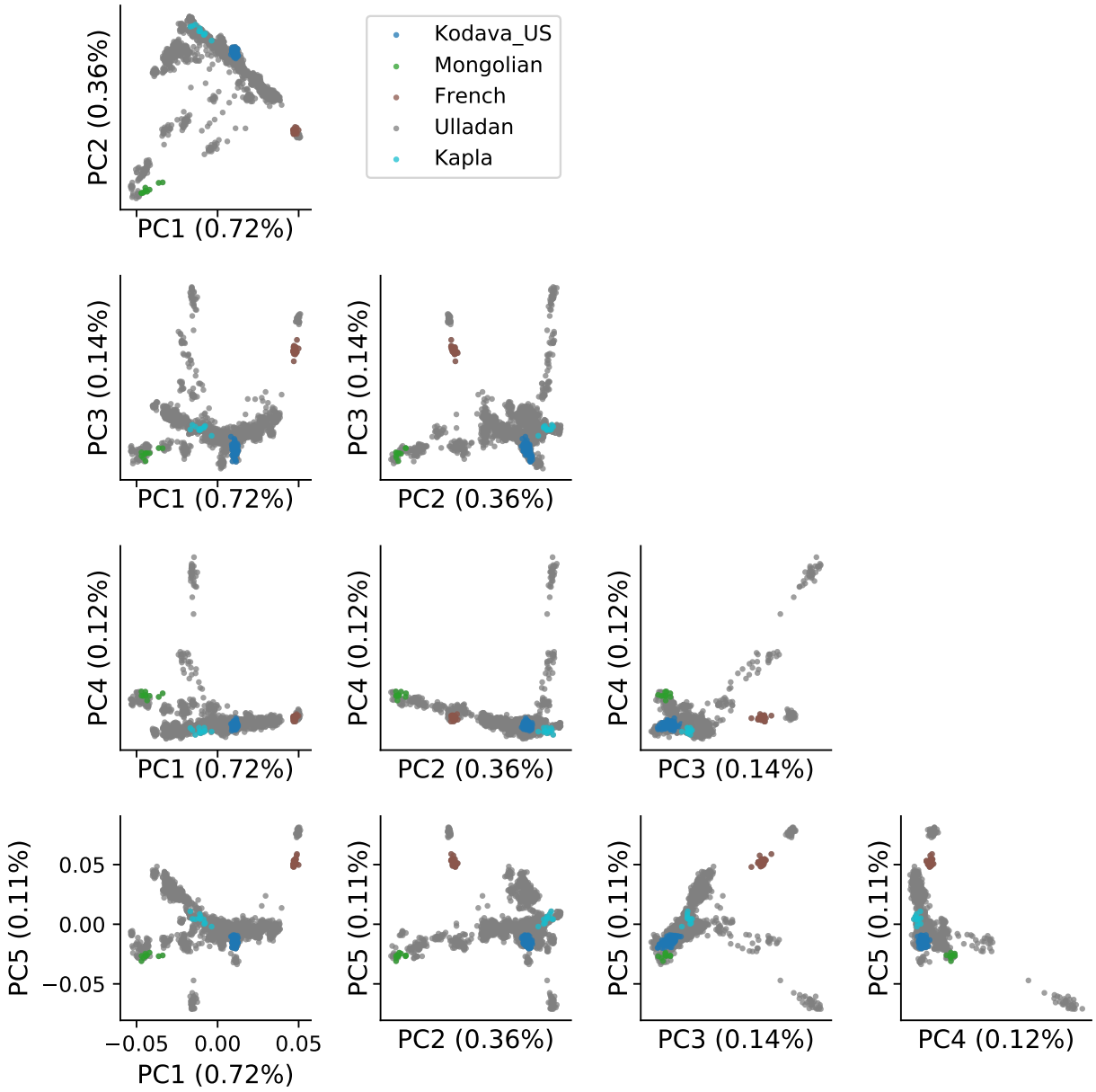

Figure 4.7: Principal components analysis across multiple PCs. We removed populations from the Andaman islands (Onge, Jarawa) to more clearly depict the population genetic structure in south India. We highlight the Kodava and the Kapla as two relevant points of reference within our dataset to explore across higher PCs.

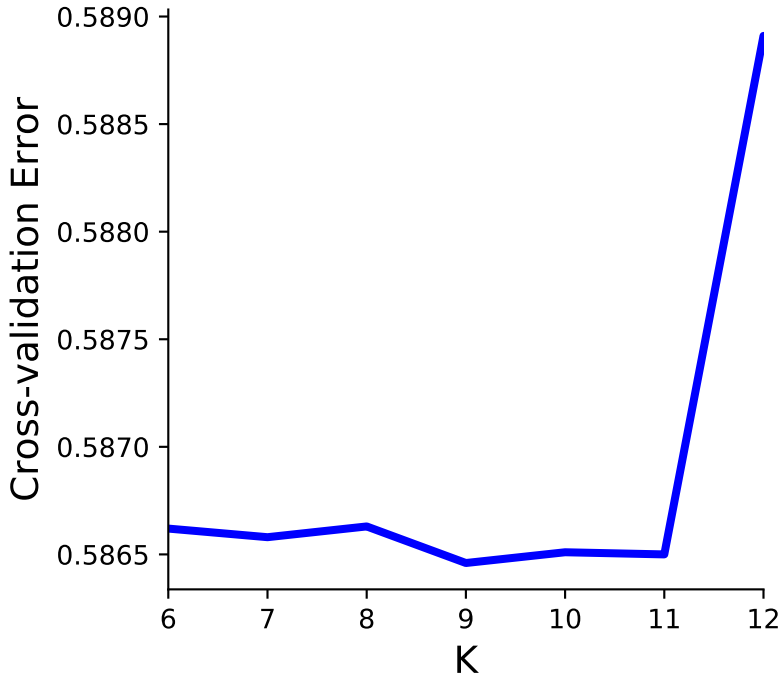

Figure 4.8: 5-fold cross validation error in the `ADMIXTURE` model as a function of K in `ADMIXTURE` from $K \in \{6, ..., 14\}$. Note that we have only shown values from $K \in \{6, ..., 12\}$ to show the minimum cross validation error at $K = 9$.

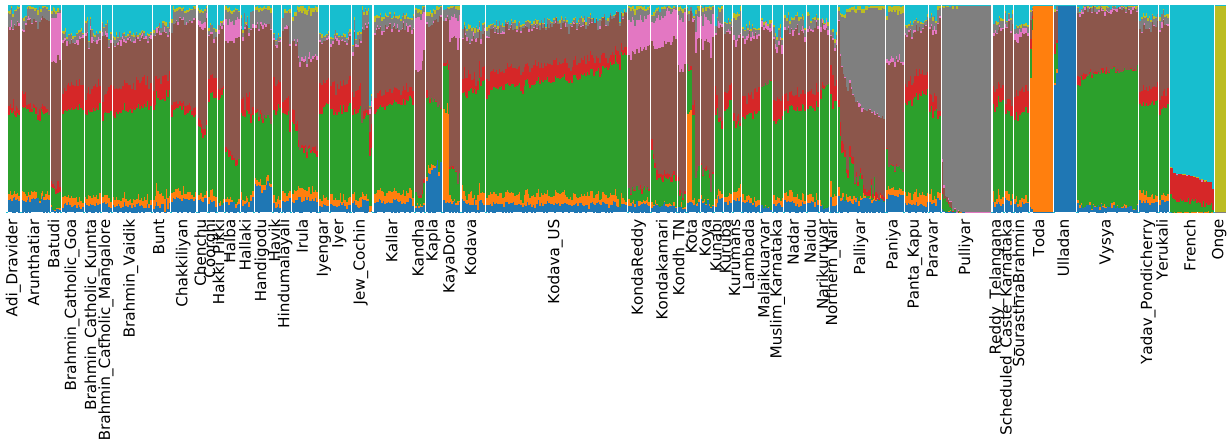

Figure 4.9: `ADMIXTURE` results for $K = 9$ across all populations in south India (see 4.6.5 for details on bounding box for south India)

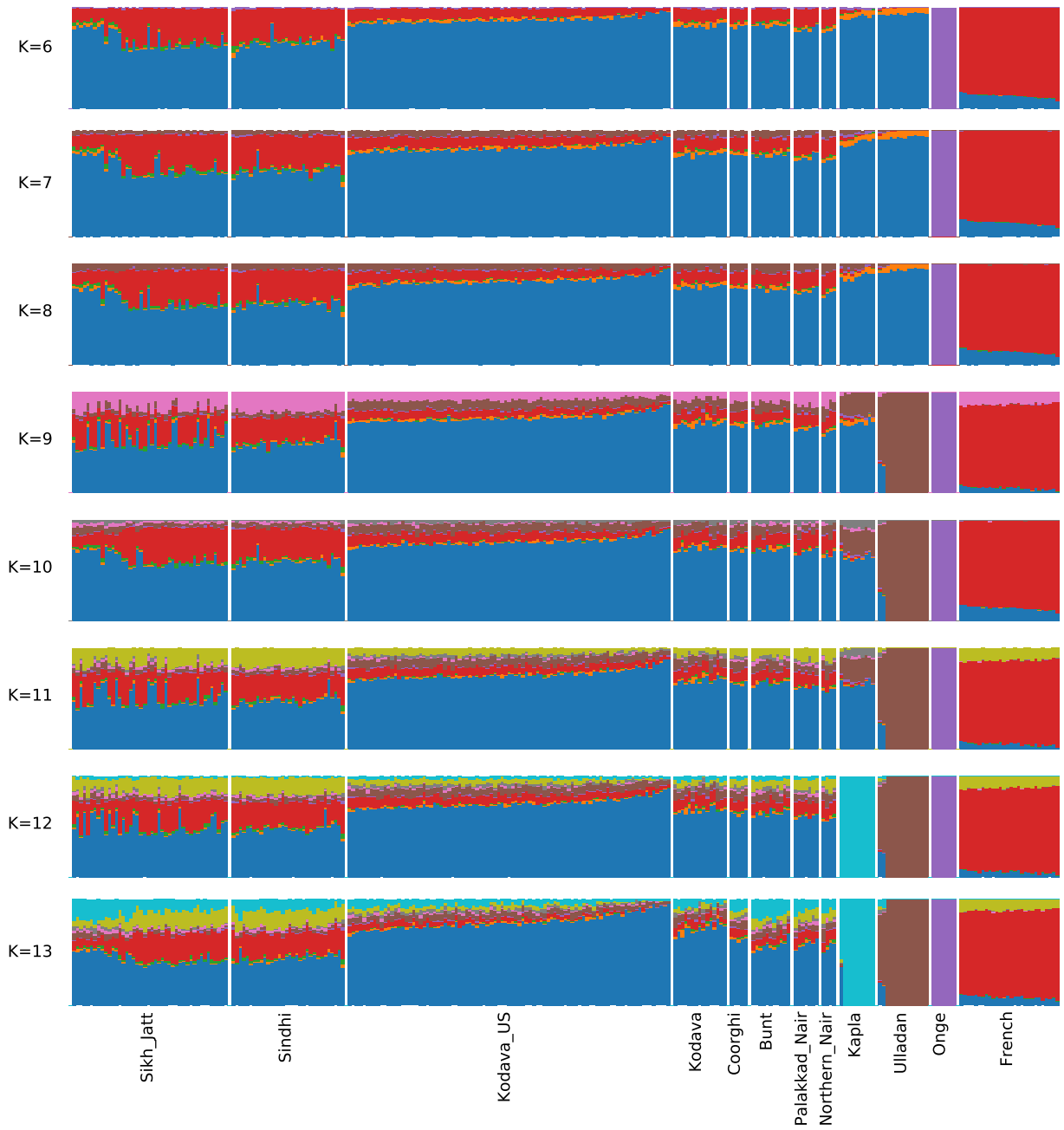

Figure 4.10: ADMIXTURE results from $K = 9$ to $K = 13$ across local south Indian populations focused on the set of samples newly typed in our dataset

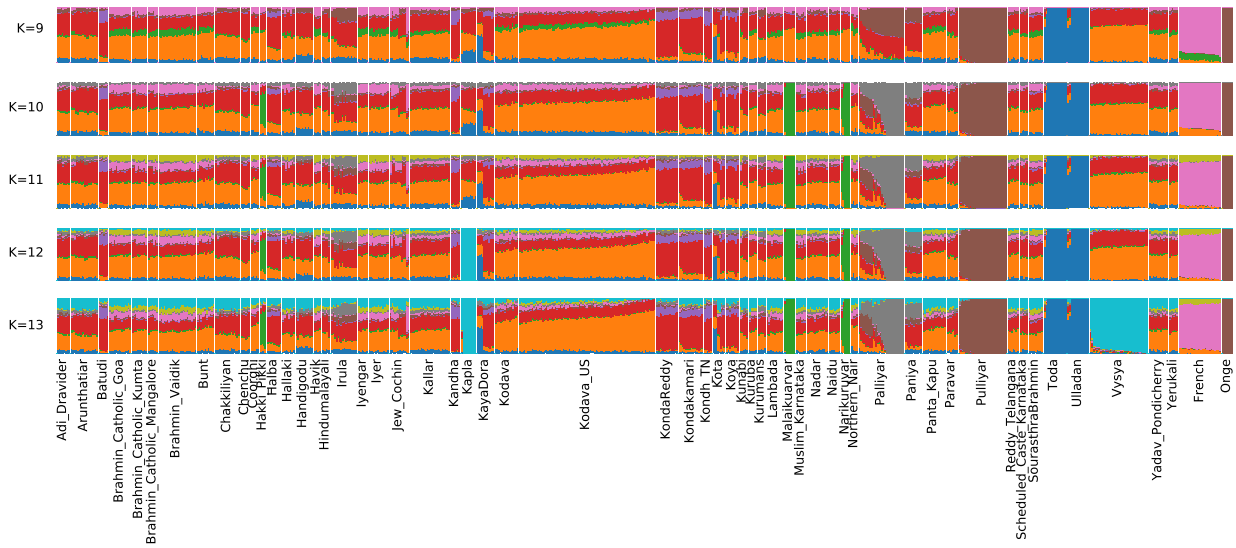

Figure 4.11: ADMIXTURE results from $K = 9$ to $K = 13$ across all local south Indian populations

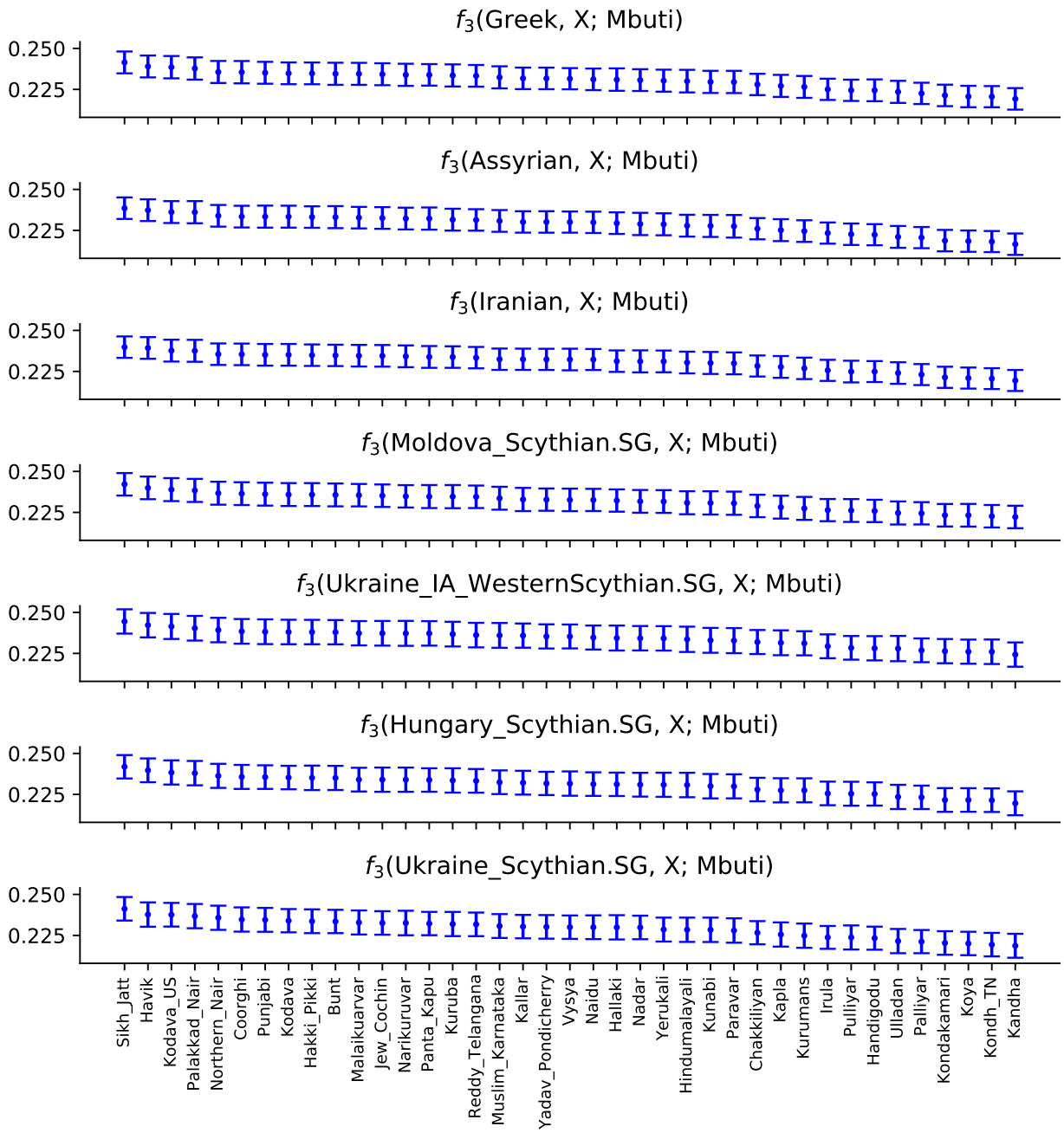

Figure 4.12: Gallery of outgroup f_3 values for potential source populations (including several ancient Scythian populations). We note that for none of these source populations, the outgroup f_3 value for the US-based Kodava cohort was significantly different from the other populations. We have also added the "Sikh_Jatt" and Punjabi populations here to check the results for Indo-European speaking populations living farther north in India.

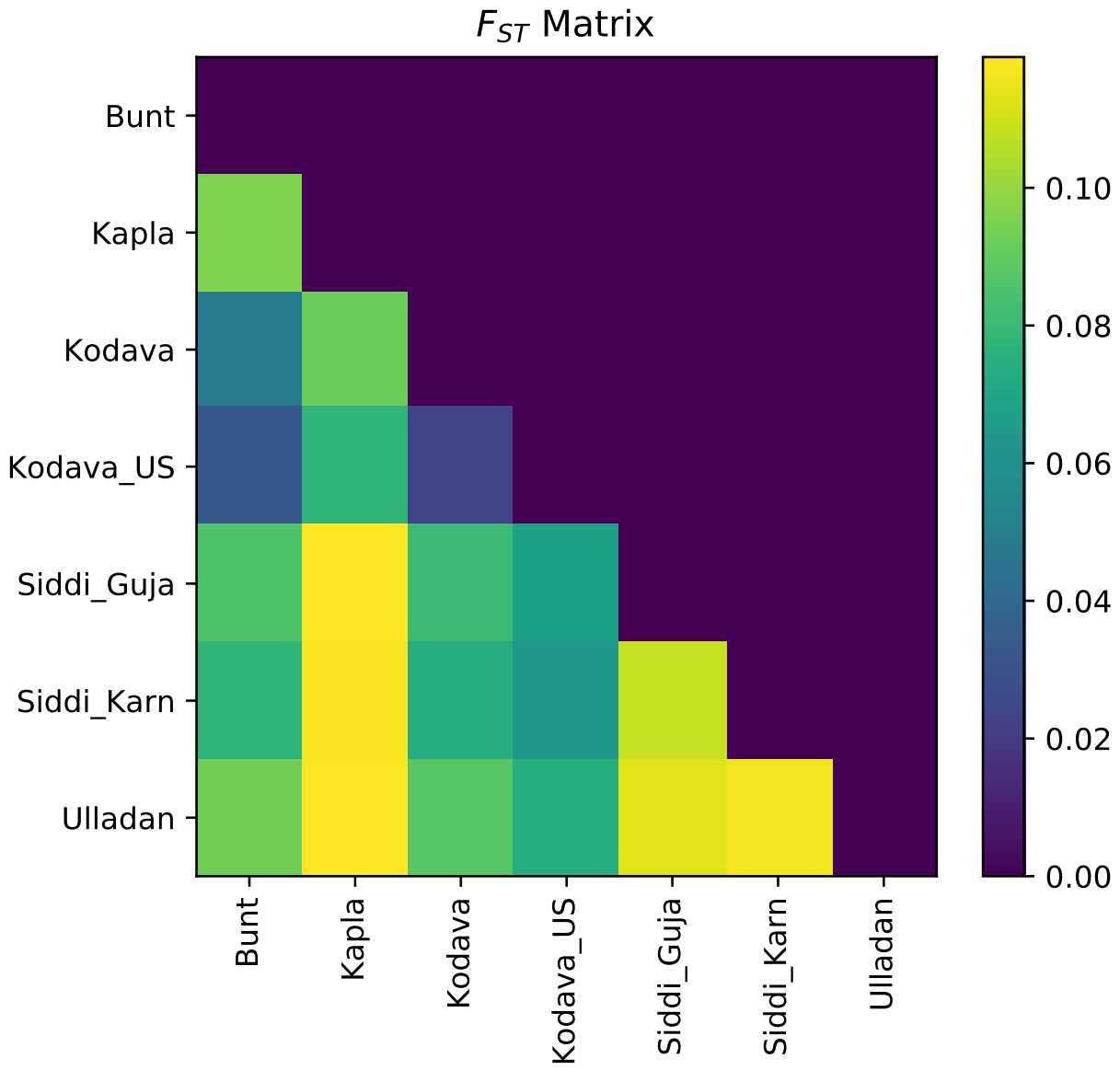

Figure 4.13: Matrix of F_{ST} values computed using ADMIXTOOLS across newly sampled populations in south India as a part of this project(Patterson et al., 2012).

4.8 Appendices

4.8.1 Kodava Population Genetics Results: Community Asked Questions

Q1. Did we test all of the theories surrounding the ancestral history of the Kodavas?

A: In our analysis we used descriptive statistics (f_3, PCA) to characterize the genetic relatedness between Kodavas, other Indian populations, and western Eurasian/Central Asian populations hypothesized to have contributed to the ancestry of present-day Kodavas. These tests serve to highlight which theories are not compatible with the levels of relatedness that we observe. The three predominant hypotheses we tested here relate to the genetic relationship of the Kodavas to the ancient Scythian culture from Central Asia, ancient Iranian farmers (modern South Asians are modeled to have Iranian farmer ancestry from the time when agriculture was introduced in the region), and modern Greeks (a proxy for Alexander the Great's army).

Q2. What did we learn that is new or in support of previous theories?

A: Our results shows that broadly Kodavas are genetically closest to neighboring populations in southern India. However, the tests we have performed cannot exclude the possibility of low-levels of non-local ancestry exclusive to the Kodava.

Q3. Could you comment on the similarity in patterns seen with our culture, DNA and language?

A: What is interesting in this case is that while the Kodava community has several unique cultural and linguistic features relative to neighboring populations in south India, the broad scale DNA evidence does not seem to indicate similar differentiation at the genetic level. We should note that cultural and genetic differentiation often occur on very different timescales. There is the possibility that more sophisticated methods may reveal more subtle structure that differentiates the Kodava community from its neighbors. Linguistically, Kodava-thak draws similarities to many other Dravidian languages (e.g. Tamil, Kannada) and therefore seems more in line with the genetic evidence as well.

Q4. How can I get access to my own data? If I do, will I get any help in interpretation like genetic counseling?

A: The goal of the project is to aggregate information at the community level and analyze

population history at the community level, rather than individual-level inferences. Considering these goals and to protect the identities of donors we do not have IRB approval to release individual-level data in order to protect the identities of donors. As a result, all samples and data have been de-identified within the study, and individuals will not be able to access their genetic data.

Q6. Did you have enough samples for the analysis?

A: We have enough samples to perform analyses like PCA and ADMIXTURE to understand the overall genetic relatedness between the Kodava and other Indian populations. For the hypothesis testing based on the outgroup f_3 statistics we would likely be able to detect more subtle amounts of non-local ancestry (and refine our conclusions) with more samples (see Section 4.6.6).

Q7. How can the community help in advancing the initiative? A: We welcome active participation from the community in a variety of ways from directly volunteering to genetic sampling, providing technical expertise or questions, and financial aid. To get involved please contact Chinnappa Dilip Kodira (dkodira@gmail.com).

Q8. Do you expect any backlashes based on the research findings?

A: We aim to provide results directly back to the community to get feedback and to present results that are not only scientifically consistent and rigorous, but also cognizant of the culture and traditions of the Kodava community.

Q9. What about the Greek theory? Is there any supporting data from your findings?

A: We have not found evidence supporting the hypothesis that the Kodava community has a substantially higher genetic affinity to the Greek populations (ancient or modern) than neighboring populations in southern India, such as the Bunt population. However, we acknowledge that the methods we have used may be under-powered to detect more subtle amounts of ancestry and more powerful methods may be required to detect potential ancestry contributions.

CHAPTER 5

CONCLUSION

The goal of this dissertation was to address some of the challenges in modeling spatially and temporally population genetic data. We describe the progress made in each of the specific sub problems and address potential future directions within the context of the problems addressed here.

In Chapter 2, we developed a visualization scheme to more prominently show the joint absolute frequencies of variants. We find particular use of these visualizations in the pedagogical setting to (1) clarify broad-scale features of human population genetic variation and (2) to avoid potential misconceptions on the extent of genetic differences between populations. However, even with modest numbers of defined populations the cumulative proportion of GeoVar categories with very low abundances is large. This highlights the complexity in visualizing high-dimensional data structures such as the joint-SFS (Gutenkunst et al., 2009; Kamm et al., 2020). Applications of recent clustering and dimensionality reduction techniques may prove to be effective successors of the visualization method we have proposed to highlight variant-centric population structure, although such techniques may also have notable drawbacks or difficulties in interpretation (Diaz-Papkovich et al., 2019)

In Chapter 3, we develop a model of the two-locus ancestral process with serial sampling. We focus on the case with two haplotypes because this simpler process is analytically tractable and yields insight into summary statistics, such as expected patterns of linkage disequilibrium (Hudson, 2001; McVean, 2002).

The two-locus genealogies describe expected patterns of linkage disequilibrium with serial samples, although we show the effect of time stratified sampling to be relatively small on the σ_d^2 metric of linkage disequilibrium. The key innovation in the case of two serially sampled haplotypes was to develop a rate matrix for the sub-process of a single modern haplotype evolving backwards in time, and using matrix exponentiation to obtain the probability

distribution on the states to "restart" the process at the ancient sampling time point.

A potential future direction for spatio-temporal inference in two-locus models would be to extend the results of Duforet-Frebourg and Slatkin (2016) on "Isolation-by-Distance-and-Time" (IBDT) to the two-locus setting. While it may not necessarily yield analytical expressions, we speculate it may be possible to obtain numerical results for this two-locus model (Hobolth et al., 2019). Directly accounting for demographic history *and* serial sampling in the two-locus process is another potential direction. To directly study the joint moments of genealogical history (e.g. the joint expectation of the tree heights at two loci, $\mathbb{E}[T_A T_B]$) under varying demographic history, it is possible to either leverage endpoint conditioned markov-chains (Hobolth and Jensen, 2011) or recently developed phase-type theory(Hobolth et al., 2019). We anticipate that these theoretical developments may lead to a richer understanding of the joint effects of migration, time separation, and recombination in shaping patterns of genetic variation at linked loci and lead to new methods for population genetic inference.

We also investigated the impact of serial sampling on the haplotype copying model of Li and Stephens (2003) to explore potential biases in the imputation of ancient human DNA. Counter-intuitively, the result that the haplotype copying jump-rate (λ) decreases when applied to European male X-chromosomes implies that the imputation accuracy of segregating common variants will likely be similar to applications with modern DNA (which is quite high (McCarthy et al., 2016, e.g.)). We observed in simultions that rapid population growth in Europe decreases the copying jump-rate as a function of sample age (Keinan and Clark, 2012; Reppell et al., 2014; Browning et al., 2018). It would be useful to investigate if similar decreases in the haplotype copying jump rate with sample age are seen in ancient DNA data from other regions of the world. This may be particularly particularly once human ancient DNA time-series are as densely sampled as in western Eurasia. In preliminary analyses of data from ancient male X-chromosomes from Africa (not shown), we found the maximum-

likelihood jump rate ($\hat{\lambda}$) to be substantially higher from modern samples. This suggests that accounting for population demographic history may impose limits on the imputation of ancient DNA at particular time-depths, in addition to potential issues of DNA sequence preservation and potential contamination. Furthermore, a principled approach to modeling the haplotype-copying model in a spatio-temporal context could directly incorporate space and time into prior transition density kernels (Yang et al., 2015; Ralph and Coop, 2013; Ringbauer et al., 2017).

Overall, the analysis of joint spatial and temporal patterns within genealogical models of linked variation is still in its infancy, due to the limited availability of ancient DNA haplotype data and limited use of such models for population genetic inference in humans. We expect that increasing data availability across both spatial and temporal dimensions will provide fruitful future opportunities for haplotype-based population genetic inference.

In Chapter 4, we explored the population genetic history of the Kodava population in south western India. Indian populations are currently under-represented in global datasets of genomic and phenotypic variation (Popejoy and Fullerton, 2016). However, there are increasing efforts to leverage the social structure and endogamy within India to learn more about the genetic basis of recessive human diseases (Nakatsuka et al., 2017; Wall et al., 2019). Spatial sampling biases are a challenge in analyzing global population genetic data, but there is a renewed effort to more widely sample genetic data across populations to develop equitable biomedical therapies (e.g. Martin et al., 2019).

We have highlighted through our study of the Kodava that genetics can also provide insight on population origin hypotheses and oral histories. These hypotheses are increasingly important to integrate with anthropological studies, leading to a broader understanding of population identity. With high linguistic and cultural diversity in India, we expect that the region will be a stage for many future investigations at the intersection of population genetics and anthropology.

Overall, the results within this dissertation have addressed specific challenges in the analysis of spatio-temporal population genetic data. By developing new data representations for visualization, extensions of classical population genetic models, and studying the population genetic history in under sampled regions of the world, we have made meaningful progress in addressing specific problems within this larger domain. As population genetic datasets continue to grow in size and density across spatial and temporal scales, further opportunities will present themselves for the development of novel population genetics theory and statistical methods.

REFERENCES

Jeffrey R. Adrion, Christopher B. Cole, Noah Dukler, Jared G. Galloway, Ariella L. Gladstein, Graham Gower, Christopher C. Kyriazis, Aaron P. Ragsdale, Georgia Tsambos, Franz Baumdicker, et al. (2020). A community-maintained standard library of population genetic models. *eLife*, 9:1–39.

Patrick K Albers and Gil McVean (2020). Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.*, 18(1):e3000586.

Anders Albrechtsen, Finn Cilius Nielsen, and Rasmus Nielsen (2010). Ascertainment biases in {SNP} chips affect measures of population divergence. *Mol. Biol. Evol.*, 27(11):2534–2547.

Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, et al. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Adam Auton and Gil McVean (2007). Recombination rate estimation in the presence of hotspots. *Genome research*, 17(8):1219–27.

Ramaswami Balakrishnan (1976). *Phonology of Kodagu with vocabulary / R. Balakrishnan*. Annamalai University Annamalainagar.

Analabha Basu, Neeta Sarkar-Roy, and Partha P. Majumder (2016). Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proceedings of the National Academy of Sciences of the United States of America*, 113(6):1594–1599.

Anders Bergström, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484).

Anders Bergström, Shane A McCarthy, Ruoyun Hui, Mohamed A Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, et al. (2019). Insights into human genetic variation and population history from 929 diverse genomes. *bioRxiv*, page 674986.

Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L Price (2013). Estimating and interpreting {FST}: the impact of rare variants. *Genome Res.*, 23(9):1514–1521.

Stephanie A Bien, Genevieve L Wojcik, Chani J Hodonsky, Christopher R Gignoux, Iona Cheng, Tara C Matise, Ulrike Peters, Eimear E Kenny, and Kari E North (2019). The

Future of Genomic Studies Must Be Globally Representative: Perspectives from {PAGE}. *Annu. Rev. Genomics Hum. Genet.*, 20:181–200.

Stephanie A Bien, Genevieve L Wojcik, Niha Zubair, Christopher R Gignoux, Alicia R Martin, Jonathan M Kocarnik, Lisa W Martin, Steven Buyske, Jeffrey Haessler, Ryan W Walker, et al. (2016). Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLoS One*, 11(12):e0167758.

Lorenzo Bomba, Klaudia Walter, and Nicole Soranzo (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.*, 18(1):77.

Gideon S Bradburd and Peter L Ralph (2019). Spatial Population Genetics: It's About Time. *Annual Review of Ecology, Evolution, and Systematics*, 50(1):427–449.

Sharon R. Browning and Brian L. Browning (2015). Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *American Journal of Human Genetics*, 97(3):404–418.

Sharon R Browning, Brian L Browning, Ying Zhou, Serena Tucci, and Joshua M Akey (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, 173(1):53—-61.e9.

Carlos D Bustamante, Esteban González Burchard, and Francisco M la Vega (2011). Genomics for the world. *Nature*, 475(7355):163–165.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

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

Census of India (2011). *District census handbook, Kodagu*, volume SERIES-12,.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7.

Hua Chen and Kun Chen (2013). Asymptotic distributions of coalescence times and ancestral lineage numbers for populations with temporally varying size. *Genetics*, 194(3):721–36.

Lu Chen, Aaron B Wolf, Wenqing Fu, Liming Li, and Joshua M Akey (2020). Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell*, 180(4):677—-687.e16.

Andrew G Clark, Melissa J Hubisz, Carlos D Bustamante, Scott H Williamson, and Rasmus Nielsen (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.*, 15(11):1496–1502.

Donald F Conrad, Mattias Jakobsson, Graham Coop, Xiaoquan Wen, Jeffrey D Wall, Noah A Rosenberg, and Jonathan K Pritchard (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.*, 38(11):1251–1260.

Jesse Dabney, Matthias Meyer, and Svante Pääbo (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, 5(7).

Chengzhen L. Dai, Mohammad M. Vazifeh, Chen Hsiang Yeang, Remi Tachet, R. Spencer Wells, Miguel G. Vilar, Mark J. Daly, Carlo Ratti, and Alicia R. Martin (2020). Population Histories of the United States Revealed through Fine-Scale Migration and Haplotype Analysis. *American Journal of Human Genetics*.

Michael DeGiorgio, Mattias Jakobsson, and Noah A Rosenberg (2009). Out of Africa: modern human origins special feature: explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci. U. S. A.*, 106(38):16057–16062.

A Di Rienzo and A C Wilson (1991). Branching pattern in the evolutionary tree for human mitochondrial {DNA}. *Proc. Natl. Acad. Sci. U. S. A.*, 88(5):1597–1601.

Kevin Dialdestoro, Jonas Andreas Sibbesen, Lasse Maretty, Jayna Raghwani, Astrid Gall, Paul Kellam, Oliver G. Pybus, Jotun Hein, and Paul A. Jenkins (2016). Coalescent inference using serially sampled, high-throughput sequencing data from intrahost HIV infection. *Genetics*, 202(4):1449–1472.

Alex Diaz-Papkovich, Luke Anderson-Trocmé, Chief Ben-Eghan, and Simon Gravel (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genetics*, 15(11):e1008432.

Brian M Donovan, Rob Semmens, Phillip Keck, Elizabeth Brimhall, K C Busch, Monica Weindling, Alex Duncan, Molly Stuhlsatz, Zoë Buck Bracey, Mark Bloom, et al. (2019). Toward a more humane genetics education: Learning about the social and quantitative complexities of human genetic variation research could reduce racial bias in adolescent and adult populations. *Science Education*, 103(3):529–560.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

Nicolas Duforet-Frebourg and Montgomery Slatkin (2016). Isolation-by-distance-and-time in a stepping-stone model. *Theoretical Population Biology*, 108:24–35.

Eric Y Durand, Nick Patterson, David Reich, and Montgomery Slatkin (2011). Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, 28(8):2239–2252.

M B Eisen, P T Spellman, P O Brown, and D Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25):14863–14868.

M B Emeneau (1967). The South Dravidian Languages. *Journal of the American Oriental Society*, 87(4):365–413.

Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.*, 48(D1):D941—-D947.

Paul Fearnhead and Peter Donnelly (2001). Estimating Recombination Rates From Population Genetic Data. *Genetics*, 159(3).

Jack N. Fenner (2005). Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *American Journal of Physical Anthropology*, 128(2):415–423.

Roald Forsberg, Alexei J. Drummond, and Jotun Hein (2005). Tree measures and the number of segregating sites in time-structured population samples. *BMC Genetics*, 6:1–14.

Andrew E. Fry, Clare J. Trafford, Martin A. Kimber, Man Suen Chan, Kirk A. Rockett, and Dominic P. Kwiatkowski (2006). Haplotype homozygosity and derived alleles in the human genome. *American Journal of Human Genetics*, 78(6):1053–1059.

Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L.F. Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare De Filippo, et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514(7253):445–449.

Cristina Gamba, Eppie R. Jones, Matthew D. Teasdale, Russell L. McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kovári, Ildikó Pap, Alexandra Anders, et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5(1):1–9.

B D Ganapathy (1967). *Kodavas (Coorgs), Their Customs and Culture*. Ms. Kodagu Ltd., Mercara.

Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi Yang Fritz, et al. (2010). A draft sequence of the neandertal genome. *Science*, 328(5979):710–722.

R. C. Griffiths (1984). Asymptotic line-of-descent distributions. *Journal of Mathematical Biology*, 21(1):67–75.

Gilles Guillot, Raphaël Leblois, Aurélie Coulon, and Alain C Frantz (2009). Statistical methods in spatial genetics. *Mol. Ecol.*, 18(23):4734–4756.

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10).

H Harpending and A Rogers (2000). Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.*, 1:361–385.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, et al. (2020). Array programming with NumPy. *Nature*.

Kelley Harris and Rasmus Nielsen (2016). The Genetic Cost of Neanderthal Introgression. *Genetics*, 203(2):881–891.

Garrett Hellenthal, George B.J. Busby, Gavin Band, James F. Wilson, Cristian Capelli, Daniel Falush, and Simon Myers (2014). A genetic atlas of human admixture history. *Science*, 343(6172):747–751.

W. G. Hill and Alan Robertson (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38(6):226–231.

Asger Hobolth and Jens Ledet Jensen (2011). Summary Statistics for Endpoint-Conditioned Continuous-Time Markov Chains. *Journal of Applied Probability*, 48(04):911–924.

Asger Hobolth and Jens Ledet Jensen (2014). Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical Population Biology*, 98:48–58.

Asger Hobolth, Arno Siri-Jégousse, and Mogens Bladt (2019). Phase-type distributions in population genetics. *Theoretical Population Biology*.

Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.*, 44(8):955–959.

Bryan N. Howie, Peter Donnelly, and Jonathan Marchini (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6):e1000529.

Richard R. Hudson (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*.

Richard R Hudson (1985). The Sampling Distribution of Linkage Disequilibrium under an Infinite Allele Model without Selection. *Genetics*, 109(3):611–631.

R R Hudson (1990). Gene genealogies and the coalescent process. volume 7, pages 1–44.

R R Hudson (2001). Two-locus sampling distributions and their application. *Genetics*, 159.

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

L A K Iyer and L A Iyer (1969). *The Coorg Tribes and Castes*. Johnson Reprint Corporation.

Ethan M. Jewett and Noah A. Rosenberg (2014). Theory and applications of a deterministic approximation to the coalescent model. *Theoretical Population Biology*, 93:14–29.

Ethan M. Jewett, Matthew Zawistowski, Noah A. Rosenberg, and Sebastian Zöllner (2012). A coalescent model for genotype imputation. *Genetics*, 191(4):1239–1255.

Julien Jouganous, Will Long, Aaron P Ragsdale, and Simon Gravel (2017). Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*.

Ivan Juric, Simon Aeschbacher, and Graham Coop (2016). The Strength of Selection against Neanderthal Introgression. *PLoS Genet.*, 12(11):e1006340.

S Kamat (1993). *Karnataka State Gazetteers: Kodagu District*. Office of the Chief Editor, Karnataka Gazetteer.

Jack Kamm, Jonathan Terhorst, Richard Durbin, and Yun S. Song (2020). Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. *Journal of the American Statistical Association*, 115(531):1472–1487.

John A. Kamm, Jeffrey P. Spence, Jeffrey Chan, and Yun S. Song (2016). Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation. *Genetics*, 203(3):1381–1399.

Alon Keinan and Andrew G Clark (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–3.

Jerome Kelleher, Alison M Etheridge, and Gilean McVean (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12(5):1–22.

Jerome Kelleher, Yan Wong, Anthony W Wohns, Chaimaa Fadil, Patrick K Albers, and Gil McVean (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

Felix M Key, Muslihudeen A Abdul-Aziz, Roger Mundry, Benjamin M Peter, Aarthi Sekar, Mauro D'Amato, Megan Y Dennis, Joshua M Schmidt, and Aida M Andrés (2018). Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLOS Genetics*, 14(5):1–22.

Adam Kiezun, Sara L. Pulit, Laurent C. Francioli, Freerk van Dijk, Morris Swertz, Dorret I. Boomsma, Cornelia M. van Duijn, P. Eline Slagboom, G. J B van Ommen, Cisca Wijmenga, et al. (2013). Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency. *PLoS Genetics*, 9(2):1–12.

M Kimura and T Ohta (1973). The age of a neutral mutant persisting in a finite population. *Genetics*, 75(1):199–212.

J F C Kingman (1982). On the genealogy of large populations. *J Appl Prob*, 19A.

Augustine Kong, Gudmar Thorleifsson, Daniel F. Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G. Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103.

Bh. Krishnamurti (1985). An Overview of Comparative Dravidian Studies since "Current Trends" 5 (1969). *Oceanic Linguistics Special Publications*, 5(20):212–231.

Mookonda Kushalappa (2013). *The Early Coorgs*. Notion Press.

Mookonda Kushalappa (2018). A tribe nestled in the hills of Kodagu. *Deccan Herald*.

Joseph Lachance and Sarah A Tishkoff (2013). {SNP} ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays*, 35(9):780–786.

Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush (2012). Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):11–17.

Daniel J Lawson, Lucy van Dorp, and Daniel Falush (2018). A tutorial on how not to over-interpret {STRUCTURE} and {ADMIXTURE} bar plots. *Nat. Commun.*, 9(1):3258.

Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413.

Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C Royrvik, Barry Cunliffe, Wellcome Trust Case Control Consortium 2, et al. (2015). The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314.

R C Lewontin (1972). The Apportionment of Human Diversity. pages 381–398.

Heng Li (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.

Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, and Richard M Myers (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, 319(5866):1100–1104.

Na Li and Matthew Stephens (2003). Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233.

Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, and Joseph K Pickrell (2013). Inferring Admixture Histories of Human Populations. *Genetics*, 193(April):1233–1254.

Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, 48:811.

Gerton Lunter (2019). Haplotype matching in large cohorts using the Li and Stephens model. *Bioinformatics*, 35(5):798–806.

Michael Lynch, Sen Xu, Takahiro Maruki, Xiaoqian Jiang, Peter Pfaffelhuber, and Bernhard Haubold (2014). Genome-wide linkage-disequilibrium profiles from single individuals. *Genetics*, 198(1):269–281.

Partha P Majumder (2010). The Human Genetic History of South Asia. *Current Biology*, 20:R184–R187.

Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206.

Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei Min Chen (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873.

Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Urko M Marigorta and Arcadi Navarro (2013). High trans-ethnic replicability of {GWAS} results implies common causal variants. *PLoS Genet.*, 9(6):e1003566.

Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.*, 100(4):635–649.

Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*.

Rui Martiniano, Lara M. Cassidy, Ros Ó'Maoldúin, Russell McLaughlin, Nuno M. Silva, Licinio Manco, Daniel Fidalgo, Tania Pereira, Maria J. Coelho, Miguel Serra, et al. (2017). The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genetics*, 13(7).

Sarabjit S. Mastana (2014). Unity in diversity: An overview of the genomic anthropology of India. *Annals of human biology*, 41:287–99.

Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, et al. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503.

Iain Mathieson and Gil McVean (2014). Demography and the age of rare variants. *PLoS Genet.*, 10(8):e1004528.

Iain Mathieson and Aylwyn Scally (2020). What is ancestry? *PLoS Genet.*, 16(3):e1008624.

Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R. Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283.

Gil McVean (2009). A genealogical interpretation of principal components analysis. *PLoS Genet.*, 5(10):e1000686.

Gilean A T McVean (2002). A genealogical interpretation of linkage disequilibrium. *Genetics*, 162(2):987–991.

Gilean A T McVean, Simon R Myers, Sarah Hunt, Panos Deloukas, David R Bentley, and Peter Donnelly (2004). The fine-scale structure of recombination rate variation in the human genome. *Science (New York, N.Y.)*, 304(5670):581–4.

Mait Metspalu, Irene Gallego Romero, Bayazit Yunusbayev, Gyaneshwer Chaubey, Chandana Basu Mallick, Georgi Hudjashov, Mari Nelis, Reedik Mägi, Ene Metspalu, Maido Remm, et al. (2011). Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *American Journal of Human Genetics*, 89(6):731–744.

Priya Moorjani, Sriram Sankararaman, Qiaomei Fu, Molly Przeworski, Nick Patterson, and David Reich (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*, pages 1514696113–.

Priya Moorjani, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh (2013). Genetic evidence for recent population mixture in India. *American Journal of Human Genetics*, 93(3):422–438.

Nathan Nakatsuka, Priya Moorjani, Niraj Rai, Biswanath Sarkar, Arti Tandon, Nick Patterson, Gandham Srilakshmi Bhavani, Katta Mohan Girisha, Mohammed S. Mustak, Sudha Srinivasan, et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. *Nature Genetics*, 49(9):1403–1407.

Ankita Narang, Pankaj Jha, Vimal Rawat, Arijit Mukhopadhayay, Debasis Dash, Analabha Basu, and Mitali Mukerji (2011). Recent admixture in an Indian population of african ancestry. *American Journal of Human Genetics*.

Vagheesh M. Narasimhan, Karen A. Hunt, Dan Mason, Christopher L. Baker, Konrad J. Karczewski, Michael R. Barnes, Anthony H. Barnett, Chris Bates, Srikanth Bellary, Nicholas A. Bockett, et al. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science*.

Vagheesh M. Narasimhan, Nick Patterson, Priya Moorjani, Nadin Rohland, Rebecca Bernardos, Swapan Mallick, Iosif Lazaridis, Nathan Nakatsuka, Iñigo Olalde, Mark Lipson, et al. (2019). The formation of human populations in South and Central Asia. *Science*, 365(6457).

Matthew R Nelson, Daniel Wegmann, Margaret G Ehm, Darren Kessner, Pamela St Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science (New York, N.Y.)*, 337(6090):100–104.

Rasmus Nielsen, Joshua M Akey, Mattias Jakobsson, Jonathan K Pritchard, Sarah Tishkoff, and Eske Willerslev (2017). Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310.

John Novembre and Benjamin M Peter (2016). Recent advances in the study of fine-scale population structure in humans. *Current Opinion in Genetics & Development*, 41:98–105.

Iñigo Olalde and Cosimo Posth (2020). Latest trends in archaeogenetic research of west Eurasians. *Current Opinion in Genetics and Development*, 62:36–43.

Diego Ortega-Del Vecchyo and Montgomery Slatkin (2018). FST between archaic and present-day samples. *Heredity*, page 1.

Luca Pagani, Daniel John Lawson, Evelyn Jagoda, Alexander Mörseburg, Anders Eriksson, Mario Mitt, Florian Clemente, Georgi Hudjashov, Michael Degiorgio, Lehti Saag, et al. (2016). Genomic analyses inform on migration events during the peopling of Eurasia. *Nature*.

Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er (2012). Length distributions of identity by descent reveal fine-scale demographic history. *American Journal of Human Genetics*, 91(5):809–822.

Aaron Panofsky and Catherine Bliss (2017). Ambiguity and scientific authority: population classification in genomic science. *Am. Sociol. Rev.*, 82(1):59–87.

Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich (2012). Ancient admixture in human history. *Genetics*, 192(3):1065–1093.

Nick Patterson, Alkes L Price, and David Reich (2006). Population structure and eigenanalysis. *PLoS Genet.*, 2(12):e190.

Joshua S. Paul, Matthias Steinrücken, and Yun S. Song (2011). An accurate sequentially markov conditional sampling distribution for the coalescent with recombination. *Genetics*, 187(4):1115–1128.

Leena Peltonen, Aarno Palotie, and Kenneth Lange (2000). Use of population isolates for mapping complex traits. *Nature Reviews Genetics*.

Trevor J. Pemberton, Fang Yuan Li, Erin K. Hanson, Niyati U. Mehta, Sunju Choi, Jack Ballantyne, John W. Belmont, Noah A. Rosenberg, Chris Tyler-Smith, and Pragna I. Patel (2012). Impact of restricted marital practices on genetic variation in an endogamous Gujarati group. *American Journal of Physical Anthropology*, 149(1):92–103.

Roseann E Peterson, Karoline Kuchenbaecker, Raymond K Walters, Chia-Yen Chen, Alice B Popejoy, Sathish Periyasamy, Max Lam, Conrad Iyegbe, Rona J Strawbridge, Leslie Brick, et al. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, 179(3):589–603.

C Phillips, J Amigo, A O Tillmar, M A Peck, M de la Puente, J Ruiz-Ram\'\irez, F Bittner, Š Idrizbegović, Y Wang, T J Parsons, and M V Lareu (2020). A compilation of tri-allelic {SNPs} from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. *Forensic Sci. Int. Genet.*, 46:102232.

Joseph K Pickrell and Jonathan K Pritchard (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics*, 8(11):1–17.

Joseph K Pickrell and David Reich (2014). Toward a new history and geography of human genes informed by ancient {DNA}. *Trends Genet.*, 30(9):377–389.

Alexander Platt, Alyssa Pivirotto, Jared Knoblauch, and Jody Hey (2019). An estimator of first coalescent time reveals selection on young variants and large heterogeneity in rare allele ages among human populations. *PLoS Genet.*, 15(8):e1008340.

K C Ponnappa (1999). *A Study of the Origins of Coorgs*. K. C. Ponappa.

Alice B. Popejoy and Stephanie M. Fullerton (2016). Genomics is failing on diversity. *Nature*.

Ryan Poplin, Valentin Ruano-Rubio, Mark DePristo, Tim Fennell, Mauricio Carneiro, Geraldine Van der Auwera, David Kling, Laura Gauthier, Ami Levy-Moonshine, David Roazen, et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, page 201178.

Alkes L. Price, Arti Tandon, Nick Patterson, Kathleen C. Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H. Beaty, Rasika Mathias, David Reich, and Simon Myers (2009). Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genetics*, 5(6):e1000519.

Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare De Filippo, et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.

Franck Prugnolle, Andrea Manica, Marie Charpentier, Jean François Guégan, Vanina Guernier, and François Balloux (2005). Pathogen-driven selection and worldwide {HLA} class {I} diversity. *Curr. Biol.*, 15(11):1022–1027.

R. C. Lewontin and Kenichi Kojima (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14(4):458–472.

L.R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Race, Ethnicity, and Genetics Working Group (2005). The use of racial, ethnic, and ancestral categories in human genetics research. *Am. J. Hum. Genet.*, 77(4):519–532.

Fernando Racimo, Sriram Sankararaman, Rasmus Nielsen, and Emilia Huerta-Sánchez (2015). Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371.

Maanasa Raghavan, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, Thomas W. Stafford, Ludovic Orlando, Ene Metspalu, et al. (2014). Upper palaeolithic Siberian genome reveals dual ancestry of native Americans. *Nature*, 505(7481):87–91.

Aaron P. Ragsdale and Simon Gravel (2019). Models of archaic admixture and recent history from two-locus statistics. *PLoS Genetics*, 15(6):e1008204.

Peter Ralph and Graham Coop (2013). The geography of recent genetic ancestry across Europe. *PLoS Biol.*, 11(5):e1001555.

Sohini Ramachandran, Omkar Deshpande, Charles C Roseman, Noah A Rosenberg, Marcus W Feldman, and L Luca Cavalli-Sforza (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci. U. S. A.*, 102(44):15942–15947.

David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh (2009). Reconstructing Indian population history. *Nature*, 461(7263):489–494.

David E. Reich, Michele Cargili, Stacey Boik, James Ireland, Pardis C. Sabeti, Daniel J. Richter, Thomas Lavery, Rose Kouyoumjian, Shelli F. Farhadian, Ryk Ward, and Eric S. Lander (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Mark Reppell, Michael Boehnke, and Sebastian Zöllner (2014). The Impact of Accelerating Faster than Exponential Population Growth on Genetic Variation. *Genetics*, 196(3):819–828.

Richard Durrett (2002). *Probability Models for DNA Sequence Evolution*.

Harald Ringbauer, Graham Coop, and Nicholas H. Barton (2017). Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351.

Allen G Rodrigo and Joseph Felsenstein (1999). Coalescent approaches to HIV population genetics. The Johns Hopkins University Press.

Alan R. Rogers and Chad Huff (2009). Linkage disequilibrium between loci with unknown phase. *Genetics*, 182(3):839–844.

Noah A Rosenberg (2011). A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum. Biol.*, 83(6):659–684.

S. Rubinacci, D.M. Ribeiro, R. Hofmeister, and O. Delaneau (2020). Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *bioRxiv*.

Chiara Sabatti and Neil Risch (2002). Homozygosity and Linkage Disequilibrium. *Genetics*, 160(4):1707 LP – 1719.

Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9(1):88.

Anish M. Shah, Rakesh Tamang, Priya Moorjani, Deepa Selvi Rani, Periyasamy Govindaraj, Gururaj Kulkarni, Tanmoy Bhattacharya, Mohammed S. Mustak, L. V.K.S. Bhaskar, Alla G. Reddy, et al. (2011). Indian siddis: African descendants with Indian admixture. *American Journal of Human Genetics*.

Katy L. Simonsen and Gary A. Churchill (1997). A Markov Chain Model of Coalescence with Recombination. *Theoretical Population Biology*, 52(1):43–59.

Sridhar Sivasubbu and Vinod Scaria (2019). Genomics of rare genetic diseases-experiences from India. *Human genomics*.

Pontus Skoglund and Iain Mathieson (2018). Ancient genomics of modern humans: The first decade. *Annual Review of Genomics and Human Genetics*, 19:381–404.

Montgomery Slatkin (1991). Inbreeding coefficients and coalescence times. *Genetical Research*, 58(2):167–175.

Montgomery Slatkin (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future.

M Slatkin and R R Hudson (1991). Pairwise comparisons of mitochondrial {DNA} sequences in stable and exponentially growing populations. *Genetics*, 129(2):555–562.

Montgomery Slatkin and Fernando Racimo (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23):6380–6387.

Joel Smith, Graham Coop, Matthew Stephens, and John Novembre (2018). Estimating time to the common ancestor for a beneficial allele. *Molecular Biology and Evolution*.

Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321–1329.

Chris C. A. Spencer, Zhan Su, Peter Donnelly, and Jonathan Marchini (2009). Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genetics*, 5(5):e1000477.

Matthias Steinrücken, Joshua S. Paul, and Yun S. Song (2013). A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical Population Biology*, 87(1):51–61.

C B Stringer and P Andrews (1988). Genetic and fossil evidence for the origin of modern humans. *Science*, 239(4845):1263–1268.

Jacob A. Tennessen, Abigail W. Bigham, Timothy D. O'Connor, Wenqing Fu, Eimear E. Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 336(6090):64–69.

Jonathan Terhorst, Christian Schlötterer, and Yun S. Song (2015). Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. *PLOS Genetics*, 11(4):e1005069.

R Thomson, J K Pritchard, P Shen, P J Oefner, and M W Feldman (2000). Recent common ancestry of human {Y} chromosomes: evidence from {DNA} sequence data. *Proc. Natl. Acad. Sci. U. S. A.*, 97(13):7360–7365.

Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. (2013). From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43(SUPL.43):11.10.1–11.10.33.

Jenna M. VanLiere and Noah A. Rosenberg (2008). Mathematical properties of the r 2 measure of linkage disequilibrium. *Theor Popul Biol.*, 74(1):130–137.

Mário Vicente and Carina M. Schlebusch (2020). African population history: an ancient DNA perspective. *Current Opinion in Genetics and Development*, 62:8–15.

Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po Ru Loh, Gaurav Bhatia, Ron Do, et al. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *American Journal of Human Genetics*, 97(4):576–592.

John Wakeley (2009). *Coalescent Theory: An Introduction*. Coalescent theory: an introduction. Roberts & Company Publishers, Greenwood Village.

John Wakeley and Sabin Lessard (2003). Theory of the effects of population structure and sampling on patterns of linkage disequilibrium applied to genomic data from humans. *Genetics*, 164(3):1043–1053.

Jeffrey D. Wall, Eric W. Stawiski, Aakrosh Ratan, Hie Lim Kim, Changhoon Kim, Ravi Gupta, Kushal Suryamohan, Elena S. Gusareva, Rikky Wenang Purbojati, Tushar Bhangale, et al. (2019). The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, 576(7785):106–111.

Samuel K. Wasser, Andrew M. Shedlock, Kenine Comstock, Elaine A. Ostrander, Benezeth Mutayoba, and Matthew Stephen (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences of the United States of America*.

D J Witherspoon, S Wooding, A R Rogers, E E Marchani, W S Watkins, M A Batzer, and L B Jorde (2007). Genetic similarities within and between human populations. *Genetics*, 176(1):351–359.

Genevieve L Wojcik, Mariaelisa Graff, Katherine K Nishimura, Ran Tao, Jeffrey Haessler, Christopher R Gignoux, Heather M Highland, Yesha M Patel, Elena P Sorokin, Christy L Avery, et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518.

Aaron B Wolf and Joshua M Akey (2018). Outstanding questions in the study of archaic hominin admixture. *PLoS Genet.*, 14(5):e1007349.

M H Wolpoff, X Z Wu, and A G Thorne (1984). Modern Homo sapiens origins : a general theory of hominid evolution involving the fossil evidence from East Asia. pages 411–483.

Naomi R. Wray (2005). Allele Frequencies and the r 2 Measure of Linkage Disequilibrium: Impact on Design and Interpretation of Association Studies . *Twin Research and Human Genetics*, 8(2):87–94.

Wen Yun Yang, Farhad Hormozdiari, Eleazar Eskin, and Bogdan Pasaniuc (2015). A spatial haplotype copying model with applications to genotype imputation. *Journal of Computational Biology*, 22(5):451–462.

X Yi, Y Liang, E Huerta-Sanchez, X Jin, Z X Cuo, J E Pool, X Xu, H Jiang, N Vinckenbosch, T S Korneliussen, et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*, 329(5987):75–78.

Ying Zhou, Sharon R. Browning, and Brian L. Browning (2020). A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *American Journal of Human Genetics*, 106(4):426–437.