THE UNIVERSITY OF CHICAGO

INFERRING INTERPRETABLE REPRESENTATIONS OF POPULATION STRUCTURE

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
JOSEPH HENRY MARCUS

CHICAGO, ILLINOIS
DECEMBER 2020

This dissertation is dedicated to my grandmother Harriet Rochlin and uncle Mitch Fink who passed away during my PhD.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

It would not have been possible to complete my PhD without the enduring support, advice, and mentorship of my family, friends, and colleagues. First and foremost, I would like to thank my partner Cassie who has been a constant source of love and comfort throughout both the ups downs of my PhD. Cassie has always been there for me and I am deeply grateful for her support throughout our relationship. I would like to thank my parents Davida and Fred. I am incredibly grateful for their encouragement, love and guidance which has been a great source of inspiration for me throughout my life and PhD. I would like to thank my brother, Jake, for his support and love as well as for being a role model of intellectual curiosity. I've appreciated our countless discussions about our personal lives, science, and statistics and machine learning. I would like to thank my grandparents and extended family who have provided so much love and support throughout my life. I would like to particularly thank my grandmother Harriet, who passed away during my PhD. I miss our discussions about life and her wisdom and enthusiasm. I aspire to have a long life full of family, friendship, and fulfilling and creative work like she did.

My scientific career began at the University of Washington in Seattle and I am deeply grateful to all of my friends and mentors who helped cultivate my interests and who kick-started my trajectory into science. In terms of scientific mentorship at the UW, I particularly want to thank Ben Kerr, Peter Conlin, Josh Nahum, Brian Connelly, Katie Dickinson, Eli Wheat, Doug Ewing, Keith Possee, Tim Billo, Ursula Valdez, Joe Felsenstein, Mary-Claire King, Jay Shendure, Evan Eichler, M.K. Raghuraman, Ben Wiggins, and many more. I particularly want to acknowledge my amazing group of friends in Los Angeles, Chicago, Sardinia and Seattle that provided on-going support and community.

One of the greatest sources of joy and pride I take from my PhD experience are the interactions with the people I have been lucky to meet and work with along the way. I am truly grateful for my advisor John Novembre. John has always believed in me and has provided so many opportunities for me to develop as a scientist, leader and person. I deeply admire John's zest for life, confidence, and diverse interests and talents. Thank you John for your constant enthusiasm and encouragement, flexibility and mentorship. I would also like to thank my committee members Rina Foygel Barber, Anna Di Rienzo, and Matthew Stephens. Thank you Rina for being an incredible teacher and mentor. Rina's statistical insights and general positivity cultivated a great working environment in our collaboration. Thank you Anna for all the honest guidance and feedback you have provided while navigating my PhD. I would like to particularly acknowledge Matthew for all of your encouragement and mentorship. I have learned so much by your approach to science and statistics and have a great admiration for your integrity and work.

I also want to thank the Novembre lab, He lab, Stephens lab, and the whole fourth floor computational biology community. I have learned from my amazingly talented colleagues and friends in this open and collaborative environment on the fourth floor and in the Department of Human Genetics more broadly. I want thank my close collaborators Hussein Al-Asadi, Wooseok Ha, Harald Ringbauer, and Jason Willwerscheid. I learned a great deal working closely with all of you and I'm very grateful for the interaction and collaboration we had. I particularly want to acknowledge my friendship with Hussein and for all our discussions and musings on life and science. Hussein had a huge impact on my trajectory of work during my PhD and overall career. I also want to thank Peter Carbonetto for many helpful conversations about computing, statistical modeling and optimization.

I want to acknowledge the amazing group of scientists at the IRGB in Sardinia, whom I had the great pleasure of working with. I particularly want to thank them for hosting me in Sardinia in the Summer of 2015. I really appreciate the friendships I made and experiences I had during that time. I would also like to thank Sriram Sankararaman and the whole UCLA computational genomics group who hosted me while spending time in Los Angeles during the later stages of my PhD. Finally, I would like to thank Sue Levison, Candice Lewis and all the amazing people who have helped along the way in in the Department of Human Genetics and University of Chicago more broadly.

I have learned and accomplished many things throughout my PhD, but what I value most are the many friendships I have made during my time in Chicago. I feel lucky and grateful to have spent this unique time period with you all and I will cherish the many great life experiences I shared with you. Thank you.

ABSTRACT

Inferring population structure is important for several applications in medical and population genetic studies. However, the output of population structure inference methods can often be challenging to interpret. The goal of this dissertation is to apply population structure inference tools to learn and visualize demographic history and develop statistical methods for interpretable population structure inference. In Chapter 2, I apply population structure inference tools to learn about the genetic history of the Mediterranean island of Sardinia using a new ancient DNA dataset. In Chapter 3, I develop a fast and flexible statistical method and optimization algorithm for inferring and visualizing non-homogeneous patterns of migration using spatially indexed population genetic data. Finally, in Chapter 4, I develop a new Bayesian matrix factorization method and variational inference algorithm for emphasizing shared evolutionary histories when representing population structure. Overall, the work presented in this dissertation aims to provide interpretable representations of population structure which, in turn, give understanding into the underlying demographic factors that shape patterns of genetic variation.

CHAPTER 1

INTRODUCTION

The focus of this dissertation is on the inference of representations of population structure from genetic variation data. In this dissertation I apply population genetic inference tools as well as develop new statistical models and optimization algorithms for population genetics researchers to use in their own empirical applications. Today, population genetics researchers are often faced with a high-dimensional genotype matrix with rows representing samples and columns representing genetic variants typed along the genome. The elements of this matrix encode genotypes which count the number of copies a sample carries, often at a Single Nucleotide Polymorphism (SNP), of some arbitrarily pre-defined allele, like the minor allele. Sometimes side information is present, like grand-parental birth locations, geographic coordinates or temporal periods (in the case of ancient DNA) which can be helpful for interpreting the output of population genetics tools or can even be incorporated directly into statistical models (Pickrell and Reich, 2014; Novembre et al., 2008; Bradburd et al., 2018). In many, if not all, datasets it is rare that the samples are independent of one-another because they share common ancestry which induces genetic relatedness, resulting in correlations in the genotypes even for distant relatives (Speed and Balding, 2015). To put it simply, the goal of inferring population structure is to infer these genetic relationships among the samples (Novembre and Peter, 2016).

Some of the earliest work in understanding population structure focused on developing simple mathematical frameworks for modeling evolution in a spatial context (Wright, 1940, 1943). In one such framework allele frequencies are imagined to evolve in a continuous spatial habitat with local individual level dispersal (Wright, 1940, 1943, 1946). One fundamental prediction under this model is that the correlation in allele frequencies between sub-populations will decay exponentially as they become more geographically distant (Malécot, 1948). This phenomenon is often referred to as “isolation-by-distance” (IBD) and has been shown to be a pervasive feature in spatial population genetic data across many different species (Dobzhansky and Wright, 1943; Meirmans, 2012). Indeed, it is no surprise that the exponential decay of the correlation of a measurement across space is also an useful prediction in many types of data beyond genetics (Cressie, 1992; Rasmussen, 2003). Since the initial proposal of models of IBD, population geneticists have developed a rich body of theory to understand the underlying mechanisms of spatial genetic processes and make more accurate predictions of patterns of genetic differentiation in structured populations.

One fundamental outcome of this subsequent work is known as the "Stepping Stone Model" (SSM) (Kimura, 1953). As contrasted to the early conceptions of IBD, the SSM describes the evolution of allele frequencies in a discrete spatial habitat, which can be conveniently modeled as a graph, specifically an infinite lattice. The SSM defines an allele frequency in every sub-population, represented at each node (colloquially referred to as demes), and models the exchange of alleles between connected sub-populations. In the simplest form of the model, the amount of exchange is parameterized through a single migration rate between neighboring sub-populations. As the level of migration is increased the meta-population becomes more panmictic while as it is decreased the meta-population becomes more structured i.e. the sub-populations are more independent. Interestingly, as the density of sub-populations approaches an infinite limit, the SSM is equivalent to the continuous spatial models of IBD (Kimura and Weiss, 1964). The SSM provided a simple conceptual framework for deriving expectations of genetic distances, such as F_{st}, in a tractable spatial model (Slatkin, 1993). Extensions of the SSM used a finite habitat with more flexible assumptions about migration rates between sub-populations. These so called, "Migration Matrix Models", allowed the use of estimated migration matrices from pedigree data to make predictions about genetic differentiation in, for example, human populations (Bodmer and Cavalli-Sforza, 1968). Later, with the ability to directly measure sequence variation in multiple populations and species (Kreitman, 1983), there became an expanded need for tools to fit stepping stone inspired models to real data.

The development of the coalescent provided a natural means to perform demographic inference from population genetic samples (Kingman, 1982; Hudson et al., 1990; Wakeley, 2009). Under coalescent-based approaches to modeling structured populations, a demographic model, typically specifying migration rates and population sizes, determine the probabilities of particular gene genealogies (Hudson et al., 1990). In turn, these genealogies structure patterns of genetic variation in a population genetic sample, conditionally independent of the demographic history. The challenge of performing inference is thus to marginalize over the set of possible gene-genealogies, directly linking the demographic history to observed genetic variation. Even for a small number of population genetic samples integrating over latent genealogies is difficult (Rasmussen et al., 2014; Li and Durbin, 2011; Schiffels and Durbin, 2014; Palacios et al., 2015). A number of methods used approximate approaches to marginalize out gene-genealogies, such as Markov Chain Monte Carlo (MCMC), and estimate migration parameters under structured coalescent models. Most notably, MIGRATE is one such example that has been widely used to estimate effective population sizes, N_e, and migration rates, m (Beerli and Felsenstein, 2001). While these approaches provide an elegant means of inferring the parameters of population genetic models from data, they are not scalable for large numbers of samples (Novembre and Peter, 2016).

With the advent of large-scale genotyping array datasets with genotypes of thousands of individuals measured at hundreds of thousands SNPs (e.g. Nelson et al., 2008; Cavalli-Sforza, 2005), more generic and scalable approaches to estimating population structure became necessary. Principal Components Analysis (PCA) and the Pritchard, Stephens, and Donnelly (PSD) model, also known as the STRUCTURE model, have been particularly impactful in these large-scale datasets (Price et al., 2006; Novembre et al., 2008; Patterson et al., 2006; Pritchard et al., 2000; Alexander et al., 2009). The goal of both approaches is to find a compact representation of a high-dimensional genotype matrix that allows researchers to compactly summarize similarities amongst individuals in a sample (Engelhardt and Stephens, 2010). Both of these approaches can be conceptualized as fitting a linear combination of a reduced set of latent variables to model an individual's expected genotype (Engelhardt and Stephens, 2010). The PSD model assumes the expected value of an individual's genotype is a convex combination of latent allele frequencies (Pritchard et al., 2000; Gopalan et al., 2016; Cabreros and Storey, 2019; Alexander et al., 2009; Tang et al., 2005). The PSD model can be interpreted in a simple admixture model where the global ancestry of an individual's genome is made up as a mixture of ancestries from K different unobserved source populations. PCA makes no assumptions about the sign or convexity of the latent variables but assumes they are orthogonal to each other, conveniently allowing for the use of the singular value decomposition (SVD) to find the optimal solution (Bishop, 2006).

The behaviors of PCA and the PSD model have been studied with regards to how they behave with data generated from more elaborate population genetic models (Novembre and Stephens, 2008; McVean, 2009; Lawson et al., 2018). In a purely spatial model, PCA will produce sinusoidal patterns in the individual PC scores which had previously been interpreted as specific signatures of directional migration events, without recognition they are a regular outcome when an underlying homogeneous spatial process generates the data (Novembre and Stephens, 2008). Interpretation of the PSD model is also difficult in a spatial context (François et al., 2006; Bradburd et al., 2018; François et al., 2019). When the data is drawn from a homogenous spatial process the PSD model will find a continuous gradient of ancestries going from each corner of the habitat, with individuals occupying the middle of the habitat showing a signature of admixture (Bradburd et al., 2018). Numerous methods have been proposed to account for spatial and temporal "confounding" in PCA and the PSD model, leading to, sometimes, more interpretable solutions but typically with higher computational cost (Bradburd et al., 2018; Frichot et al., 2012; François et al., 2019).

The introduction of the Estimating Effective Migration Surfaces (EEMS) method bridged a gap between computational tractability and interpretability with respect to spatial population genetic models (Petkova et al., 2016). Like in "Migration Matrix Models", EEMS assumes individuals occupy a discrete spatial habitat, specifically a triangular lattice embedded in geographic space (Bodmer and Cavalli-Sforza, 1968). The input data is a set of spatial coordinates for each individual as well as a genetic distance matrix computed across all individuals in the dataset. EEMS uses a resistance distance approximation to compute the expected genetic distance between sub-populations under a coalescent-based stepping stone model (CSSM) (McRae, 2006; Slatkin, 1993). The edge weights of the graph parameterize this resistance distance and thus provide a likelihood for observed genetic distances. EEMS uses a hierarchical Bayesian model and a Voronoi tessellation based prior to encourage spatial smoothness in the fitted edge weights, specifically piece-wise constant smoothness. EEMS uses MCMC to find the posterior distribution of the latent variables and outputs a visualization of the posterior mean for effective migration and genetic diversity for every spatial position of the focal habitat. Regions with relatively low effective migration can be interpreted to have restricted gene-flow over time whereas regions with relatively high migration can be interpreted as having elevated gene-flow. The EEMS framework has been recently extended to use genetic distances computed from identity-by-descent blocks, using a random walk approximation to the CSSM for computation of expected distances. This framework allows for the ability to estimate migration rates and population sizes in different temporal periods, by binning identity-by-descent blocks by length (Al-Asadi et al., 2019).

To help ground the work in this dissertation, I look back to a enduring review written in 1982 by Joseph Felsenstein (Felsenstein, 1982). In this review titled "*How can we infer geography and history from gene frequencies?*", Felsenstein posed a number of open statistical problems in the field of population genetics, some of which I restate here,

- *Problem 1:* For any given covariance matrix, is there a corresponding migration matrix which would be expected to lead to it? If so, how can we find it?

- *Problem 2:* How can we characterize the set of possible migration matrices which are compatible with a given set of observed covariances?

- *Problem 3:* How can we confine our attention to migration patterns which are consistent with the known geometric co-ordinates of the populations

- *Problem 4:* How can we make valid statistical estimates of parameters of stepping stone models?

- *Problem 5:* How can we parameterize the space of historical patterns of migration so that we know which sets of patterns are consistent with our data?

These problems remain largely unsolved in the field nearly 40 years later. Addressing some of these problem in both real-word applications of population structure inference methods as well as the development of new statistical methods is the primary focus of this dissertation.

CHAPTER 2

GENETIC HISTORY FROM THE MIDDLE NEOLITHIC TO PRESENT ON THE MEDITERRANEAN ISLAND OF SARDINIA

This chapter has been published as part of a large collaborative project of which I am a co-first author. In this chapter I include the components I led or substantially contributed to. For the complete manuscript and full author list see Marcus et al. (2020c), and for a listing of author contributions see section 2.9 below. All co-authored material included is used with permission.

2.1 Abstract

The island of Sardinia has been of particular interest to geneticists for decades. The current model for Sardinia's genetic history describes the island as harboring a founder population that was established largely from the Neolithic peoples of southern Europe which remained isolated from later Bronze Age expansions on the mainland. To evaluate this model, we generated genome-wide ancient DNA data for 70 individuals from more than 20 Sardinian archaeological sites spanning the Middle Neolithic through the Medieval period. The earliest individuals show a strong affinity to western Mediterranean Neolithic populations, followed by an extended period of genetic continuity on the island through the Nuragic period (second millennium BCE). Beginning with individuals from Phoenician/Punic sites (first millennium BCE), we observe spatially-varying signals of admixture with sources principally from the eastern and northern Mediterranean. Overall, our analysis sheds light on the genetic history of Sardinia, revealing how relationships to mainland populations shifted over time.

2.2 Introduction

The whole-genome sequencing in 2012 of “Ötzi”, an individual who was preserved in ice for over 5,000 years near the Italo-Austrian border, revealed a surprisingly high level of shared ancestry with present-day Sardinian individuals (Keller et al., 2012; Sikora et al., 2014). Subsequent work on genome-wide variation in ancient Europeans found that most “early European farmer” individuals, even when from geographically distant locales (e.g. from Sweden, Hungary and Spain) have their highest genetic affinity with present-day Sardinian individuals (Skoglund et al., 2012, 2014a; Gamba et al., 2014; Olalde et al., 2015). Accumulating ancient DNA (aDNA) results have provided a framework for understanding how early European farmers show such genetic affinity to modern Sardinians.

In this framework, Europe was first inhabited by Paleolithic and later Mesolithic hunter-gatherer groups. Then, starting about 7,000 BCE, farming peoples arrived from the Middle East as part of a Neolithic transition (Lazaridis et al., 2014), spreading through Anatolia and the Balkans (Hofmanová et al., 2016; Mathieson et al., 2018) while progressively admixing with local hunter-gatherers (Lipson et al., 2017). Major movements from the Eurasian Steppe, beginning about 3,000 BCE, resulted in further admixture throughout Europe (Allentoft et al., 2015; Haak et al., 2015; Olalde et al., 2018, 2019). These events are typically modeled in terms of three ancestry components: western hunter gatherers (“WHG”), early European farmers (“EEF”), and Steppe pastoralists (“Steppe”). Within this broad framework, the island of Sardinia is thought to have received a high level of EEF ancestry early on and then remained mostly isolated from the admixture occurring on mainland Europe (Keller et al., 2012; Sikora et al., 2014). However, this specific model for Sardinian population history has not been tested with genome-wide aDNA data from the island.

The oldest known human remains on Sardinia date to ~20,000 years ago (Melis, 2002). Archaeological evidence suggests Sardinia was not densely populated in the Mesolithic, and experienced a population expansion coinciding with the Neolithic transition in the sixth millennium BCE (Luglië, 2018). Around this time, early Neolithic pottery assemblages were spreading throughout the western Mediterranean, including Sardinia, in particular vessels decorated with Cardium shell impressions (variably described as Impressed Ware, Cardial Ware, Cardial Impressed Ware)(Barnett, 2000), with radio-carbon dates indicating a rapid westward maritime expansion around 5,500 BCE (Zilhão, 2001). In the later Neolithic, obsidian originating from Sardinia is found throughout many western Mediterranean archaeological sites (Tykot, 1996), indicating that the island was integrated into a maritime trade network. In the middle Bronze Age, about 1,600 BCE, the "Nuragic" culture emerged, named for the thousands of distinctive stone towers, *nuraghi* (Webster, 2016). During the late Nuragic period, the archaeological and historical record shows the direct influence of several major Mediterranean groups, in particular the presence of Mycenaean, Levantine and Cypriot traders. The Nuragic settlements declined throughout much of the island as, in the late 9th and early 8th century BCE, Phoenicians originating from present-day Lebanon and northern Palestine established settlements concentrated along the southern shores of Sardinia (Moscati, 1966). In the second half of the 6th century BCE, the island was occupied by Carthaginians (also known as Punics), expanding from the city of Carthage on the North African coast of present-day Tunisia, which was founded in the late 9th century by Phoenicians (Van Dommelen, 2006; Guirguis et al., 2017). Sardinia was occupied by Roman forces in 237 BCE, and turned into a Roman province a decade later (Dyson and Rowland, 2007). Throughout the Roman Imperial period, the island remained closely aligned with both Italy and central North Africa. After the fall of the Roman empire, Sardinia became increasingly autonomous (Dyson and Rowland, 2007), but interaction with the Byzantine Empire, the maritime republics of Genova and Pisa, the Catalan and Aragonese Kingdom, and the Duchy of Savoy and Piemonte continued to influence the island (Ortu, 2011; Mastino, 2005).

The population genetics of Sardinia has long been studied, in part because of its importance for medical genetics (Calò et al., 2008; Lettre and Hirschhorn, 2015). Pioneering studies found evidence that Sardinia is a genetic isolate with appreciable population substructure (Siniscalco et al., 1966; Contu et al., 1992; Lampis et al., 2000). Recently, Chiang et al. (Chiang et al., 2018) analyzed whole genome sequences (Sidore et al., 2015) together with continental European aDNA. Consistent with previous studies, they found the mountainous Ogliastra region of central/eastern Sardinia carries a signature of relative isolation and subtly elevated levels of WHG and EEF ancestry.

Four previous studies have analyzed aDNA from Sardinia using mitochondrial DNA. Ghirotto et al. (Ghirotto et al., 2009) found evidence for more genetic turnover in Gallura (a region in northern Sardinia with cultural/linguistic connections to Corsica) than Ogliastra. Modi et al. (Modi et al., 2017) sequenced mitogenomes of two Mesolithic individuals and found support for a model of population replacement in the Neolithic. Olivieri et al. (Olivieri et al., 2017) analyzed 21 ancient mitogenomes from Sardinia and estimated the coalescent times of Sardinian-specific mtDNA haplogroups, finding support for most of them originating in the Neolithic or later, but with a few coalescing earlier. Finally, Matisoo-Smith et al. (Matisoo-Smith et al., 2018) analyzed mitogenomes in a Phoenician settlement on Sardinia and inferred continuity and exchange between the Phoenician population and broader Sardinia. One additional study recovered β-thalessemia variants in three aDNA samples and found one carrier of the cod39 mutation in a necropolis used in the Punic and Roman periods (Viganó et al., 2017). Despite the initial insights these studies reveal, none of them analyze genome-wide autosomal data, which has proven to be useful for inferring population history (Pickrell and Reich, 2014).

Here, we generated genome-wide data from the skeletal remains of 70 Sardinian individuals radiocarbon dated to between 4,100 BCE - 1,500 CE. We investigated three aspects of Sardinian population history: First, the ancestry of individuals from the Sardinian Neolithic (ca. 5,700-3,400 BCE) – who were the early peoples expanding onto the island at this time? Second, the genetic structure through the Sardinian Chalcolithic (i.e. Copper Age, ca. 3,400-

2,300 BCE) to the Sardinian Bronze Age (ca. 2,300-1,000 BCE) – were there genetic turnover events through the different cultural transitions observed in the archaeological record? And third, the post-Bronze Age contacts with major Mediterranean civilizations and more recent Italian populations – have they resulted in detectable gene flow?

Our results reveal insights about each of these three periods of Sardinian history. Specifically, our earliest samples show affinity to the early European farmer populations of the mainland, then we observe a period of relative isolation with no significant evidence of admixture through the Nuragic period, after which we observe evidence for admixture with sources from the Northern and Eastern Mediterranean.

2.3 Results

2.3.1 Ancient DNA from Sardinia

We organized a collection of skeletal remains (Supp. Fig. 1) from: 1) a broad set of previously excavated samples initially used for isotopic analysis (Lai et al., 2013), 2) the Late Neolithic to Bronze Age Seulo cave sites of central Sardinia (Skeates et al., 2013), 3) the Neolithic Sites Noedalle and S'isterridolzu (Germanà, 1980), 4) the Phoenician-Punic sites of Monte Sirai (Guirguis et al., 2017) and Villamar (Pompianu and Murgia, 2017), 5) the Imperial Roman period site at Monte Carru (Alghero) (La Fragola and Rovina, 2018), 6) medieval remains from the site of Corona Moltana (Meloni, 2004), 7) medieval remains from the necropolis of the Duomo of San Nicola (Rovina and Fiori M. Olia, 2013). We sequenced DNA libraries enriched for the complete mitochondrial genome as well as a targeted set of 1.2 million single nucleotide polymorphisms (SNPs) (Fu et al., 2015). After quality control, we arrived at a final set of 70 individuals with an average coverage of 1.02× (ranging from 0.04× to 5.39× per individual) and a median number of 466,049 sites covered at least once per individual. We obtained age estimates by either direct radiocarbon dating ($n = 53$), previously reported radiocarbon dates (n = 13), or archaeological context and radiocarbon dates from the same burial site (n = 4). The estimated ages range from 4,100 years BCE to 1,500 years CE (Fig. 2.1, Supp. Data 1A). We pragmatically grouped the data into broad periods: Middle / Late Neolithic ('Sar-MN', 4,100-3,500 BCE, n = 6), Early Copper Age ('Sar-ECA', 3,500-2,500 BCE, n = 3), Early Middle Bronze Age ('Sar-EMBA', 2,500-1,500 BCE, n = 27) and Nuragic ('Sar-Nur', 1,500-900 BCE, n = 16). For the post-Nuragic sites, there is substantial genetic heterogeneity within and among sites, and so we perform analysis per site when grouping is necessary ('Sar-MSR' and 'Sar-VIL' for the Phoenician and Punic sites of Monte Sirai, n = 2; and Villamar, n = 6; 'Sar-ORC002' for a Punic period individual from the interior site of S'Orcu 'e Tueri, n = 1; 'Sar-AMC' for the Roman period site of Monte Carru near Alghero, n = 3; 'Sar-COR' for the early medieval individuals from the site of Corona Moltana, n = 2; and 'Sar-SNN' for the medieval San Nicola necropoli, n = 4). Figure 2.1 provides an overview of the sample.

To assess the relationship of the ancient Sardinian individuals to other ancient and present-day west Eurasian and north African populations we analyzed our individuals alongside published autosomal DNA data (ancient: 972 individuals (Mathieson et al., 2015; Lazaridis et al., 2016, 2017; Mathieson et al., 2018; Lipson et al., 2017; Olalde et al., 2018); modern: 1,963 individuals from outside Sardinia (Lazaridis et al., 2014) and 1,577 individuals from Sardinia (Sidore et al., 2015; Chiang et al., 2018)). For some analyses, we grouped the modern Sardinian individuals into eight geographic regions (see inset in panel C of Figure 2 for listing and abbreviations, also see Supp. Data 1E) and for others we subset the more isolated Sardinian region of Ogliastra ('Sar-Ogl', n =419) and the remainder ('Sar-non Ogl', n =1,158). As with other human genetic variation studies, population annotations are important to consider in the interpretation of results.

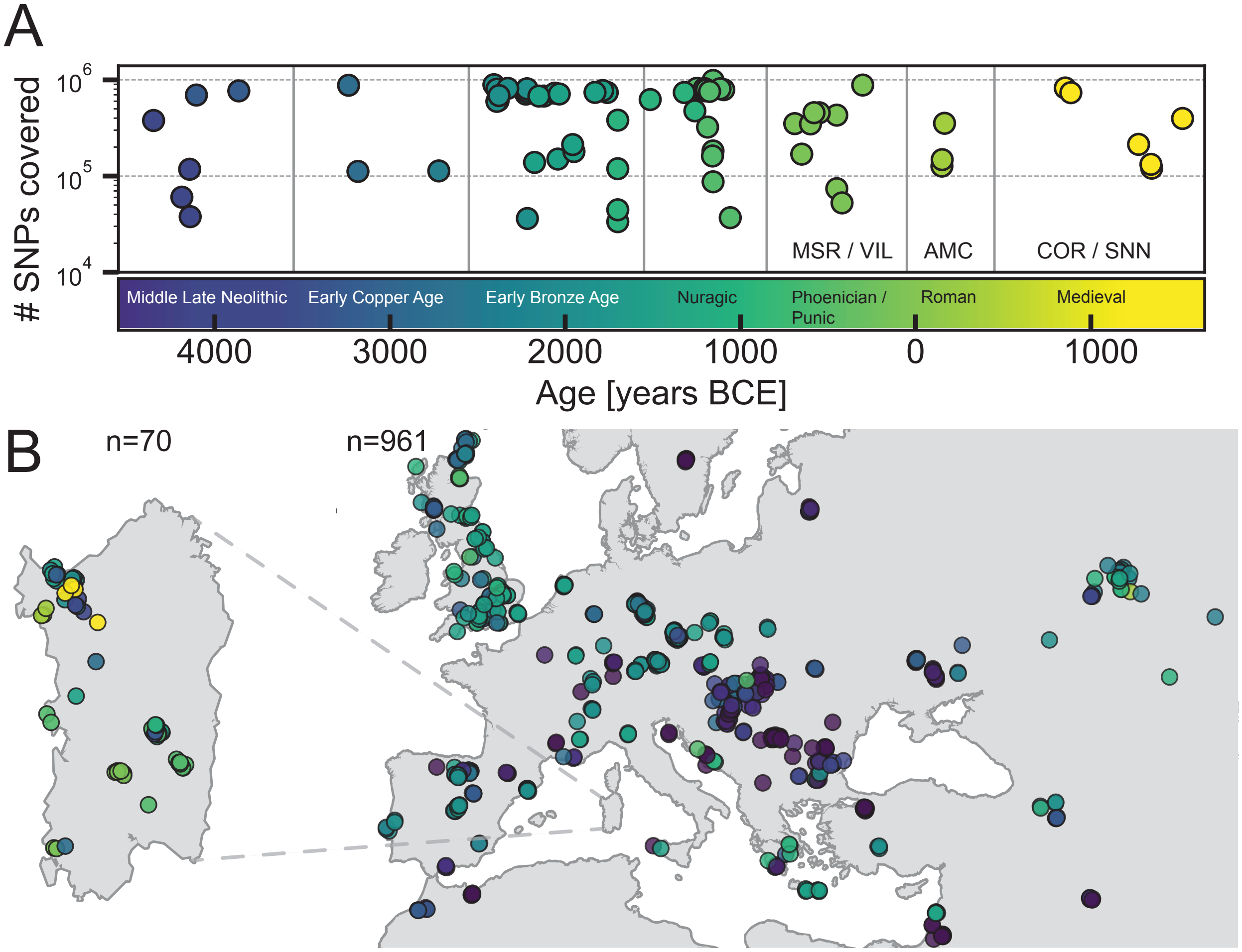


Figure 2.1: **Number of SNPs covered, sampling locations and ages of ancient individuals.** A: The number of SNPs covered at least once and age (mean of 2σ radiocarbon age estimates) for the 70 ancient Sardinian individuals. B: The sampling locations of ancient Sardinian individuals and a reference dataset of 961 ancient individuals from across western Eurasia and North Africa.

2.3.2 Similarity to western mainland Neolithic populations

We found low differentiation between Middle/Late Neolithic Sardinian individuals and Neolithic western mainland European populations, in particular groups from Spain (Iberia-EN) and southern France (France-N). When projecting ancient individuals onto the top two principal components (PCs) defined by modern variation, the Neolithic ancient Sardinian individuals sit between early Neolithic Iberian and later Copper Age Iberian populations, roughly on an axis that differentiates WHG and EEF populations, and embedded in a cluster that additionally includes Neolithic British individuals (Fig. 2.2). This result is also evident in terms of genetic differentiation, with low pairwise $F_{ST} \approx 0.005$-0.008, between Middle/Late Neolithic and Neolithic western mainland European populations (Fig. 2.3). Pairwise outgroup-f_3 analysis shows a similar pattern, with the highest values of f_3 (i.e. most shared drift) being with Western European Neolithic and Copper Age populations (Fig. 2.3), gradually dropping off for populations more distant in time or space (Supp. Fig. 10).

Ancient Sardinian individuals are shifted towards WHG individuals in the top two PCs relative to early Neolithic Anatolians (Fig. 2.2). Analysis using qpAdm shows that a two-way admixture model between WHG and Neolithic Anatolian populations is consistent with our data (e.g. $p = 0.376$ for Sar-MN, Tab. 2.1), similar to other western European populations of the early Neolithic (Supp. Tab. 1). The method estimates ancient Sardinian individuals harbor HG ancestry ($\approx 17 \pm 2\%$) that is higher than early Neolithic mainland populations (including Iberia, 8.7±1.1%), but lower than Copper Age Iberians (25.1±0.9%) and about the same as Southern French Middle-Neolithic individuals (21.3±1.5%) (Tab. 2.1, Supp. Fig. 13) (± denotes plus and minus one standard error).

In explicit models of continuity (using qpAdm, see Methods) the southern French Neolithic individuals (France-N) are consistent with being a single source for Middle / Late Neolithic Sardinia ($p = 0.38$ to reject the model of one population being the direct source of the other); followed by other western populations high in EEF ancestry, though with poor

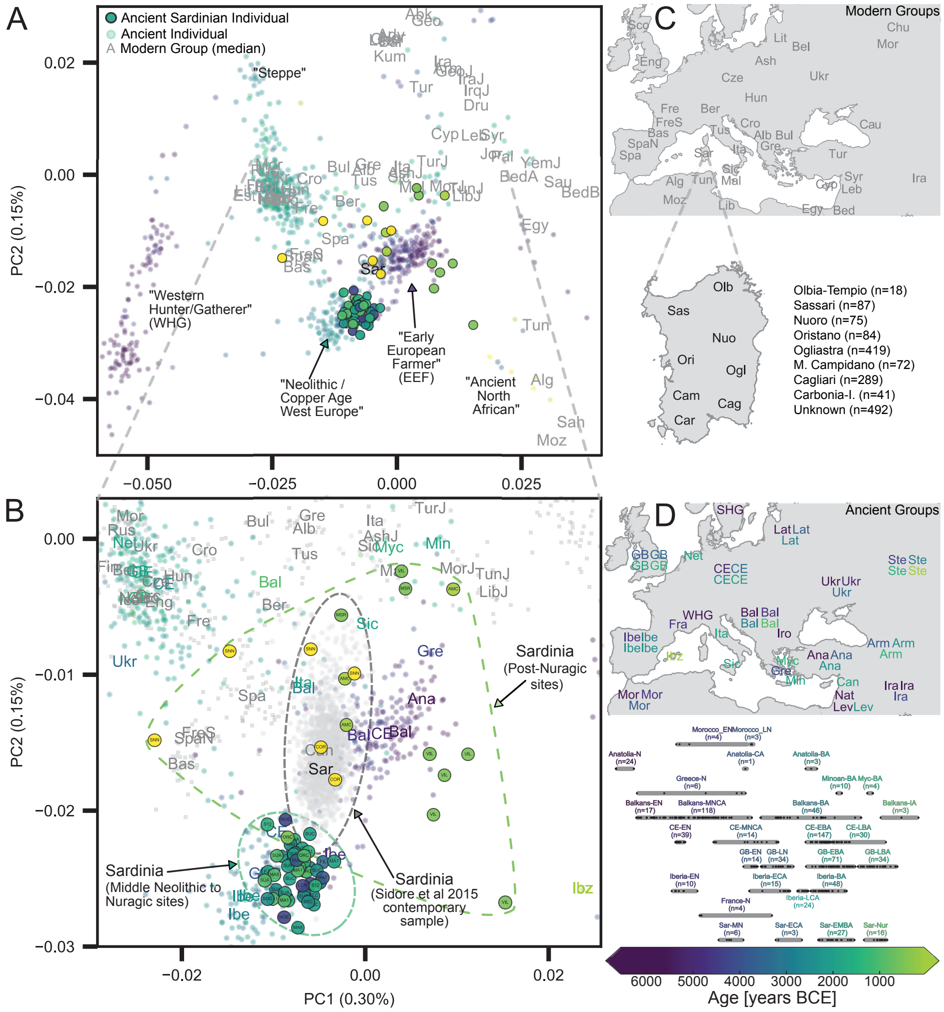


Figure 2.2: **Principal Components Analysis based on the Human Origins dataset.** A: Projection of ancient individuals' genotypes onto principal component axes defined by modern Western Eurasians and North Africans (gray labels, see panel C for legend for all abbreviations but 'Can', for Canary Islands.). B: Zoom into the region most relevant for Sardinia. Each projected ancient individual is displayed as a transparent colored point in panel A and B, with the color determined by the age of each sample (see panel D for legend). In panel B, median PC1 and PC2 values for each population are represented by three-letter abbreviations, with black or gray font for moderns and a color-coded font based on the mean age for ancient populations. Ancient Sardinian individuals are plotted as circles with edges, color-coded by age, and with the first three letters of their sample ID (which typically indicates the archaeological site). Modern individuals from the Sidore et al sample of Sardinia are represented with gray circles and modern individuals from the reference panel with grey squares. See Figure 5 for a zoomed in representation with detailed province labels for Sardinian individuals. The full set of labels and abbreviations are described in Supp. Data 1E and 1F. C: Geographic legend of present-day individuals from the Human Origins and our Sardinian reference dataset. D: Timeline of selected ancient groups. Note: The same geographic abbreviation can appear multiple times with different colors to represent groups with different median ages.

fit (qpAdm p-values $< 10^{-5}$, Supp. Tab. 2). France-N may result in improved fits as it is a better match for the WHG and EEF proportions seen in Middle / Late Neolithic Sardinia (Supp. Tab. 1). As we discuss below, caution is necessary as there is a lack of aDNA from other relevant populations of the same period (such as mainland Italian Neolithic cultures and neighboring islands).

For our sample from the Middle Neolithic through the Nuragic ($n = 52$ individuals), we were able to infer mtDNA haplotypes for each individual and Y haplotypes for 30 out of 34 males. The mtDNA haplotypes belong to macro-haplogroups HV ($n = 20$), JT (n =19), U (n =12) and X (n =1), a composition broadly similar to other European Neolithic populations. For Y haplotypes, we found at least one carrier for each of three major Sardinia-specific Y founder clades (within the haplogroups I2-M26, G2-L91 and R1b-V88) that were identified previously based on modern Sardinian data (Francalacci et al., 2013). More than half of the 31 identified Y haplogroups were R1b-V88 or I2-M223 (n =11 and n =8, respectively, Supp. Fig. 6, Supp. Data 1B), both of which are also prevalent in Neolithic Iberians (Olalde et al., 2019). Compared to most other ancient populations in our reference dataset, the frequency of R1b-V88 (Supp. Note 3, Supp. Fig. 6) is relatively high, but as we observed clustering of Y haplogroups by sample location (Supp. Data 1B) caution should be exercised with interpreting our results as estimates for island-wide Y haplogroup frequencies. The oldest individuals in our reference data carrying R1b-V88 or I2-M223 were Balkan hunter-gatherer and Neolithic individuals, and both haplogroups later appear also in western Neolithic populations (Supp. Fig. 7-9).

2.3.3 Continuity from the Middle Neolithic through the Nuragic

We found several lines of evidence supporting genetic continuity from the Sardinian Middle Neolithic into Bronze Age and Nuragic times. Importantly, we observed low genetic differentiation between ancient Sardinian individuals from various time periods (F_{ST} =

0.0055 ± 0.0014 between Middle / Late Neolithic and late Bronze Age, Fig. 2.3). Furthermore, we did not observe temporal substructure within the ancient Sardinian individuals in the top two PCs – they form a coherent cluster (Fig. 2.2). In stark contrast, ancient individuals from mainland regions such as central Europe show large movements over the first two PCs from the Neolithic to the Bronze Age, and also have higher pairwise differentiation (e.g. F_{ST} =0.0200± 0.0004 between Neolithic and Bronze Age individuals from central Europe, Supp. Fig. 11). A qpAdm analysis cannot reject a model of Middle / Late Neolithic Sardinian individuals being a direct predecessor of Nuragic Sardinian individuals ($p = 0.15$, Supp. Tab. 2, also see results for f_4 statistics, Supp. Data 2). Our qpAdm analysis further shows that the WHG ancestry proportion, in a model of admixture with Neolithic Anatolia, remains stable at 17 ± 2% through the Nuragic period (Tab. 2.1A). When using a three-way admixture model, we do not detect significant Steppe ancestry in any ancient Sardinian group from the Middle / Late Neolithic to the Nuragic, as is inferred, for example, in later Bronze Age Iberians (Tab. 2.1B, Supp. Fig. 13). Finally, in a 5-way model with Iran Neolithic and Moroccan Neolithic samples added as sources, neither source is inferred to contribute ancestry during the Middle Neolithic to Nuragic (point estimates are statistically indistinguishable from zero, Supp. Fig 14).

2.3.4 From the Nuragic period to present-day Sardinia: Signatures of admixture

We found multiple lines of evidence for gene flow into Sardinia after the Nuragic period. The present-day Sardinian individuals from the Sidore et al sample are shifted from the Nuragic period ancients on the western Eurasian / north African PCA (Fig. 2.2). Using a "shrinkage" correction for the projection is key for detecting this shift (see Supp. Fig. 23 for an evaluation of different PCA projection techniques). In the ADMIXTURE results (Fig. 2.4), present-day Sardinian individuals carry a modest "Steppe-like" ancestry component (but

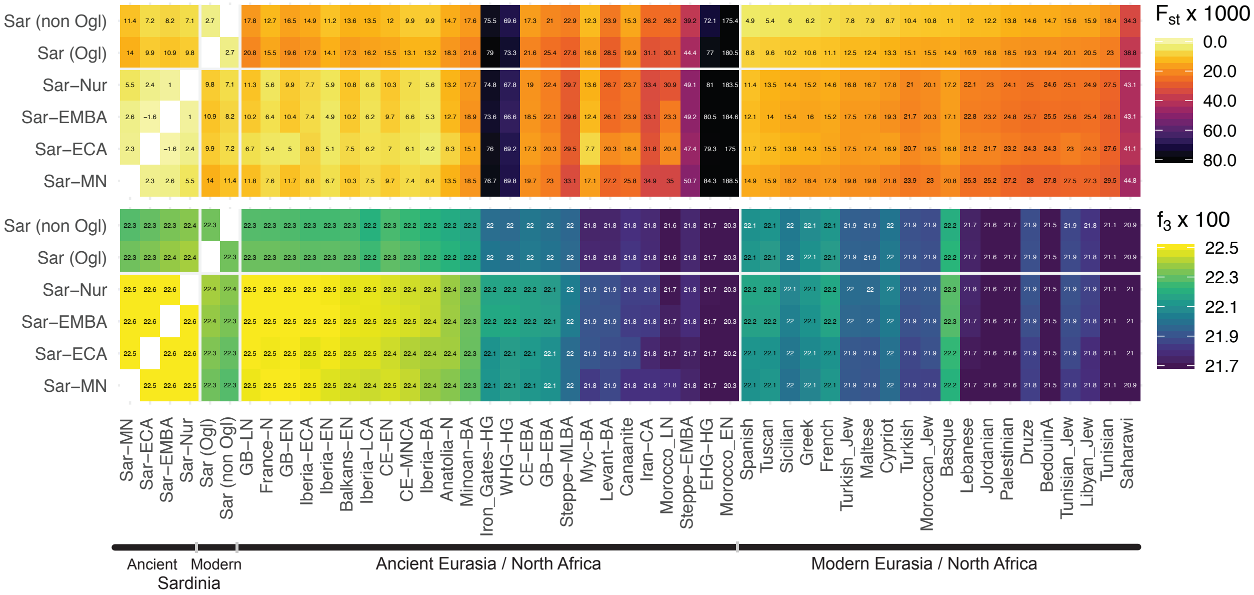


Figure 2.3: **Genetic similarity matrices.** We calculated F_{ST} (upper panel) and outgroup-f_3 (lower panel) of ancient Sardinian (Middle / Late Neolithic to Nuragic periods) and modern Sardinian individuals (grouped into within and outside the Ogliastra region) with each other (left), various ancient (middle), and modern populations (right) of interest. The full sharing matrices can be found in Supp. Fig. 10/11, where we also include post-Nuragic sites.

generally less than continental present-day European populations), and an appreciable "eastern Mediterranean" ancestry component (also inferred at a high fraction in other present-day Mediterranean populations, such as Sicily and Greece) relative to Nuragic period and earlier Sardinian individuals.

To further refine this recent admixture signal, we considered two-way, three-way, and four-way models of admixture with qpAdm (Tab. 2, Supp. Fig. 15-18, Supp. Tab. 3-5). We find three-way models fit well ($p > 0.01$) that contain admixture between Nuragic Sardinia, one northern Mediterranean source (e.g. individuals with group labels Lombardy, Tuscan, French, Basque, Spanish) and one eastern Mediterranean source (e.g. individuals with group labels Turkish-Jew, Libyan-Jew, Maltese, Tunisian-Jew, Moroccan-Jew, Lebanese, Druze, Cypriot, Jordanian, Palestinian) (Table 2C,D). Maltese and Sicilian individuals can provide two-way model fits (Tab. 2B), but appear to reflect a mixture of N. Mediterranean and E. Mediterranean ancestries, and as such they can serve as single-source proxies in two-way admixture models with Nuragic Sardinia. For four-way models including N. African ancestry, the inferences of N. African ancestry are negligible (though as we show below, forms of N. African ancestry were already likely present in the eastern Mediterranean components).

Because of limited sample sizes and ancestral source mis-specification, caution is warranted when interpreting inferred admixture fractions; however, the results indicate that complex post-Nuragic gene flow has likely played a role in the population genetic history of Sardinia.

2.3.5 Refined signatures of post-Nuragic admixture and heterogeneity

To more directly evaluate the models of post-Nuragic admixture, we obtained aDNA from 17 individuals sampled from post-Nuragic sites. The post-Nuragic individuals spread across a wide range of the PCA, and many shift towards the "eastern" and "northern" Mediterranean sources posited above (Fig. 2.2). We confidently reject qpAdm models of continuity from

		Proxy Source Populations				**Admixture Fractions**			**Standard Error**		
	Target	**a**	**b**	**c**	**p-value**	**a**	**b**	**c**	**a**	**b**	**c**
A	Sar-MN	WHG	Anatolia-N	-	0.376	0.177	0.823	-	0.014	0.014	-
	Sar-ECA	WHG	Anatolia-N	-	0.268	0.161	0.839	-	0.020	0.020	-
	Sar-EMBA	WHG	Anatolia-N	-	0.049	0.161	0.839	-	0.007	0.007	-
	Sar-Nur	WHG	Anatolia-N	-	0.134	0.163	0.837	-	0.009	0.009	-
B	Sar-MN	WHG	Anatolia-N	Steppe	0.265	0.177	0.823	0.000	0.016	0.023	0.026
	Sar-ECA	WHG	Anatolia-N	Steppe	0.18	0.164	0.836	0.000	0.023	0.032	0.036
	Sar-EMBA	WHG	Anatolia-N	Steppe	0.032	0.162	0.838	0.000	0.009	0.012	0.013
	Sar-Nur	WHG	Anatolia-N	Steppe	0.089	0.163	0.837	0.000	0.010	0.014	0.016
C	France-N	WHG	Anatolia-N	Steppe	0.093	0.213	0.787	0.000	0.018	0.023	0.027
	Iberia-EN	WHG	Anatolia-N	Steppe	0.243	0.087	0.913	0.000	0.012	0.017	0.019
	Iberia-LCA	WHG	Anatolia-N	Steppe	0.045	0.251	0.749	0.000	0.012	0.015	0.018
	Iberia-BA	WHG	Anatolia-N	Steppe	$6.0 \cdot 10^{-3}$	0.239	0.689	0.072	0.010	0.014	0.016
	CE-EN	WHG	Anatolia-N	Steppe	0.656	0.046	0.954	0.000	0.007	0.010	0.012
	CE-LBA	WHG	Anatolia-N	Steppe	0.105	0.128	0.403	0.468	0.008	0.011	0.013

Table 2.1: **Results from fitting models of admixture with `qpAdm` for Middle Neolithic to Nuragic period**. A) Two-way models of admixture for ancient Sardinia using Western Hunter-Gatherer (WHG) and Neolithic Anatolia (Anatolia-N) individuals as proxy sources. B) Three-way models of admixture for ancient Sardinia using Western Hunter-Gatherer (WHG), Neolithic Anatolia (Anatolia-N), and Early Middle Bronze Age Steppe (Steppe-EMBA, abbreviated Steppe in table), individuals as proxy sources. C) Three-way models for select comparison populations on the European mainland. Full results are reported in Supp. Info. 4.

the Nuragic period for all of these post-Nuragic samples, apart from a sample from S'Orcu 'e Tueri (ORC002, Tab. 2.2E, Supp. Tab. 6). The ADMIXTURE results concur, most post-Nuragic individuals show the presence of novel ancestry components not inferred in any of the more ancient individuals (Fig. 2.4).

Consistent with an influx of novel ancestry, we observed that haplogroup diversity increases after the Nuragic period. In particular, we identified one carrier of the mtDNA haplogroup L2a at both the Punic Villamar site and the Roman Monte Carru site. At present, this mtDNA haplogroup is common across Africa, but so far undetected in samples from Sardinia (Olivieri et al., 2017). We also found several Y haplogroups absent in our Neolithic trough the Nuragic period sample (Supp. Fig. 6). R1b-M269, at about 15 % within modern Sardinian males (Francalacci et al., 2013), appears in one Punic (VIL011) and two Medieval individuals (SNN002 and SNN004). We also observed J1-L862 in one individual from a Punic site (VIL007) and E1b-L618 in one medieval individual (SNN001). Notably, J1-

		Proxy Source Populations				Admixture Fractions			Standard Error		
	Target	**a**	**b**	**c**	**p-value**	**a**	**b**	**c**	**a**	**b**	**c**
A	Sar-ECA	Sar-MN	-	-	0.175	-	-	-	-	-	-
	Sar-EMBA	Sar-ECA	-	-	0.769	-	-	-	-	-	-
	Sar-Nur	Sar-EMBA	-	-	0.765		-	-	-	-	-
	Cagliari	Sar-Nur	-	-	$< 10^{-30}$	-	-	-		-	-
B	Cagliari	Sar-Nur	Sicilian	-	0.011	0.545	0.455	-	0.021	0.021	-
	Cagliari	Sar-Nur	Maltese	-	0.011	0.573	0.427	-	0.019	0.019	-
	Cagliari	Sar-Nur	Turkish	-	$2.4 \cdot 10^{-3}$	0.699	0.301	-	0.014	0.014	-
	Cagliari	Sar-Nur	Tuscan	-	$2.3 \cdot 10^{-4}$	0.522	0.478	-	0.024	0.024	-
C	Cagliari	Sar-Nur	N Mediterranean	Turkish-Jew	0.168	0.512	0.237	0.251	0.021	0.042	0.034
	Cagliari	Sar-Nur	N Mediterranean	Libyan-Jew	0.046	0.515	0.299	0.186	0.023	0.038	0.028
	Cagliari	Sar-Nur	N Mediterranean	Tunisian-Jew	0.037	0.498	0.312	0.191	0.022	0.035	0.028
	Cagliari	Sar-Nur	N Mediterranean	Druze	0.031	0.542	0.287	0.170	0.022	0.037	0.024
	Cagliari	Sar-Nur	N Mediterranean	Moroccan-Jew	0.026	0.519	0.275	0.206	0.022	0.042	0.032
	Cagliari	Sar-Nur	N Mediterranean	Cypriot	0.026	0.520	0.297	0.184	0.021	0.035	0.026
	Cagliari	Sar-Nur	N Mediterranean	Maltese	0.025	0.541	0.132	0.327	0.025	0.069	0.058
	Cagliari	Sar-Nur	N Mediterranean	Lebanese	0.023	0.550	0.299	0.151	0.024	0.038	0.023
	Cagliari	Sar-Nur	N Mediterranean	Sicilian	0.021	0.526	0.113	0.361	0.022	0.064	0.058
	Cagliari	Sar-Nur	N Mediterranean	Jordanian	$10.0 \cdot 10^{-3}$	0.542	0.319	0.138	0.023	0.035	0.021
	Cagliari	Sar-Nur	N Mediterranean	Greek	$8.7 \cdot 10^{-3}$	0.551	0.000	0.449	0.037	0.217	0.191
	Cagliari	Sar-Nur	N Mediterranean	Palestinian	$7.0 \cdot 10^{-3}$	0.542	0.331	0.127	0.024	0.035	0.019
	Cagliari	Sar-Nur	N Mediterranean	Turkish	$6.3 \cdot 10^{-3}$	0.637	0.136	0.228	0.033	0.067	0.039
	Cagliari	Sar-Nur	N Mediterranean	BedouinA	$2.0 \cdot 10^{-3}$	0.540	0.351	0.109	0.024	0.033	0.017
	Cagliari	Sar-Nur	N Mediterranean	Egyptian	$1.4 \cdot 10^{-4}$	0.529	0.389	0.082	0.025	0.031	0.014
	Cagliari	Sar-Nur	N Mediterranean	Tunisian	$3.7 \cdot 10^{-5}$	0.524	0.404	0.072	0.025	0.031	0.014
D	Cagliari	Sar-Nur	E Mediterranean	Lombardy	0.168	0.512	0.251	0.237	0.021	0.034	0.042
	Cagliari	Sar-Nur	E Mediterranean	Tuscan	0.09	0.527	0.198	0.275	0.020	0.047	0.053
	Cagliari	Sar-Nur	E Mediterranean	Greek	0.079	0.548	0.151	0.302	0.018	0.054	0.057
	Cagliari	Sar-Nur	E Mediterranean	French	0.05	0.560	0.324	0.116	0.019	0.027	0.023
	Cagliari	Sar-Nur	E Mediterranean	Basque	0.034	0.533	0.340	0.128	0.021	0.025	0.025
	Cagliari	Sar-Nur	E Mediterranean	Spanish	0.023	0.540	0.309	0.151	0.020	0.029	0.030
	Cagliari	Sar-Nur	E Mediterranean	Sicilian	0.013	0.544	0.000	0.456	0.029	0.226	0.245
	Cagliari	Sar-Nur	E Mediterranean	Maltese	0.012	0.572	0.000	0.428	0.026	0.261	0.266
	Cagliari	Sar-Nur	E Mediterranean	Turkish	$1.2 \cdot 10^{-3}$	0.700	0.000	0.300	0.058	0.190	0.135
	Cagliari	Sar-Nur	E Mediterranean	Cypriot	$3.8 \cdot 10^{-5}$	0.587	0.413	0.000	0.047	0.310	0.275
E	Sar-VIL	Sar-Nur	-	-	$< 10^{-30}$	-	-	-	-	-	-
	Sar-MSR	Sar-VIL	-	-	$1.8 \cdot 10^{-6}$	-	-	-	-	-	-
	Sar-AMC	Sar-MSR	-	-	0.203	-	-	-	-	-	-
	Sar-SNN	Sar-MSR	-	-	0.037	-	-	-	-	-	-
	Sar-COR	Sar-AMC	-	-	0.014	-	-	-	-	-	-
	Sar-SNN	Sar-AMC	-	-	0.124	-	-	-	-	-	-
F	Cagliari	Sar-VIL	-	-	$9.1 \cdot 10^{-12}$	-	-	-	-	-	-
	Cagliari	Sar-MSR	-	-	0.078	-	-	-	-	-	-
	Cagliari	Sar-AMC	-	-	0.012	-	-	-	-	-	-
	Cagliari	Sar-COR	-	-	0.16	-	-	-	-	-	-
	Cagliari	Sar-SNN	-	-	0.037	-	-	-	-	-	-
	Ogliastra	Sar-VIL	-	-	$8.6 \cdot 10^{-14}$	-	-	-	-	-	-
	Ogliastra	Sar-MSR	-	-	0.044	-	-	-	-	-	-
	Ogliastra	Sar-AMC	-	-	$2.2 \cdot 10^{-3}$	-	-	-	-	-	-
	Ogliastra	Sar-COR	-	-	0.261	-	-	-	-	-	-
	Ogliastra	Sar-SNN	-	-	0.016	-	-	-	-	-	-

Table 2.2: **Results from fitting models of admixture with `qpAdm` and Sardinian aDNA as sources.** A) Single-source models to assess continuity of each Sardinian period with the previous one (see main text for guide to abbreviations). B) Results of two-way models of admixture for Cagliari (a representative present-day sample). C) Results of three-way models showing multiple eastern Mediterranean populations that can produce viable models (Results shown with individuals from Lombardy [Bergamo in the Human Origins array (HOA) dataset, see Materials and Methods] as one of several possible proxies for north Mediterranean ancestry, see part C). C) Results of three-way models showing multiple north Mediterranean populations that can produce viable models (Results shown with Jewish individuals from Turkey ['Turkish-Jew' in the dataset] used as one of several possible proxies for east Mediterranean ancestry, see part B). E) Results of single-source models to assess continuity among post Nuragic sites. F) Results of single-source models to assess continuity between the Medieval period samples and present-day samples (Cagliari and Ogliastra taken as representatives). Full results are reported in Supp. Info. 4.

L862 first appears in Levantine Bronze Age individuals within the ancient reference dataset and is at about 5% frequency in Sardinia today.

We used individual-level qpAdm models to further investigate the presence of these new ancestries (Supp. Data 3). In addition to the original Neolithic Anatolian (Anatolia-N) and Hunter Gatherer (WHG) sources that were sufficient to model ancient Sardinians through the Nuragic period, we fit models with representatives of Steppe (Steppe-EMBA), Neolithic Iranian (Iran-N), and Neolithic North African (Morocco-EN) ancestry as sources. We observe the presence of the Steppe-EMBA (point estimates ranging $0-20\%$) and Iran-N components (point estimates ranging $0-25\%$) in many of the post-Nuragic individuals (Supp. Fig. 14). All six individuals from the Punic Villamar site were inferred to have substantial levels of ancient North African ancestry (point estimates ranging $20-35\%$, Supp. Fig. 14, also see ADMIXTURE and PCA results, Fig. 2.4 and 2.2). When fit with the same 5-way admixture model, present-day Sardinians have a small but detectable level of North African ancestry (Supp. Fig. 14, also see ADMIXTURE analysis, Fig. 2.4).

Models with direct continuity from Villamar to the present are rejected (Tab. 2.2F, Supp. Tab. 6). In contrast, nearly all the other post-Nuragic sites produce viable models as single sources for the modern Sardinians (e.g. Sar-COR qpAdm p-values of 0.16 and 0.261 for Cagliari and Ogliastra respectively; Sar-SNN qpAdm p-values of 0.037 and 0.016, similarly Tab. 2.2F,Supp. Tab. 6). We found some evidence of sub-structure: Sar-ORC002 (from an interior site) is more consistent with being a single source for Ogliastra than Cagliari, whereas Sar-AMC shows an opposite pattern (Supp. Tab. 6).

We also carried out 3-way admixture models for each post-Nuragic Sardinian individual using the Nuragic sample as a source or outgroup, and potential sources from various ancient samples that are representative of different regions of the Mediterranean. We found a range of models can be fit for each individual (Supp. Tab. 7-8). For the models with Nuragic as a source, by varying the proxy populations, one can obtain fitted models that vary widely in

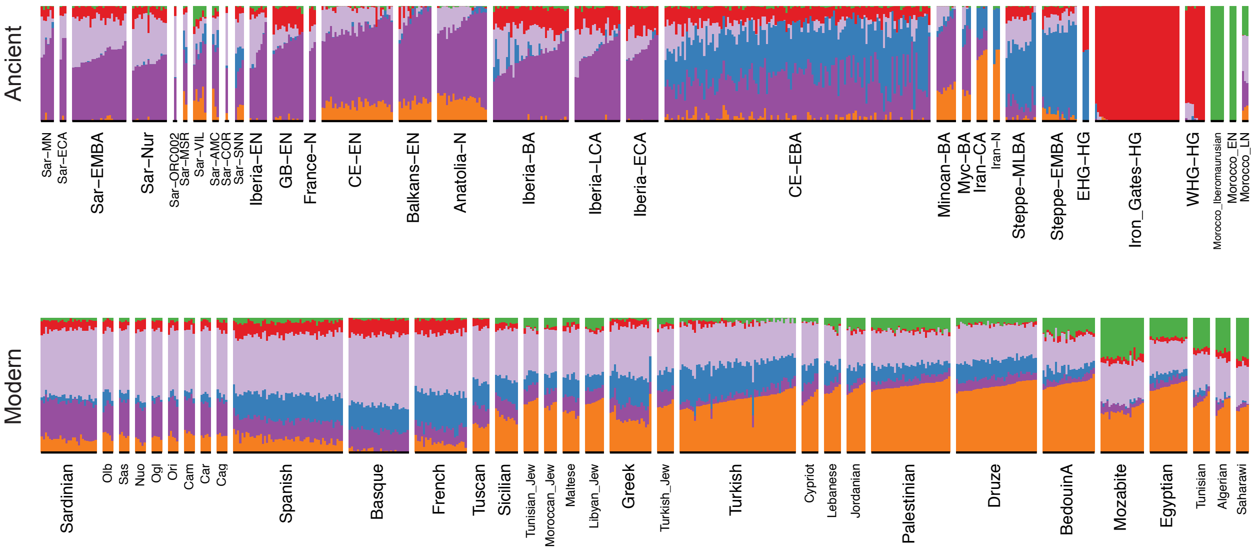


Figure 2.4: **Admixture coefficients estimated by ADMIXTURE** ($K = 6$). Each stacked bar represents one individual and color fractions depict the fraction of the given individual's ancestry coming from a given "cluster". For $K = 6$ (depicted here), Sardinian individuals up until the Nuragic share similar admixture proportions as other western European Neolithic individuals. Present-day as well as most post-Nuragic ancient Sardinian individuals have elevated Steppe-like ancestry (blue), and an additional ancestry component prevalent in Near Eastern / Levant populations (orange). An ancient North - African component (green) appears at low fraction in many present-day Mediterranean populations, and somewhat stronger in samples from the Sardinian Punic site Villamar. ADMIXTURE results for all K=2,...,11 are depicted in the supplement (Supp. Fig. 19)

.

the inferred Nuragic component (e.g. individual COR002 has a range from 4.4% to 87.8% across various fitted models; similarly, individual AMC001, with North African mtDNA haplogroup U6a, had a range form 0.2% to 43.1%, see Supp. Tab. 7-8). The ORC002 sample had the strongest evidence of Nuragic ancestry (range from 62.8% to 96.3%, see Supp. Tab. 7-8). Further, the VIL, MSR, and AMC individuals can be modeled with Nuragic Sardinian individuals included as a source or as an outgroup, while the two COR and ORC002 individuals can only be modeled with Nuragic individuals included as a source. One individual from the medieval period San Nicola Necropoli (SNN001) was distinct in that we found their ancestry can be modeled in a single source model as descendant of a population represented by present-day Basque individuals (Supp. Tab. 8). When we apply the same approach to present-day Sardinian individuals, we find models with the Nuragic sample as an out-group fail in most cases (Supp. Tab. 9). For models that include Nuragic as a possible source, each present-day individual is consistent with a wide range of Nuragic ancestry. The models with the largest p-values return fractions of Nuragic ancestry that are close to, or higher than 50% (Supp. Tab. 9), similar to observed in our population-level modeling (Tab. 2).

2.3.6 Fine-scale structure in contemporary Sardinia

Finally, we assessed our results in the context of spatial substructure within modern Sardinia which suggested elevated levels of WHG and EEF ancestry in Ogliastra (Chiang et al., 2018).

In the PCA of modern west Eurasian and north African variation, the ancient Sardinian individuals are placed closest to individuals from Ogliastra and Nuoro (see Fig. 2.2, Fig. 2.5A). At the same time, in a PCA of just the modern Sardinian sample, the ancient individuals project furthest from Ogliastra (Fig. 2.5B). Interestingly, individual ORC002, dating from the Punic period and from a site in Ogliastra, projects towards Ogliastra individuals relative to other ancient individuals.

Further, in the broad PCA results, the median of the province of Olbia-Tempio (north-

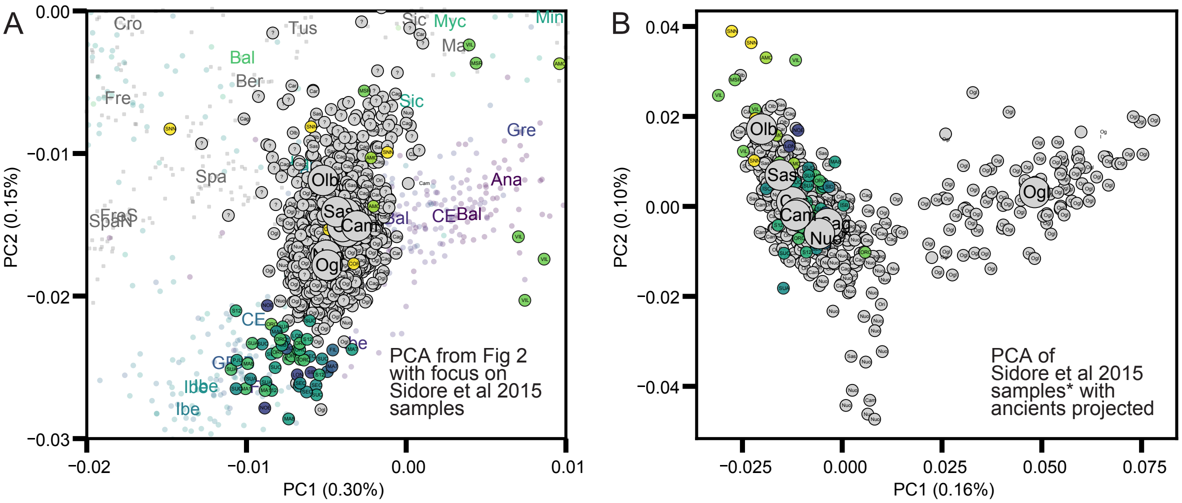


Figure 2.5: **Present-day genetic structure in Sardinia reanalyzed with aDNA.** A: Scatter plot of the first two principal components from Figure 2A with a zoom-in on present-day Sardinia diversity in our sample (Sidore et al., 2015). Median PC values for each Sardinian region are depicted as large circles. B: PCA results based on present-day Sardinian individuals, subsampling Cagliari and Ogliastra to 100 individuals to avoid effects of unbalanced sampling. In both panels, each individual is labeled with an abbreviation that denotes the source location if at least 3 grandparents were born in the same geographical location ("small" three letter abbreviations) or if grand-parental ancestry is missing with question mark. We also projected each ancient Sardinian individual on to the top two PCs (points color-coded by age, see Figure 1 for the color scale).

east Sardinia) is shifted towards mainland populations of southern Europe, and the median for Campidano (southwest Sardinia) shows a slight displacement towards the eastern Mediterranean (Fig. 2.5A). A three-way admixture model fit with qpAdm suggests differential degrees of admixture, with the highest eastern Mediterranean ancestry in the southwest (Carbonia, Campidano) and the highest northern Mediterranean ancestry in the northeast of the island (Olbia, Sassari, Supp. Fig. 17). These observations of sub-structure among contemporary Sardinian individuals contrast our results from the Nuragic and earlier, which forms a relatively tight cluster on the broad PCA 2.2 and for which the top PCs do not show any significant correlations with latitude, longitude, or regional geographic labels after correcting for multiple testing (Supp. Fig. 24-33).

2.4 Discussion

Our analysis of genome-wide data from 70 ancient Sardinian individuals has generated insights regarding the population history of Sardinia and the Mediterranean. First, our analysis provides more refined DNA-based support for the Middle Neolithic of Sardinia being related to the early Neolithic peoples of the Mediterranean coast of Europe. Middle/Late Neolithic Sardinian individuals fit well as a two-way admixture between mainland EEF and WHG sources, similar to other EEF populations of the western Mediterranean. Further, we detected Y haplogroups R1b-V88 and I2-M223 in the majority of the early Sardinian males. Both haplogroups appear earliest in the Balkans among Mesolithic hunter-gatherers and then Neolithic groups (Mathieson et al., 2018) and later in EEF Iberians (Olalde et al., 2019), in which they make up the majority of Y haplogroups, but have not been detected in Neolithic Anatolians or more western WHG individuals. These results are plausible outcomes of substantial gene flow from Neolithic populations that spread westward along the Mediterranean coast of southern Europe around 5,500 BCE (a “Cardial/Impressed” ware expansion, Introduction). We note that we lack autosomal aDNA from earlier than the Middle Neolithic in Sardinia and from key mainland locations such as Italy, which leaves some uncertainty about timing and the relative influence of gene flow from the Italian mainland versus from the north or west. The inferred WHG admixture fraction of Middle Neolithic Sardinians was higher than that of early mainland EEF populations, which could suggest a time lag of the influx into Sardinia (as HG ancestry increased through time on the mainland) but also could result from a pulse of initial local admixture or continued gene flow with the mainland. Genome-wide data from Mesolithic and early Neolithic individuals from Sardinia and potential source populations will help settle these questions.

From the Middle Neolithic onward until the beginning of the first millennium BC, we do not find evidence for gene flow from distinct ancestries into Sardinia. That stability contrasts with many other parts of Europe which had experienced substantial gene flow from central Eurasian Steppe ancestry starting about 3,000 BCE (Haak et al., 2015; Allentoft et al., 2015) and also with many earlier Neolithic and Copper age populations across mainland Europe, where local admixture increased WHG ancestry substantially over time (Lipson et al., 2017). We observed remarkable constancy of WHG ancestry (close to 17%) from the Middle Neolithic to the Nuragic period. While we cannot exclude influx from genetically similar populations (e.g. early Iberian Bell Beakers), the absence of Steppe ancestry suggests genetic isolation from many Bronze Age mainland populations - including later Iberian Bell Beakers (Olalde et al., 2018). As further support, the Y haplogroup R1b-M269, the most frequent present-day western European haplogroup and associated with expansions that brought Steppe ancestry into Britain (Olalde et al., 2018) and Iberia (Olalde et al., 2019) about 2,500-2,000 BCE, remains absent in our Sardinian sample through the Nuragic period (1,200-1,000 BCE). Larger sample sizes from Sardinia and alternate source populations may discover more subtle forms of admixture, but the evidence appears strong that Sardinia was isolated from major mainland Bronze Age gene flow events through to the local Nuragic period . As the archaeological record shows that Sardinia was part of a broad Mediterranean trade network during this period (Tykot, 1996), such trade was either not coupled with gene flow or was only among proximal populations of similar genetic ancestry. In particular, we find that the Nuragic period is not marked by shifts in ancestry, arguing against hypotheses that the design of the Nuragic stone towers was brought with an influx of people from eastern sources such as Mycaeaneans.

Following the Nuragic period, we found evidence of gene flow with both northern and eastern Mediterranean sources. We observed eastern Mediterranean ancestry appearing first in two Phoenician-Punic sites (Monte Sirai, Villamar). The northern Mediterranean ancestry became prevalent later, exemplified most clearly by individuals from a north-western Medieval site (San Nicola Necropoli). Many of the post-Nuragic individuals could be modeled as direct immigrants or offspring from new arrivals to Sardinia, while others had higher fractions of local Nuragic ancestry (Corona Moltana, ORC002). Substantial uncertainty exists here as the low differentiation among plausible source populations makes it challenging to exclude alternate models, especially when using individual-level analysis. Overall though, we find support for increased variation in ancestry after the Nuragic period, and this echoes other recent aDNA studies in the Mediterranean that have observed fine-scale local heterogeneity in the Iron Age and later (Olalde et al., 2019; Fernandes et al., 2019; Feldman et al., 2019; Antonio et al., 2019).

In addition, we found present-day Sardinian individuals sit within the broad range of ancestry observed in our ancient samples. A similar pattern is seen in Iberia (Olalde et al., 2019) and central Italy (Antonio et al., 2019), where variation in individual ancestry increased markedly in the Iron Age, and later decreased until present-day. In terms of the fine-scale structure within Sardinia, we note the median position of modern individuals from the central regions of Ogliastra and Nuoro on the main PCA (Fig. 2.5A) are less shifted towards novel sources of post-Nuragic admixture, which reinforces a previous result that Ogliastra shows higher levels of EEF and HG ancestry than other regions (Chiang et al., 2018). At the same time, in the PCA of within Sardinia variation (Fig. 2.5B), differentiation of Ogliastra from other regions and other ancient individuals is apparent, likely reflecting a recent history of isolation and drift. The northern provinces of Olbia-Tempio, and to a lesser degree Sassari, appear to have received more northern Mediterranean immigration after the Bronze Age; while the southwestern provinces of Campidano and Carbonia carry more eastern Mediterranean ancestry. Both of these results align with known history: the major Phoenician and Punic settlements in the first millennium BCE were situated principally along the south and west coasts, and Corsican shepherds, speaking an Italian-Corsican dialect (Gallurese), immigrated to the northeastern part of Sardinia (Le Lannou, 1941).

Our inference of gene flow after the second millennium BCE seems to contradict previous models emphasizing Sardinian isolation (Haak et al., 2015). These models were supported by admixture tests that failed to detect substantial admixture (Chiang et al., 2018), likely because of substantial drift and a lack of a suitable proxy for the Nuragic Sardinian ancestry component. However, compared to other European populations (Sarno et al., 2017; Lazaridis et al., 2017), we confirm Sardinia experienced relative genetic isolation through the Bronze Age/Nuragic period. Additionally, we find that subsequent admixture appears to derive mainly from Mediterranean sources that have relatively little Steppe ancestry. Consequently, present-day Sardinian individuals have retained an exceptionally high degree of EEF ancestry and so they still cluster with several mainland European Copper Age individuals such as Ötzi (Sikora et al., 2014), even as they are shifted from ancient Sardinian individuals of a similar time period (Fig. 2.2).

The Basque people, another population high in EEF ancestry, were previously suggested to share a genetic connection with modern Sardinian individuals (Günther et al., 2015; Chiang et al., 2018). We observed a similar signal, with modern Basque having, of all modern samples, the largest pairwise outgroup-f3 with most ancient and modern Sardinian groups (Fig. 2.3). While both populations have received some immigration, seemingly from different sources (e.g., Fig. 2.4, (Olalde et al., 2019)), our results support that the shared EEF ancestry component could explain their genetic affinity despite their geographic separation.

Beyond our focal interest in Sardinia, the results from individuals from the Phoenician-Punic sites Monte Sirai and Villamar shed some light on the ancestry of a historically impactful Mediterranean population. Notably, they show strong genetic relationships to ancient North African and eastern Mediterranean sources. These results mirror other emerging ancient DNA studies (Zalloua et al., 2018; Matisoo-Smith et al., 2018), and are not unexpected given that the Punic center of Carthage on the North African coast itself has roots in the eastern Mediterranean. Interestingly, the Monte Sirai individuals, predating the Villamar individuals by several centuries, show less North African ancestry. This could be because they harbor earlier Phoenician ancestry and North African admixture may have been unique to the later Punic context, or because they were individuals from a different ancestral background altogether. Estimated North African admixture fractions were much lower in later ancient individuals and present-day Sardinian individuals, in line with previous studies that have observed small but significant African admixture in several present-day South European populations, including Sardinia (Hellenthal et al., 2014; Loh et al., 2013; Chiang et al., 2018).

As ancient DNA studies grow, a key challenge will be fine-scale sampling to aid the interpretation of shifts in ancestry. Our sample from Sardinia's post-Nuragic period highlights the complexity, as we simultaneously observe examples of individuals that appear as novel immigrant ancestries (e.g. from Villamar and San Nicola) and of individuals that look more contiguous to the past and to the present (e.g., the two Corona a Moltana siblings, the ORC002 individual, several of the Alghero Monte Carru individuals). This variation is likely driven by differential patterns of contact — as might arise between coastal versus interior villages, central trading centers versus remote agricultural sites, or even between neighborhoods and social strata in the same village. We also note that modern populations are collected with different biases than ancient individuals (e.g. the sub-populations sampled by medical genetics projects (Sidore et al., 2015) versus the sub-populations that are accessible at archaeological sites). As such, caution should be exercised when generalizing from the sparse sampling typical for many aDNA studies, including this one.

With these caveats in mind, we find that genome-wide ancient DNA provides unique insights into the population history of Sardinia. Our results are consistent with gene flow being minimal or only with genetically similar populations from the Middle Neolithic until the late Bronze Age. In particular, the onset of the Nuragic period is not being characterized by influx of a distinct ancestry. The data also link Sardinia from the Iron Age onwards to the broader Mediterranean in what seems to have been a period of new dynamic contact throughout much of the Mediterranean. A parallel study focusing on islands of the western Mediterranean provides generally consistent results and both studies make clear the need to add complexity to simple models of sustained isolation that have dominated the genetic literature on Sardinia (Fernandes et al., 2019). Finally, our results suggest some of the current sub-structure seen on the island (e.g. Ogliastra) has emerged due to recent genetic drift. Overall, the history of isolation, migration, and genetic drift on the island has given rise to an unique constellation of allele frequencies, and illuminating this history will help future efforts to understand genetic-disease variants prevalent in Sardinia and throughout the Mediterranean, such as beta-thalassemia and G6PD deficiency.

2.5 Methods

2.5.1 Archaeological sampling

The archaeological samples used in this project derive from several collection avenues. The first was a sampling effort led by co-author Luca Lai, leveraging a broad base of samples from different existing collections in Sardinia, a subset of which were previously used in isotopic analyses to understand dietary composition and change in prehistoric Sardinia (Lai et al., 2013). The second was from the Seulo Caves project (Skeates et al., 2013), an on-going project on a series of caves that span the Middle Neolithic to late Bronze Age near the town of Seulo. The project focuses on the diverse forms and uses of caves in the prehistoric culture of Sardinia. The Neolithic individuals from Sassari province as well as the post-Nuragic individuals were collected from several co-authors as indicated in Supplemental Information Section 1. The third was a pair of Neolithic sites Noedalle and S'isterridolzu (Germanà, 1980). The fourth are a collection of post-Nuragic sites spanning from the Phoenician to the Medieval time. All samples were handled in collaboration with local scientists and with the approval of the local Sardinian authorities for the handling of archaeological samples (Ministero per i Beni e le Attività Culturali, Direzione Generale per i beni Archeologici, request dated 11 August 2009; Soprintendenza per i Beni Archeologici per le province di Sassari e Nuoro, prot. 12993 dated 20 Dec. 2012; Soprintendenza per i Beni Archeologici per le province di Sassari e Nuoro, prot. 10831 dated 27 Oct. 2014; Soprintendenza per i Beni Archeologici per le province di Sassari e Nuoro, prot. 12278 dated 05 Dec. 2014; Soprintendenza per i Beni Archeologici per le Provincie di Cagliari e Oristano, prot. 62, dated 08 Jan 2015; Soprintendenza Archeologia, Belle Arti e Paesaggio per le Provincie di Sassari, Olbia-Tempio e Nuoro, prot. 4247 dated 14 March 2017; Soprintendenza per i Beni Archeologici per le Provincie di Sassari e Nuoro, prot. 12930 dated 30 Dec. 2014; Soprintendenza Archeologia, Belle arti e Paesaggio per le Provincie di Sassari e Nuoro, prot. 7378 dated 9 May, 2017; Soprintendenza per i Beni Archeologici per le Provincie di Cagliari e Oristano, prot. 20587, dated 05 Oct. 2017; Soprintendenza Archeologia, Belle Arti e Paesaggio per le Provincie di Sassari e Nuoro, prot. 15796 dated 25 October, 2017; Soprintendenza Archeologia, Belle Arti e Paesaggio per le Provincie di Sassari e Nuoro, prot. 16258 dated 26 Nov. 2017; Soprintendenza per i Beni Archeologici per le province di Sassari e Nuoro, prot. 5833 dated 16 May 2018; Soprintendenza Archeologia, Belle Arti e

Paesaggio per la città metropolitana di Cagliari e le province di Oristano e Sud Sardegna, prot. 30918 dated 10 Dec 2019). For more, detailed description of the sites please see Supplemental Information Section 1.

2.5.2 Initial sample screening and sequencing

The ancient DNA (aDNA) workflow was implemented in dedicated facilities at the Palaeogenetic Laboratory of the University of Tübingen and at the Department of Archaeogenetics of the Max Planck Institute for the Science of Human History in Jena. The only exception was for four samples from the Seulo Cave Project which had DNA isolated at the Australian Centre for Ancient DNA and capture and sequencing carried out in the Reich lab at Harvard University. Different skeletal elements were sampled using a dentist drill to generate bone and tooth powder respectively. DNA was extracted following an established aDNA protocol (Dabney et al., 2013) and then converted into double-stranded libraries retaining (Meyer and Kircher, 2010) or partially reducing (Rohland et al., 2015) the typical aDNA substitution pattern resulting from deaminated cytosines that accumulate towards the molecule's termini. After indexing PCR (Meyer and Kircher, 2010) and differential amplification cycles, the DNA was shotgun sequenced on Illumina platforms. Samples showing sufficient aDNA preservation where captured for mtDNA and ≈ 1.24 million SNPs across the human genome chosen to intersect with the Affymetrix Human Origins array and Illumina 610-Quad array (Fu et al., 2015). The resulting enriched libraries were also sequenced on Illumina machines in single-end or paired-end mode. Sequenced data were pre-processed using the `EAGER` pipeline (Peltzer et al., 2016). Specifically, DNA adapters were trimmed using `AdapterRemoval` v2 (Schubert et al., 2016) and paired-end sequenced libraries were merged. Sequence alignment to the mtDNA (RSRS) and nuclear (hg19) reference genomes was performed with `BWA` (Li and Durbin, 2009) (parameters –n 0.01, seeding disabled), duplicates were removed with `DeDup` (Peltzer et al., 2016) and a mapping quality filter was applied (MQ$\geqslant$ 30). For genetic sexing, we compared relative X and Y-chromosome coverage to the autosomal coverage with a custom script. For males, nuclear contamination levels were estimated based on heterozygosity on the X-chromosome with the software `ANGSD` (Korneliussen et al., 2014).

After applying several standard ancient DNA quality control metrics, retaining individuals with endogenous DNA content in shotgun sequencing >0.2%, mtDNA contamination <4% (average 1.6%) and nuclear contamination <6% (average 1.1%) and after inspection of contamination patterns (Supp. Fig. 2-5), we generated genotype calls for downstream population genetic analyses for a set of 70 individuals. To account for sequencing errors we first removed any reads that overlapped a SNP on the capture array with a base quality score less than 20. We also removed the last 3-bp on both sides of every read to reduce the effect of DNA damage on the resulting genotype calls (Al-Asadi et al., 2018). We used custom python scripts (`https://github.com/mathii/gdc3`) to generate pseudo-haploid genotypes by sampling a random read for each SNP on the capture array and setting the genotype to be homozygous for the sampled allele. We then screened for first degree relatives using a pairwise relatedness statistic, and identified one pair of siblings and one parent-offspring pair within our sample (Supp. Fig. 12).

2.5.3 Processing of mtDNA data

Data originating from mtDNA capture was processed with `schmutzi` (Renaud et al., 2015), which jointly estimates mtDNA contamination and reconstructs mtDNA consensus sequences that were assigned to the corresponding mtDNA haplogroups using `Haplofind` (Vianello et al., 2013) (Supp. Data 1C). The consensus sequences were also compared with rCRS (Andrews et al., 1999) to build a phylogenetic tree of ancient Sardinian mitogenomes (Supp. Data 1D) using a maximum parsimony approach with the software mtPhyl (`http://eltsov.org/mtphyl.aspx`). We assigned haplogroups following the nomenclature proposed by the PhyloTree database build 17 (`http://www.phylotree.org`) (Van Oven and Kayser, 2009) and for Sardinian Specific Haplogroups (Olivieri et al., 2017).

2.5.4 Inference of Y haplogroups

To determine the haplotype branch of the Y chromosome of male ancient individuals, we analyzed informative SNPs on the Y-haplotype tree. For reference, we used markers from `https://isogg.org/tree` (Version: 13.238, 2018). We merged this data with our set of calls and identified markers available in both to create groups of equivalent markers for sub-haplogroups. Our targeted sequencing approach yielded read count data for up to 32,681 such Y-linked markers per individual. As the conventions for naming of haplogroups are subject to change, we annotated them in terms of carrying the derived state at a defining SNP. We analyzed the number of derived and ancestral calls for each informative marker for all ancient Sardinian individuals and reanalyzed male ancient West Eurasians in our reference data set. Refined haplotype calls were based on manual inspection of ancestral and derived read counts per haplogroup, factoring in coverage and error estimates.

2.5.5 Merging newly generated data with published data

Ancient DNA datasets from Western Eurasia and North Africa: We downloaded and processed BAM files from several ancient datasets from continental Europe and the Middle-east (Mathieson et al., 2015; Lazaridis et al., 2016, 2017; Mathieson et al., 2018; Lipson et al., 2017; Olalde et al., 2018). To minimize technology-specific batch effects in genotype calls and thus downstream population genetic inference, we focused on previously published ancient samples that had undergone the capture protocol on the same set of SNPs targeted in our study. We processed these samples through the same pipeline and filters described above, resulting in a reference dataset of 972 ancient samples. Throughout our analysis, we used a subset of n =1,088,482 variants that was created by removing SNPs missing in more than 90% of all ancients individuals (Sardinian and reference dataset) with at least 60% of all captured SNPs covered.

This ancient dataset spans a wide geographic distribution and temporal range. Ancient individuals are associated with a variety of different cultures, which provides rich context for interpreting downstream results. Our reference ancient dataset is comprised of many individuals sampled from a particular geographic locale, such as Germany or Hungary, in a transect of multiple cultural changes through time (Fig. 2.2). For the PCA (Fig. 2.2), we additionally included a single low-coverage ancient individual (label "Pun") dated to 361-178 BCE from a Punic necropolis on the west Mediterranean island of Ibiza (Zalloua et al., 2018). We merged individuals into groups (Supp. Data 1E,F). For ancient samples, these groups were chosen manually, trying to strike a balance between reducing overlap in the PCA and keeping culturally distinct populations separate. We used geographic location to first broadly group samples into geographic areas (such as Iberia, Central Europe and Balkans), and then further annotated each of these groups by different time periods.

Contemporary DNA datasets from Western Eurasia and North Africa.: We downloaded and processed the Human Origins dataset to characterize a subset of Eurasian and north African human genetic diversity at 594,924 autosomal SNPs (Lazaridis et al., 2014). We focused on a subset of 837 individuals from Western Eurasia and north Africa.

Contemporary DNA dataset from Sardinia.: We merged in a whole-genome sequence Sardinian dataset (1,577 individuals (Chiang et al., 2018)) and called genotypes on the Human Origin autosomal SNPs to create a dataset similar to the other modern reference populations. For analyses on province level, we used a subset where at least three grandparents originate from the same geographical location and grouped individuals accordingly (Fig. 2.2C, n=1,085 in total).

2.5.6 Principal Components Analysis

We performed Principal Components Analysis (PCA) on two large-scale datasets of modern genotypes from Western Eurasia and North Africa (837 individuals from the Human Origins dataset) and Sardinia (1,577 individuals from the SardiNIA project). For both datasets, we normalized the genotype matrix by mean-centering and scaling the genotypes at each SNP using the inverse of the square-root of heterozygosity (Patterson et al., 2006). We additionally filtered out rare variants with minor allele frequency ($p_{\text{min}} < 0.05$).

To assess population structure in the ancient individuals, we projected them onto the pre-computed principal axes using only the non-missing SNPs via a least-squares approach, and corrected for the shrinkage effect observed in high-dimensional PC score prediction (Lee et al., 2010) (see Supp. Note 7, Supp. Fig. 23).

We also projected a number of out-sample sub-populations from Sardinia onto our PCs. Reassuringly, these out-of-sample Sardinian individuals project very close to Humans Origins Sardinian individuals (Fig. 2.2). Moreover, the test-set Sardinia individuals with grand-parental ancestry from Southern Italy cluster with reference individuals with ancestry from Sicily (not shown).

2.5.7 ADMIXTURE Analysis

We applied `ADMIXTURE` to an un-normalized genotype matrix of ancient and modern samples (Alexander et al., 2009). `ADMIXTURE` is a maximum-likelihood based method for fitting the Pritchard, Stephens and Donnelly model (Pritchard et al., 2000) using sequential quadratic programming. We first LD pruned the data matrix based off the modern Western Eurasian and North African genotypes, using `plink1.9` with parameters [`--indep-pairwise 200 25 0.4`]. We then ran 5 replicates of `ADMIXTURE` for values of $K = 2, \ldots, 11$. We display results for the replicate that reached the highest log-likelihood after the algorithm converged (Supp. Fig. 19-22).

2.5.8 Estimation of f-statistics

We measured similarity between groups of individuals through computing an outgroup-f_3 statistic (Patterson et al., 2012) using the `scikit-allel` packages's function `average_patterson_f3`, `http://doi.org/10.5281/zenodo.3238280`). The outgroup-f_3 statistic can be interpreted as a measure of the internal branch length of a three-taxa population phylogeny and thus does not depend on genetic drift or systematic error in one of the populations that are being compared (Patterson et al., 2012).

We used the ancestral allelic states as an outgroup, inferred from a multi-species alignment from Ensembl Compara release 59, as annotated in the 1000 Genomes Phase3 sites `vcf` (1000 Genomes Project Consortium et al., 2015). We fixed the ancestral allele counts to $n = 10^6$ to avoid finite sample size correction when calculating outgroup f_3.

The f_3- and f_4-statistics that test for admixture were computed with `scikit-allel` using the functions `average_patterson_f3` and `average_patterson_d` that implement standard estimators of these statistics (Patterson et al., 2012). We estimated standard errors with a block jack-knife over 1000 markers (`blen=1000`). For all f-statistics calculations, we analyzed only one allele of ancient individuals that were represented as pseudo-haploid genotypes to avoid an artificial appearance of genetic drift - that could for instance mask a negative f_3 signal of admixture.

2.5.9 Estimation of F_{ST}-coefficients

To measure pairwise genetic differentiation between two populations [remark=R2.3](rather than shared drift from an outgroup as the out-group f_3 statistic does), we estimated average pairwise F_{ST} and its standard error via block-jackknife over 1000 markers, using `average_patterson_fst` from the package `scikit-allel`. When analyzing ancient individuals that were represented as pseudo-haploid genotypes, we analyzed only one allele. For this analysis, we removed first degree relatives within each population. Another estimator, `average_hudson_fst` gave highly correlated results ($r^2 = 0.89$), differing mostly for populations with very low sample size ($n \leqslant 5$) and did not change any qualitative conclusions.

2.5.10 Estimation of admixture proportions and model testing with qpAdm

We estimated admixture fractions of a selected target population as well as model consistency for models with one to up to five source populations as implemented in `qpAdm` (version 810), which relates a set of "left" populations (the population of interest and candidate ancestral sources) to a set of "right" populations (diverse out-groups) (Haak et al., 2015). To assess the robustness of our results to the choice of right populations, we ran one analysis with a previously used set of modern populations as outgroup (Haak et al., 2015), and another analysis with a set of ancient Europeans that have been previously used to disentangle divergent strains of ancestry present in Europe (Lazaridis et al., 2017). In the same `qpAdm` framework, we use a likelihood-ratio test (LRT) to assess whether a specific reduced-rank model, representing a particular admixture scenario, can be rejected in favor of a maximal rank ("saturated") model for the matrix of f_4-values (Haak et al., 2015). We report p-values under the approximation that the LRT statistic is χ^2 distributed with degrees of freedom determined by the number of "left" and "right" populations used in the f_4 calculation and by the rank implied by the number of admixture components. The p-values we report are not corrected for multiple testing. Formal correction is difficult as the tests are highly correlated due to shared population data used across them; informally, motivated by a Bonferroni correction of a nominal 0.01 p-value with 10 independent tests, we suggest only taking low p-values ($< 10^{-3}$) to represent significant evidence to reject a proposed model. The full qpAdm results are discussed in Supp. Note 5.

2.6 Data Availability

The aligned sequences from the data generated in this study are available through the European Nucleotide Archive (ENA, accession number PRJEB35094). Processed read counts and pseudo-haploid genotypes are available via the European Variation Archive (EVA, accession number PRJEB36033) in variant call format (VCF). The contemporary Sardinia data used to support this study have allele frequency summary data deposited to EGA (accession number EGAS00001002212). The disaggregated individual-level sequence data (n=1,577) used in this study is a subset of 2,105 samples (adult volunteers of the SardiNIA cohort longitudinal study) from Sidore et al (2015) and are available from dbGAP under project identifier phs000313 (v4.p2). The remaining individual-level sequence data originate from a case-control study of autoimmunity from across Sardinia, and per the obtained consent and local IRB, these data are available for collaboration by request from the project leader (Francesco Cucca, Consiglio Nazionale delle Ricerche, Italy).

2.7 Code Availability

The code used to process the raw-reads and create the figures in this manuscript can be found at `https://github.com/NovembreLab/ancient-sardinia`. The code to perform bias correction in predicting out of sample PC scores is publicly available `https://github.com/jhmarcus/pcshrink`.

2.8 Acknowledgements

We thank Maanasa Rhagavan for in-depth feedback on drafts, Anna Di Rienzo and Goncalo Abecasis for helpful discussions, and Magdalena Zoledziewska for useful comments and early assistance. We would like to thank Antje Wissgott, Cäcilia Freund and other members of the wet laboratory and computationally teams at MPI-SHH in Jena. We thank Nadin

Rohland, Éadaoin Harney, Shop Mallick, and Alan Cooper for contributing to generating the data for the four samples processed at the Australian Centre for Ancient DNA and in D.R.'s ancient DNA laboratory. We also thank Dan Rice and members of the Novembre lab for helpful discussion and feedback. IRGB-CNR would like to thank the Consortium SA CORONA ARRUBIA DELLA MARMILLA for making available equipment and scientific instruments within the program "Laboratori Dna del Museo Naturalistico del Territorio – Giovanni Pusceddu". This study was supported in part by the Max Planck Society, the University of Sassari, the National Science Foundation via fellowships DGE-1746045 to J.H.M and DGE-1644869 to T.A.J. and HOMINID grant BCS-1032255 to D.R., the National Institute of General Medical Sciences via training grant T32GM007197 support for J.H.M. and grant RO1GM132383 to J.N., the National Human Genome Research Institute via grant R01HG007089 to J.N., the Intramural Research Program of the National Institute on Aging via contracts N01-AG-1-2109 and HHSN271201100005C to F.C., Fondazione di Sardegna via grants U1301.2015/AI.1157 BE Prat. 2015-1651 to F.C., the Australian Research Council via grant DP130102158 to W.H., the Howard Hughes Medical Institute (D.R), the University of Pavia INROAd Program (A.O.), the Italian Ministry of Education, University and Research (MIUR) Dipartimenti di Eccellenza Program (2018–2022, A.O.), the Fondazione Cariplo via project 2018-2045 to A.O., the British Academy (R.S.), and the American Society of Prehistoric Research (P.v.D.).

2.9 Author contributions

We annotate author contributions using the CRediT Taxonomy labels (`https://casrai.org/credit/`). Where multiple individuals serve in the same role, the degree of contribution is specified as 'lead', 'equal', or 'supporting'.

- Conceptualization (Design of study) – lead: FC, JN, JK, LL; supporting: CS, CP, DS, JHM, GA

- Investigation (Collection of skeletal samples) – lead: LL, RS; supporting: JB, MGG, CDS, CP, VM, EP, CM, ALF, DRo, MG, RPO, NT, PVD, SR, PM, RB, RMS, PB (minor contribution from CS, JN)

- Investigation (Ancient DNA isolation and sequencing) – lead: CP, AF, RR, MM; supporting: CDS, WH, JK, DRe*

- Data Curation (Data quality control and initial analysis) – lead: JHM, CP, HR; supporting: CS, CC, KD, HA, AO

- Formal Analysis (General population genetics) – lead: JHM, HR; supporting: TAJ, CL

- Writing (original draft preparation) – lead: JHM, HR, JN; supporting: CP, RS, LL, FC, PVD

- Writing (review and editing) – input from all authors*

- Supervision – equal: FC, JK, JN

- Funding acquisition – lead: JK, FC, JN; supporting: RS

*: D.R. contributed data for four samples and reviewed the description of the data generation for these samples. As he is also senior author on a separate manuscript that reports data on a non-overlapping set of ancient Sardinians and his group and ours wished to keep the two studies intellectually independent, he did not review the entire manuscript until after it was accepted.

2.10 Supplementary Information

2.10.1 Supplementary Note 2: Validating quality and contamination of aDNA

Joseph H. Marcus, Kushal Dey, Hussein Al-asadi, Harald Ringbauer, Cosimo Posth

Postmortem Damage Filtering

Individuals with high levels of mtDNA or X chromosome based contamination estimates were removed in our main analysis. However, population genetic analyses could still be affected by more subtle modern contamination. To assess this possibility, we filter out reads that do not show a signature of post-mortem damage (pmd), as reads that carry a damage signature are less likely to be introduced by modern contamination (Skoglund et al., 2014b). We used `pmdtools` (`https://github.com/pontussk/PMDtools`) to compute a likelihood-based damage score for each read and subsequently removed reads which showed little evidence of being damaged. We then generated pseudo-haploid genotype calls on these "pmd-filtered" individuals and projected them onto the PCs computed in the modern west Eurasian and north African individuals from the Human Origins dataset, as described in the Materials and Methods. We corrected for the regression towards the mean effect in high-dimensional PCA using a simple jackknife estimator (see Supp. Info. 7).

We found that all the samples we analyzed in the main results and that have enough covered SNPs after pmd filtering show little difference between the pmd-filtered and corresponding unfiltered PC scores (Fig. 2.6). This observation supports our sample filtering criteria, and as such our population genetic analyses are unlikely to be strongly affected by contamination. A few individuals had too few covered SNPs after pmd filtering to make accurate predictions of their ancestry, which possibly explains larger observed differences between the pmd-filtered and the original projection.

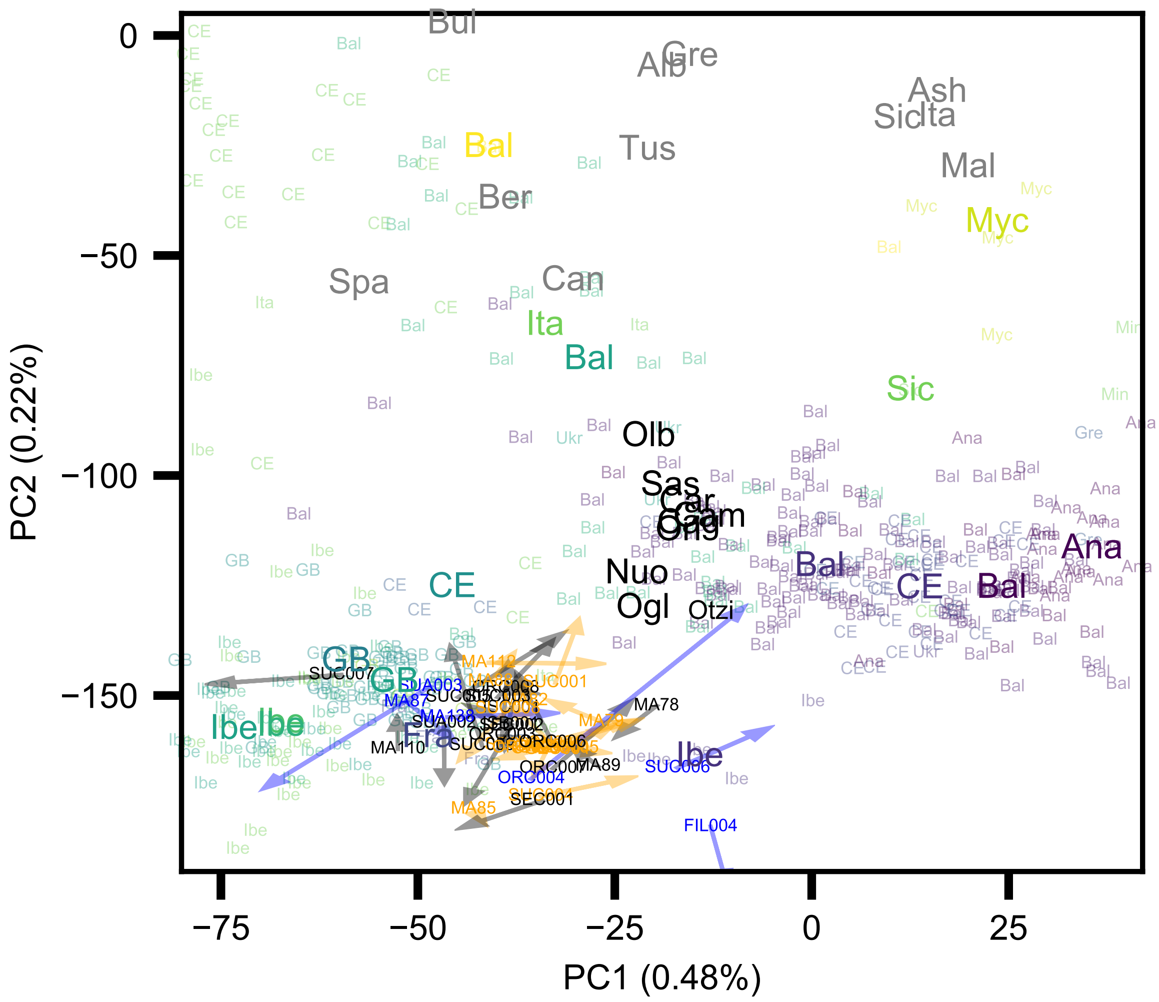


Figure 2.6: **Impact of PMD filtering on PCA projections of ancient individuals.** The figure shows a visualization of PC1 and PC2 computed on modern individuals and projecting ancient individuals from our study, alongside ancients from previously published literature. Each arrow represents an ancient Sardinian individual, where the head of the arrow is the "pmd-filtered" projection and the tail is the "non-pmd-filtered" projection. We color each arrow given the following criteria: black for male individuals that had low X-based contamination estimates (<=0.05); orange for male individuals with high X-based contamination estimates (>0.05) or females; and blue for any remaining individuals with less than 35 thousand covered SNPs after PMD filtering.

aRchaic

We estimate DNA damage profiles using `aRchaic` (Al-Asadi et al., 2018). In the `aRchaic` model, mismatches are counted across all of an individual's sequence reads with corresponding measured features, including mismatch type, mismatch position on the read, flanking reference nucleotides, and strand orientation. Each mismatch is modeled as originating from a mixture of K profiles defined by these features and adaptively learned during inference via an EM algorithm. Maximum-likelihood estimates of mixture proportions for each individual are then displayed on a stacked bar chart where each row is a different sample and each colored bar represents the proportion of the ith individual's mismatches coming from the kth mismatch profile 2.7. We applied `aRchaic` to the ancient Sardinia data for $K = 2$, $K = 3$, and $K = 4$. In (Fig. 2.7, Fig. 2.8, Fig. 2.9), we observe the typical pattern representative of ancient DNA, an enrichment of cytosine to thymine mismatches occurring preferentially at the ends of the read (Ginolhac et al., 2011; Jónsson et al., 2013). These observations helps to authenticate our data as being truly ancient. Ancient and modern individuals show distinctive mismatch profiles and as we increase $K = 2$ to $K = 4$ finer resolution substructure is revealed between subgroups of individuals. Specifically, samples treated with a protocol, UDGhalf treatment, to partially remove some of this damage signature from each sample, cluster distinctly from both untreated ancient and modern individuals. The modern individuals we included are a sub-sample of 43 low-coverage whole genome sequences from the 1000 Genomes Project Phase3 (PUR, FIN, CHS, GBR, CDX, MXL) and 7 deeply sequenced individuals from the Simons Genome Diversity project. The modern individuals show membership in two damage profile clusters. In previous analyses (Al-Asadi et al., 2018), such results have been found to arise from differing sample preparation protocols, and other plausible factors (for example differences in initial DNA quality). Importantly, none of the ancient samples show membership in the damage profile clusters found in the modern samples. This result is concordant with the low contamination estimates obtained for these samples (note: the samples analyzed with `aRchaic` here all passed our initial contamination rate filters based on mtDNA and nuclear contamination estimates, see Materials and Methods).

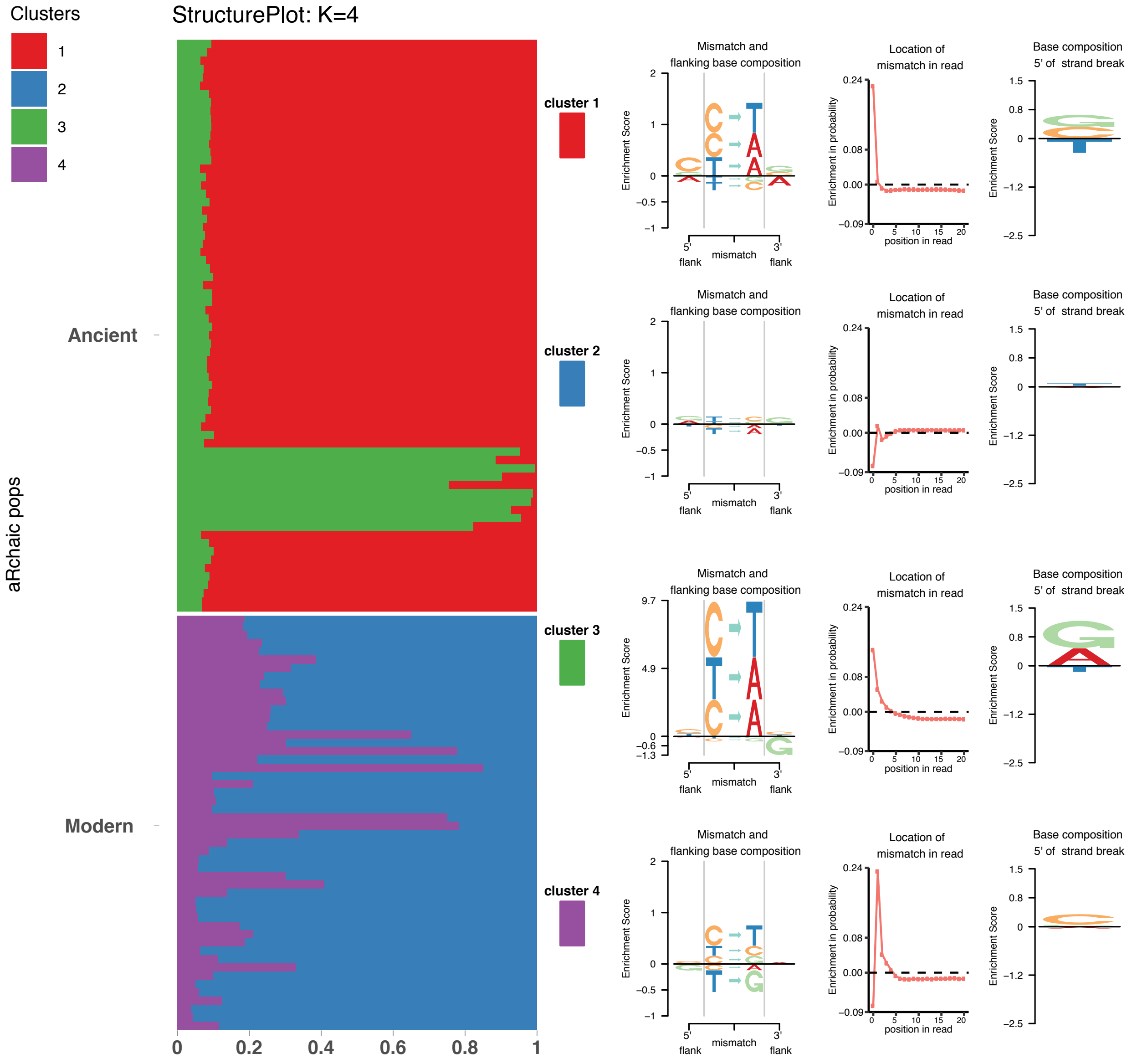


Figure 2.7: **Results of the package `aRchaic` for clustering mismatch profiles in sequence read libraries for (K=4).** On the left we plot a stacked bar-chart where each row represents a `bam` file and the colored portions of each bar represent the mixture proportion for a given cluster. As we can see each `bam` file's mixture proportions must be non-negative and sum to one. On the right we display representations of the inferred latent variables that define each cluster. The left most plot displays the enrichment or depletion of different mismatch types, the middle plot displays the enrichment probability of observing a mismatch at a particular position along the read, and finally the right hand plot displays an enrichment score for the mismatch type observed at a strand break on the 5' end of the fragment. All together these plots visualize both how the `bam` files are loaded on to each cluster as well as what defines each cluster.

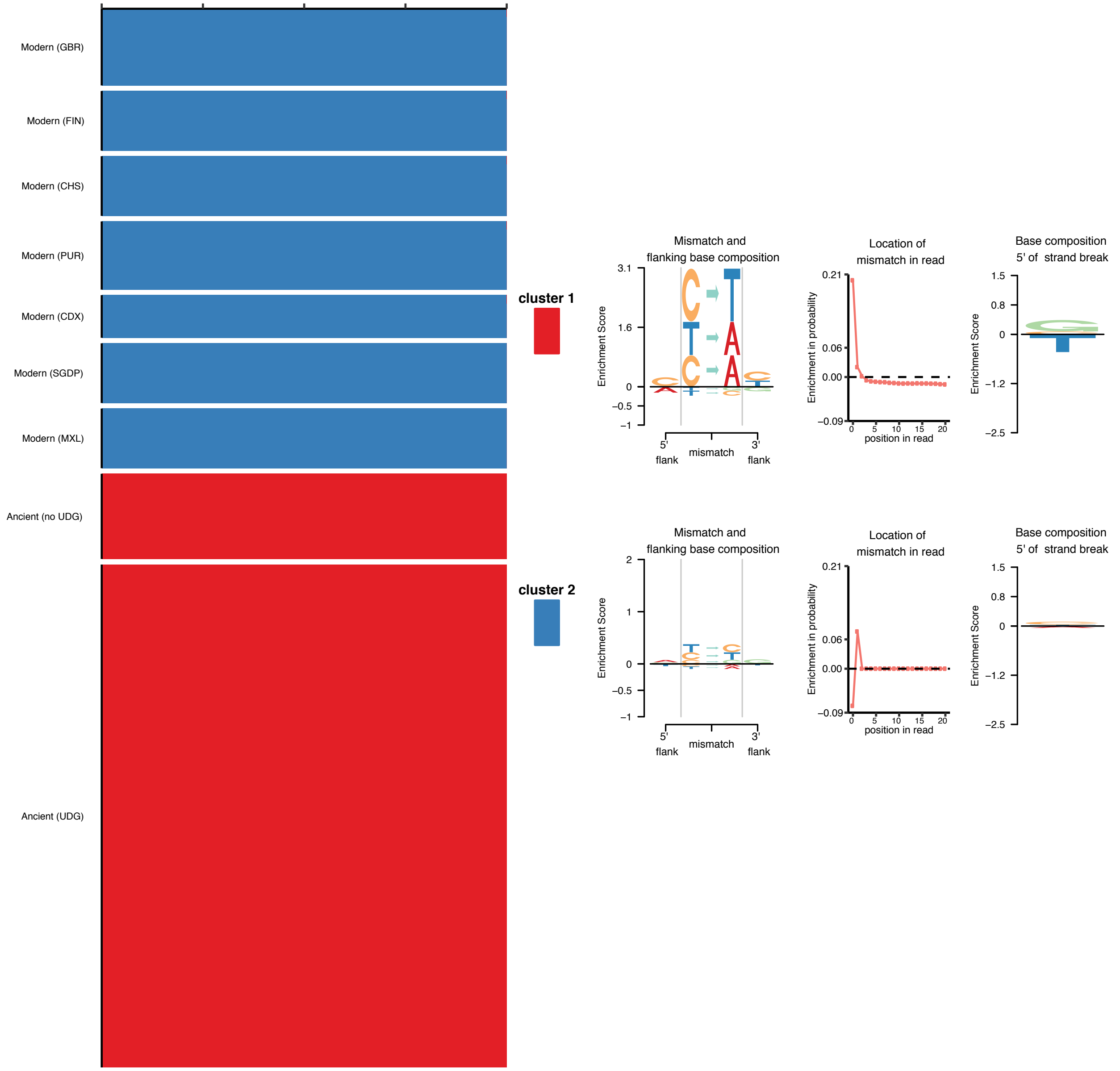


Figure 2.8:
Results of the package aRchaic (K=2) for clustering mismatch profiles in sequence read libraries. See (Fig. 2.7) for a detailed description of the panels.

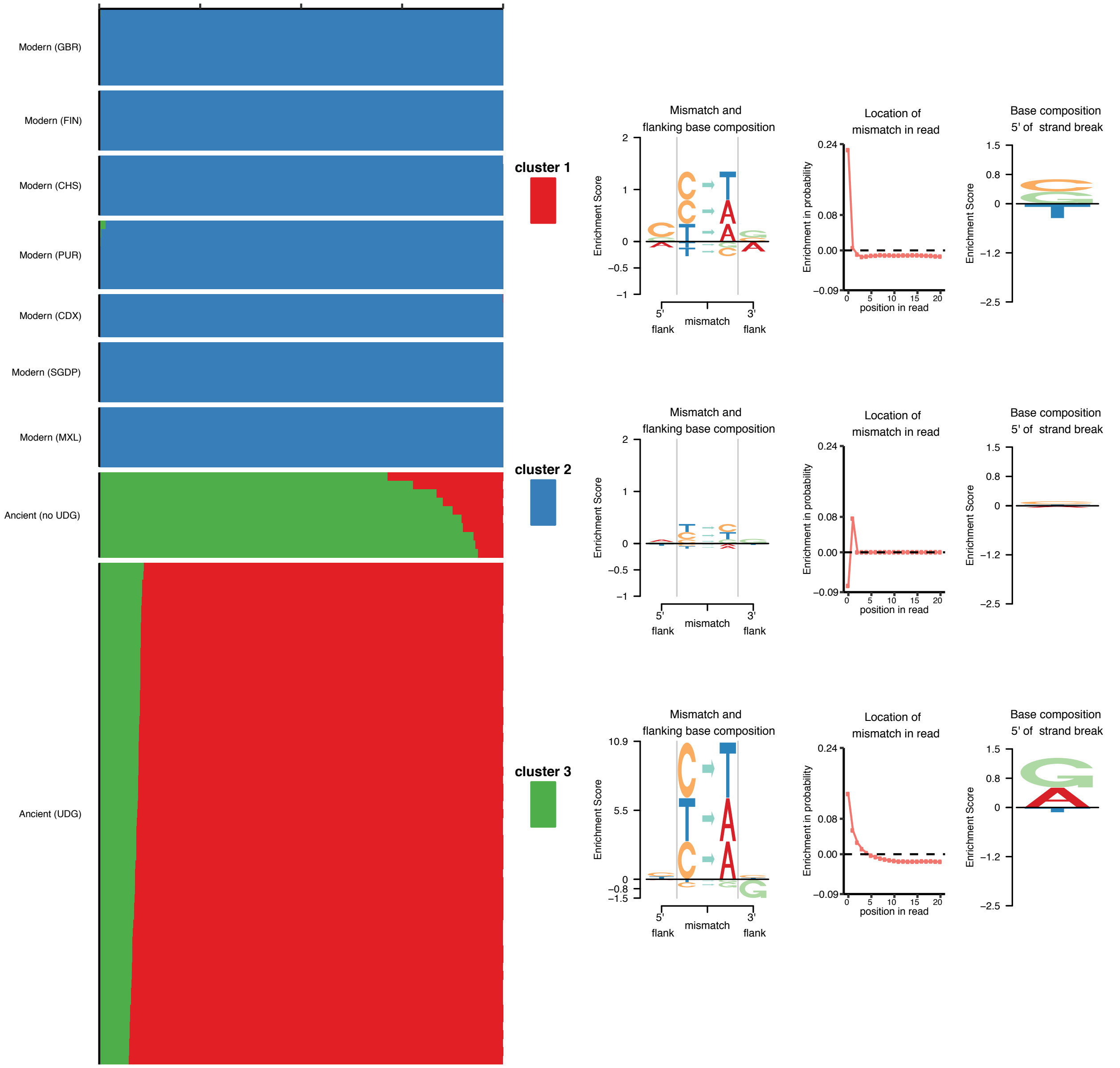


Figure 2.9:
Results of the package `aRchaic` **(K=3) for clustering mismatch profiles in sequence read libraries.** See (Fig. 2.7) for a detailed description of the panels.

2.10.2 Supplementary Note 4: Pairwise similarity statistics

Harald Ringbauer, Joseph H. Marcus

Measures of pairwise genetic differentiation.

Supp. Fig. 2.10 and Supp. Fig. 2.11 depict the matrix of genetic similarities calculated using f_3-outgroup statistics and F_{ST} as described in the main text (Materials and Methods). Numerical values are reported in Supp. Data 2A and B.

Pairwise Relatedness.

To identify close relatives within our dataset of ancient Sardinians, we directly assessed pairwise relatedness. We first filtered to markers that have calls in at least 20 of the ancient Sardinian individuals, and within those for markers with (pseudo-haploid) minor allele frequency (MAF) > 0.2. For the resulting set of $n = 351,967$ markers, we calculated pairwise correlation of allelic state (relatedness) between individuals i and j:

$$f(i,j) = \frac{(p_i - \bar{p}) \cdot (p_j - \bar{p})}{\bar{p} \cdot (1 - \bar{p})}, \tag{2.1}$$

averaged over all markers with available (pseudo-haploid) calls for both individuals. Mean allele frequencies ($\bar{p}$) were calculated from the full set of ancient Sardinians. Estimation of the allele frequency from the sample itself as well as population structure can have the effect that this pairwise correlation deviates from 0 even for unrelated pairs of individuals. To increase interpretability, we here centered this statistic by subtracting the mean of all pairwise values.

Using this basic approach enabled us to identify two pairs of first-degree relatives (expected $f(i,j) = 0.25$). The remainder of ancient Sardinian samples likely do not contain any first or second degree relatives (Supp. Fig. 2.12).

The first pair consists of a female and male sample, SUC002 and SUC003, both sampled from the Su Crucifissu Mannu site. The broadly uniform value of estimated $f(i,j)$ around 0.25 throughout the genome (Supp. Fig. 2.12) suggests that these two samples are a parent-offspring pair, as full siblings would vary between $f = 0$, 0.25 and 0.5, depending on whether 0, 1 or 2 allele were co-inherited. Both samples had identical mtDNA haplogroup J1c3, providing some evidence that these pair of samples represent a mother and son.

The second pair consists of a female and male sample, COR001 and COR002, both sampled from the Corona Moltana site. The estimated value of $f(i,j)$ centered around 0.25 with a broad range ranging from $0 - 0.5$ throughout the genome (Supp. Fig. 2.12), suggesting that these two samples are a full-sibling pair. Both samples carry an identical mtDNA haplogroup, K1b1a1.

In addition, we detected three genetic duplicates of three of our individuals (ISB001, LON003, MA82) by identifying pairs which had $f(i,j)$ close to 0.5 (results not shown). Identical uni-parentally inherited haplotypes and sampling location, close proximity on the two-dimensional PCA, as well as overlapping radiocarbon ages corroborated this result. We therefore combined the three identical pairs into three single samples, and merged their reads for all subsequent analysis and reported data.

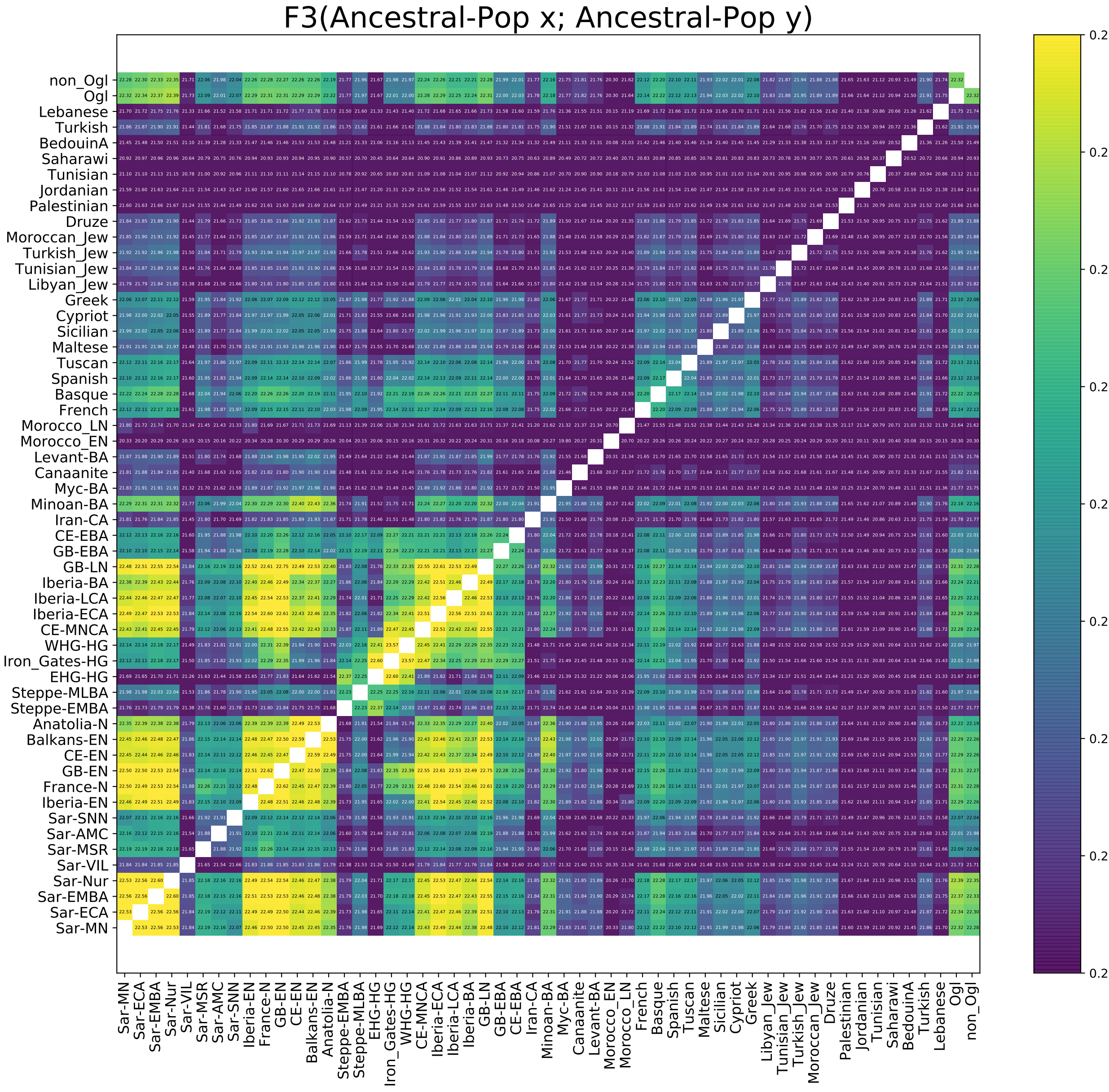


Figure 2.10: **Matrix of f3-outgroup "shared genetic drift" metrics of pairwise similarity.** Populations are ordered broadly by period and geography (See Sup. Data 1G for legend to abbreviations). For post-Nuragic Sardinians, we group individuals by sample site. The two individuals from Corona-Moltana are not shown, as only two first-degree relatives are available.

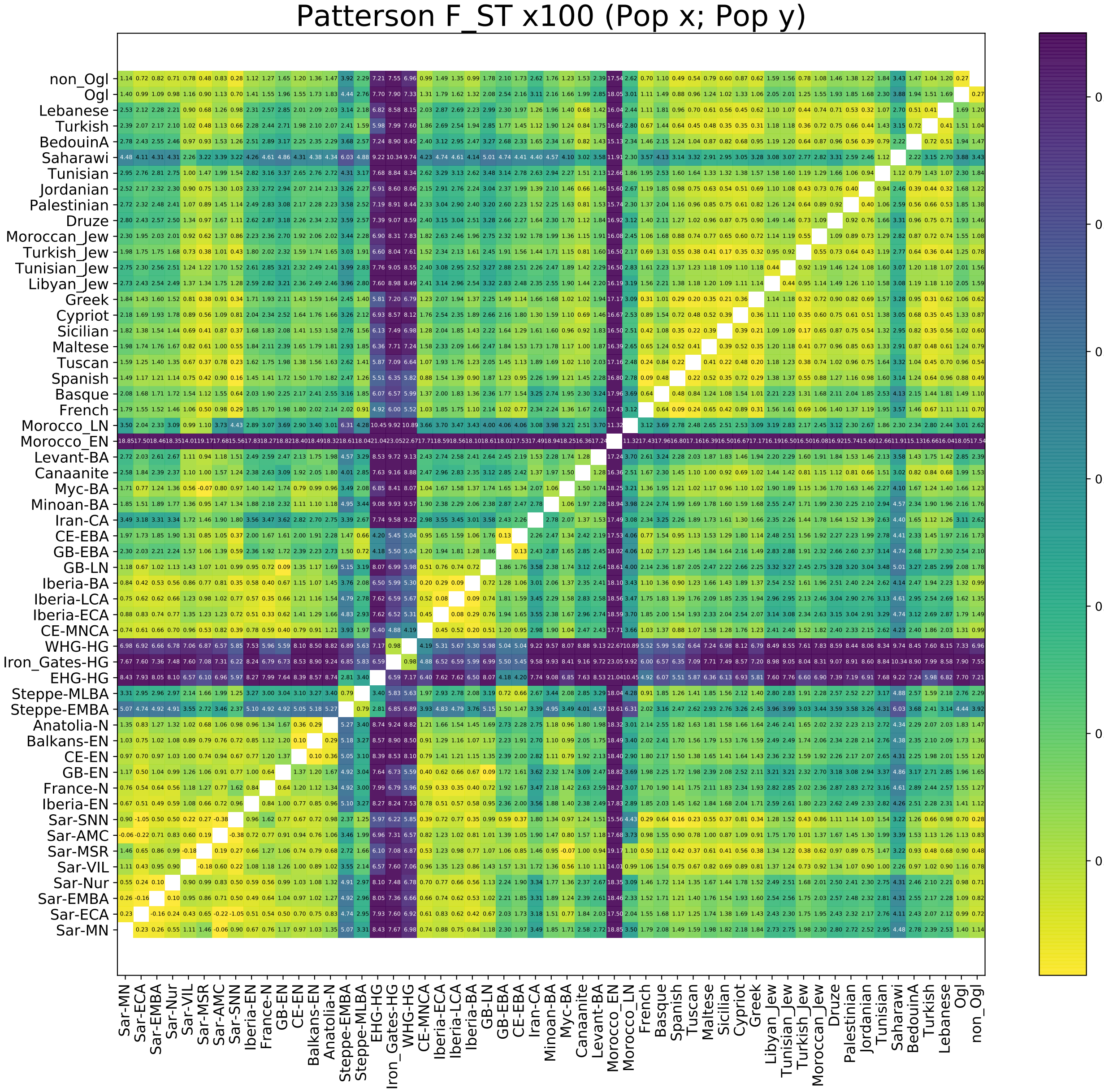


Figure 2.11: **Matrix of F_{ST} metrics of pairwise differentiation**. Populations are ordered broadly by period and geography (See Sup. Data 1G for legend to abbreviations). For post-Nuragic Sardinians, we group individuals by sample site. These sites have typically low number of individuals and low coverage. This increases estimation variance and explains the sometimes negative values (which are possible for an unbiased F_{ST} estimators such as the one used here). The two individuals from Corona-Moltana are not shown, as only two first-degree relatives are available.

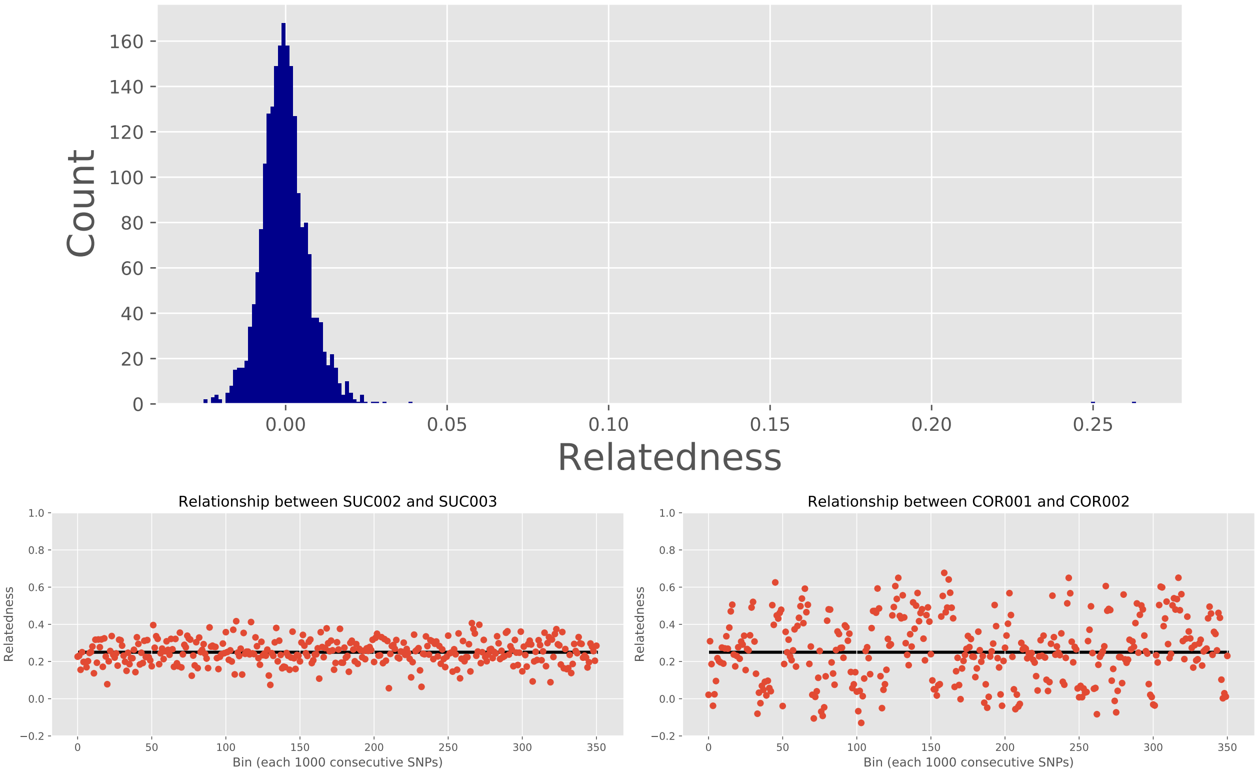


Figure 2.12: **Pairwise Relatedness estimates in ancient Sardinians.** Upper panel: Histogram of all pairwise relatedness estimates for ancient Sardinians, plotted for all pairs with more than $10,000$ intersecting called SNPs. Only two pairs of samples have significantly elevated relatedness (SUC002/SUC003 and COR001/COR002, counts seen at just larger than 0.25). Lower panel: Estimated $f(i,j)$ for these two putatively related sample pairs, calculated genome-wide using bins of 1000 consecutive SNPs ordered along the reference genome.

2.10.3 Supplementary Note 6: Population Structure Models

Joseph H. Marcus and Tyler A. Joseph

Here we describe extended results applying variants of the Pritchard, Stephens, and Donnelly model (Pritchard et al., 2000) to the dataset of ancient individuals and modern individuals from west Eurasia and north Africa.

ADMIXTURE

We applied `ADMIXTURE` to a joint dataset of ancient and modern samples, as described in the Materials and Methods (Alexander et al., 2009). In (Fig. 2.13, Fig. 2.14) we display a gallery plot, i.e. the typical stacked bar plot for $K = 2$ through 8. For each K, we run 5 replicates of `ADMIXTURE` and plot the the runs reaching the highest log likelihood.

DyStruct

We compared the results from `ADMIXTURE` to a time-aware population structure model: `DyStruct` (Joseph and Pe'er, 2019a). `DyStruct` implements a novel variational inference algorithm based on the Pritchard, Stephens, Donnelly model (Pritchard et al., 2000) that incorporates fluctuations in allele frequencies due to differences in sample times. Specifically, `DyStruct` defines a normal approximation to genetic drift that serves as a prior for allele frequency estimates for different time points. At each time point the model is equivalent to the PSD model, but allele frequency estimates between time points are regularized by the prior to ensure allele frequencies estimated from samples nearby in time are closer than allele frequencies from samples further apart. This corrects for genetic drift in populations between samples, potentially leading to different conclusions than `ADMIXTURE`.

We applied `DyStruct` to an un-normalized genotype matrix of ancient and modern samples. Sample times were converted to generation times assuming a 25 year generation time,

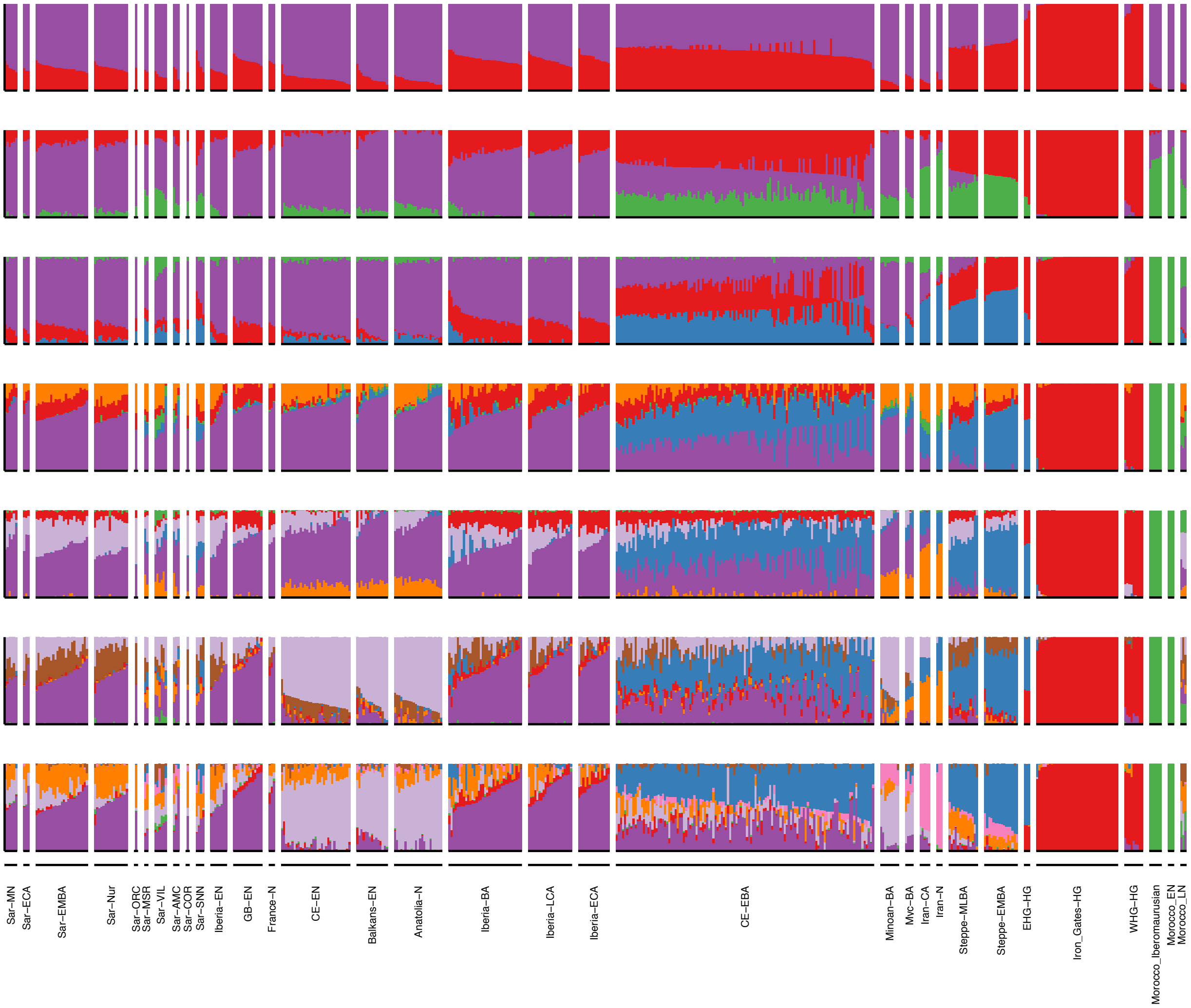


Figure 2.13: **Visualization of admixture coefficients estimated by ADMIXTURE for a subset of ancient individuals**. Each row corresponds to the highest likelihood run of five replicates of ADMIXTURE on both ancient and modern individuals from $K = 2$ to $K = 8$. Here we display admixture coefficients for just a subset ancient individuals.

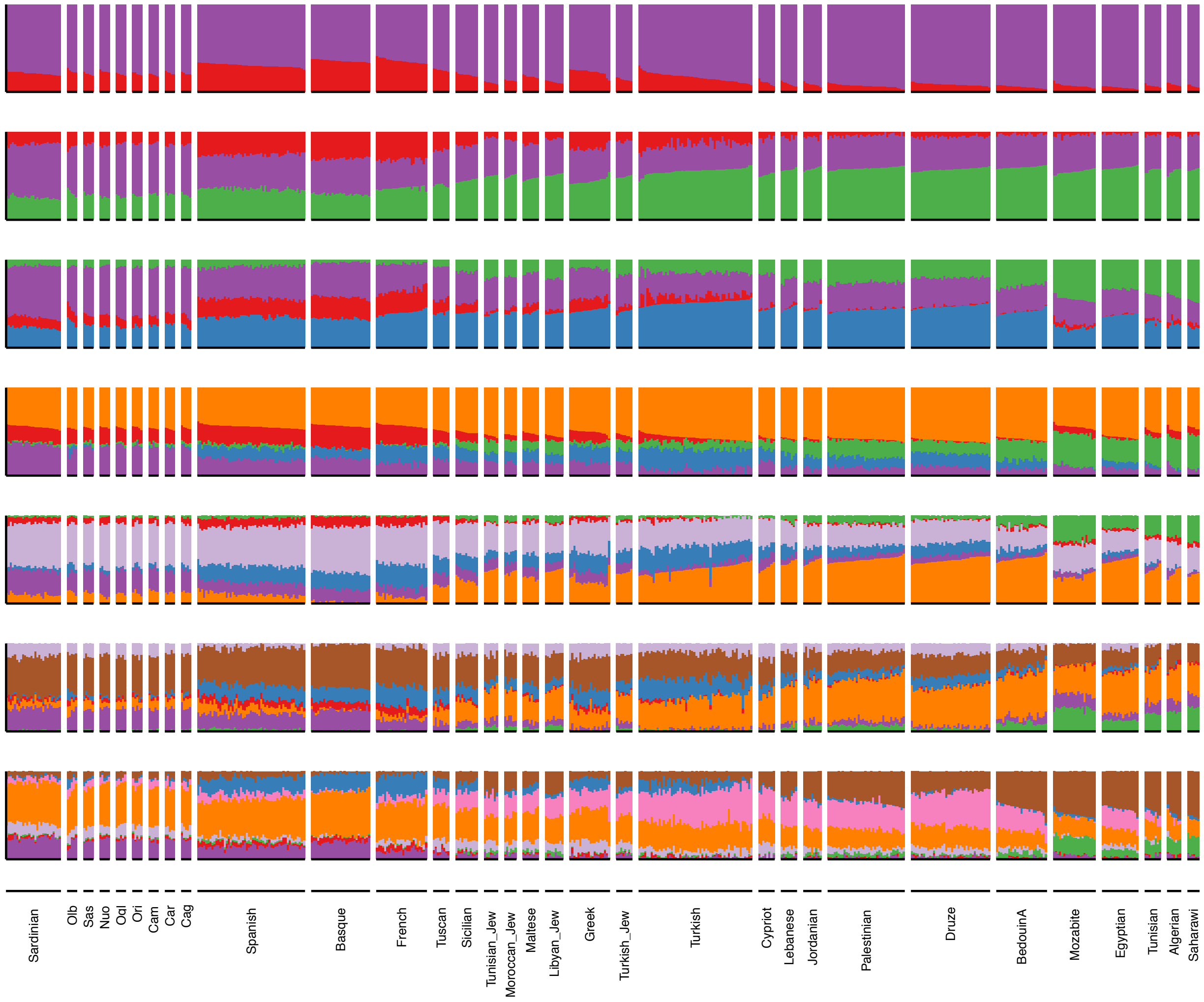


Figure 2.14: **Visualization of admixture coefficients estimated by ADMIXTURE for a subset of modern individuals**. Each row corresponds to the highest likelihood run of five replicates of ADMIXTURE on both ancient and modern individuals from $K = 2$ to $K = 8$. Here we display admixture coefficients for just a subset modern individuals.

and provided as input to DyStruct. We used default prior settings and set the effective population size hyper-parameter to 15000. (Fig. 2.15, Fig. 2.16) displays the fitted admixture coefficients for $K = 2$ to $K = 8$. Qualitatively, DyStruct appears to place emphasis on explaining modern populations as mixtures of ancient populations by assigning singular clusters to ancient samples, and describing modern samples as mixtures of these ancient clusters. Hence, ancient samples in DyStruct appear as more "extreme" versions of their cluster assignments in ADMIXTURE. Consequently, estimates of the genetic contribution from ancient samples into modern populations are different between both models. For instance, modern Sardinian individuals in DyStruct appear to inherit a larger fraction of early European farmer ancestry, Steppe/EHG ancestry instead of WHG ancestry, and a smaller portion of shared ancestry from Neolithic Iran and Neolithic Levant.

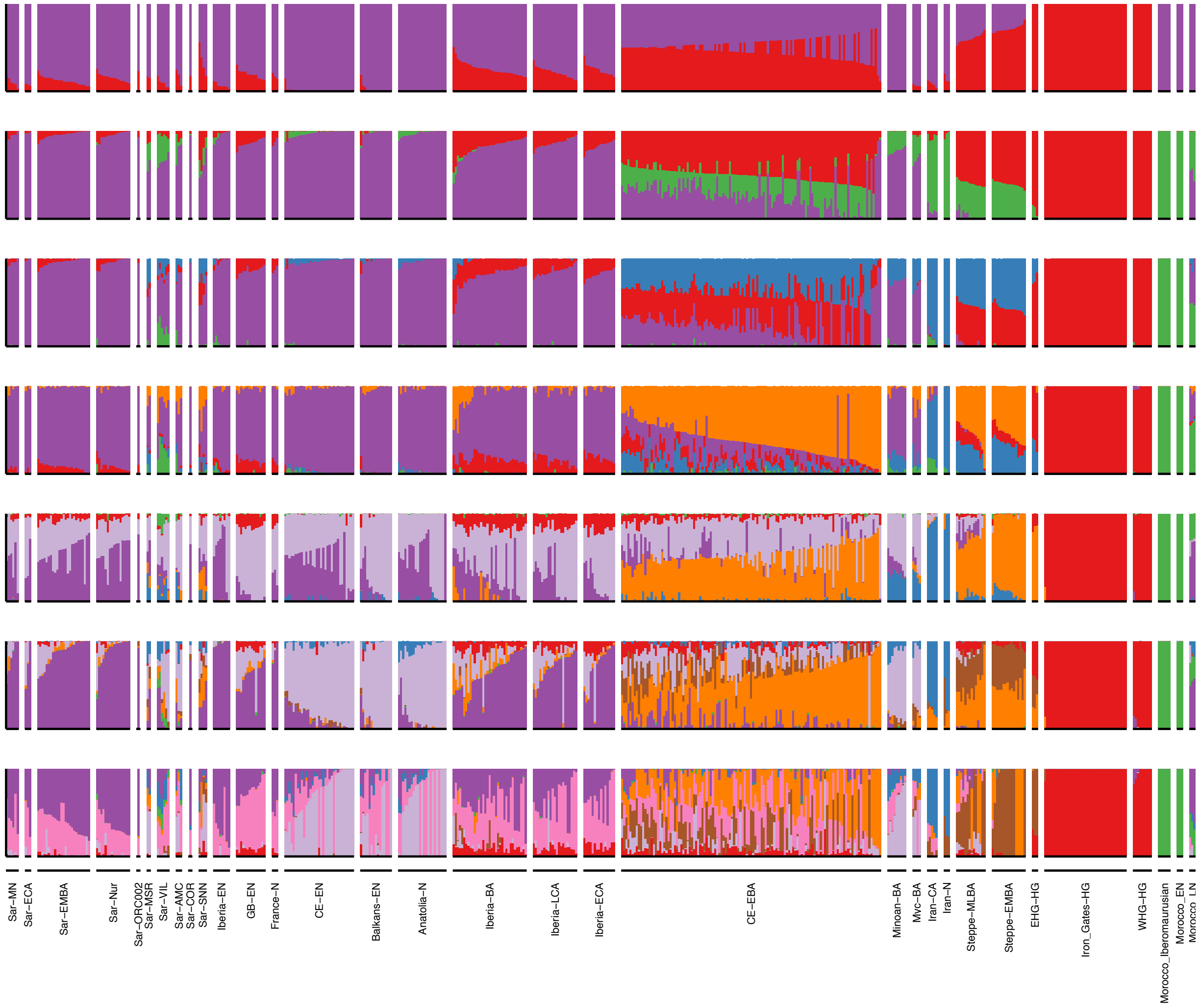


Figure 2.15: **Visualization of admixture coefficients estimated by DyStruct on a subset of ancient individuals**. Each row corresponds to a run of DyStruct on both ancient and modern individuals from $K = 2$ to $K = 8$. Here we display admixture coefficients for a subset ancient individuals.

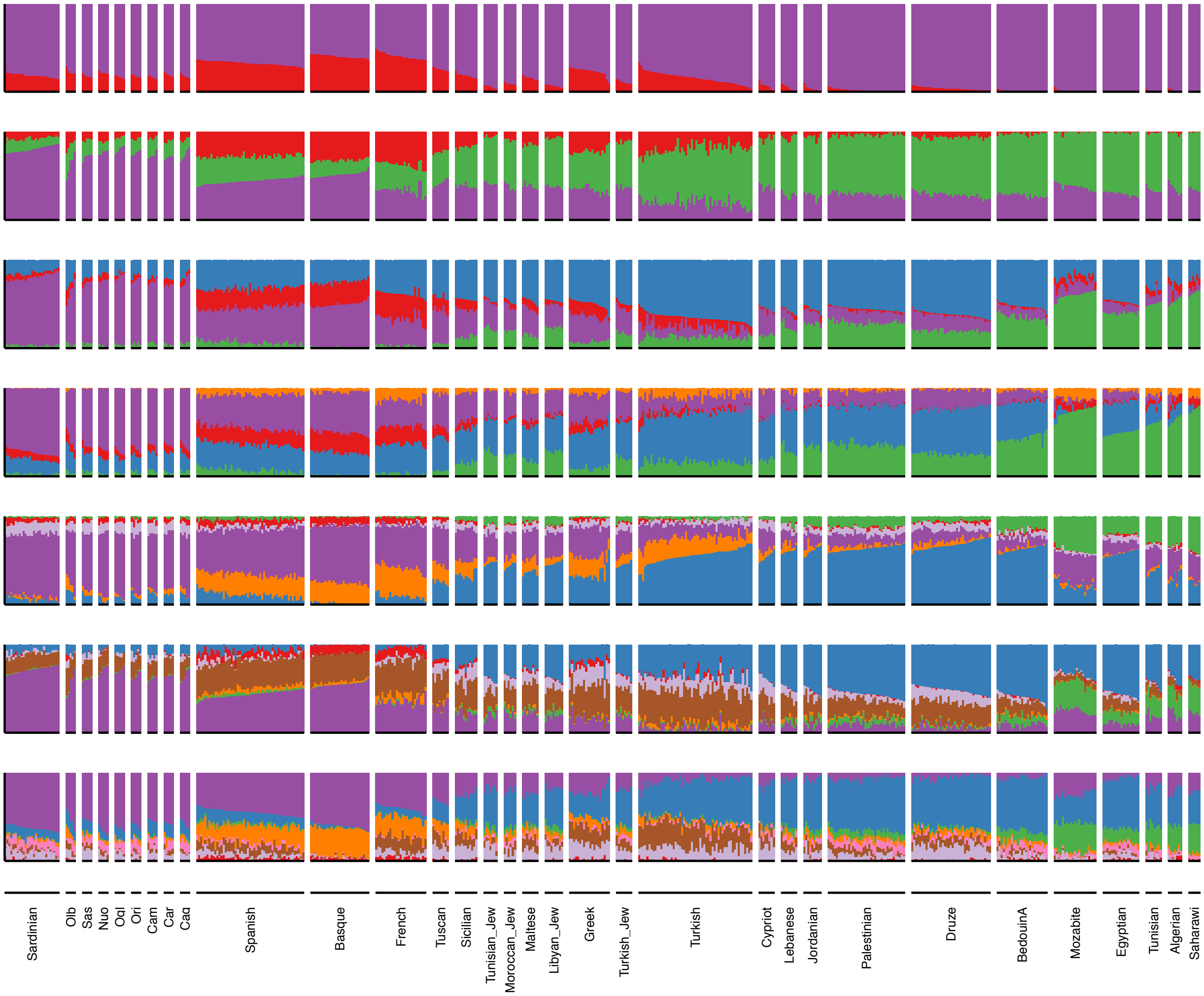


Figure 2.16: **Visualization of admixture coefficients estimated by DyStruct on a subset of modern individuals**. Each row corresponds to a run of DyStruct on both ancient and modern individuals from $K = 2$ to $K = 8$. Here we display admixture coefficients for a subset of modern individuals.

2.10.4 Supplementary Note 7: Shrinkage correction in PC score prediction

Joseph Marcus

In Fig. 2 of the main text, we perform principal components analysis (PCA) on contemporary west Eurasian and north African individuals and project each ancient individual onto modern PCs, one at a time, by solving a simple least squares problem. It is known that the estimated principal scores are biased and exhibit a regression towards the mean effect (shrinkage towards 0 if the data is mean centered) for high dimensional data i.e. when the number of features (SNPs) is much greater than the number of samples (individuals) (Lee et al., 2010; Wang et al., 2015; Liu et al., 2018). To correct for this shrinkage effect when predicting PC scores for out of sample individuals, we implemented a shrinkage correction factor through a jackknife re-sampling approach proposed previously (Lee et al., 2010) (for computational experiments see `https://github.com/jhmarcus/pcshrink/blob/master/notebook/patterson-example.ipynb`).

The procedure was performed through the following steps: (1) We compute a rank-K truncated SVD on the full dataset to obtain a first set of uncorrected PC scores. (2) We remove each individual from the dataset and compute a rank-K truncated SVD on the remaining individuals (3) We project the held-out individual on to the PCA computed from the dataset of step (2). Using the eigenvectors computed for each individual, we then constructed a jackknife estimator of the bias. We then applied this correction factor to the ancient individuals' PC scores to create our final visualization. For comparison we also applied two correction procedures, “shrinkmode” and “autoshrink”, implemented in `smartpca` (Patterson et al., 2006). In Supp. Fig. 2.17, we see no major qualitative differences between the corrected ancient PC scores for the three correction approaches.

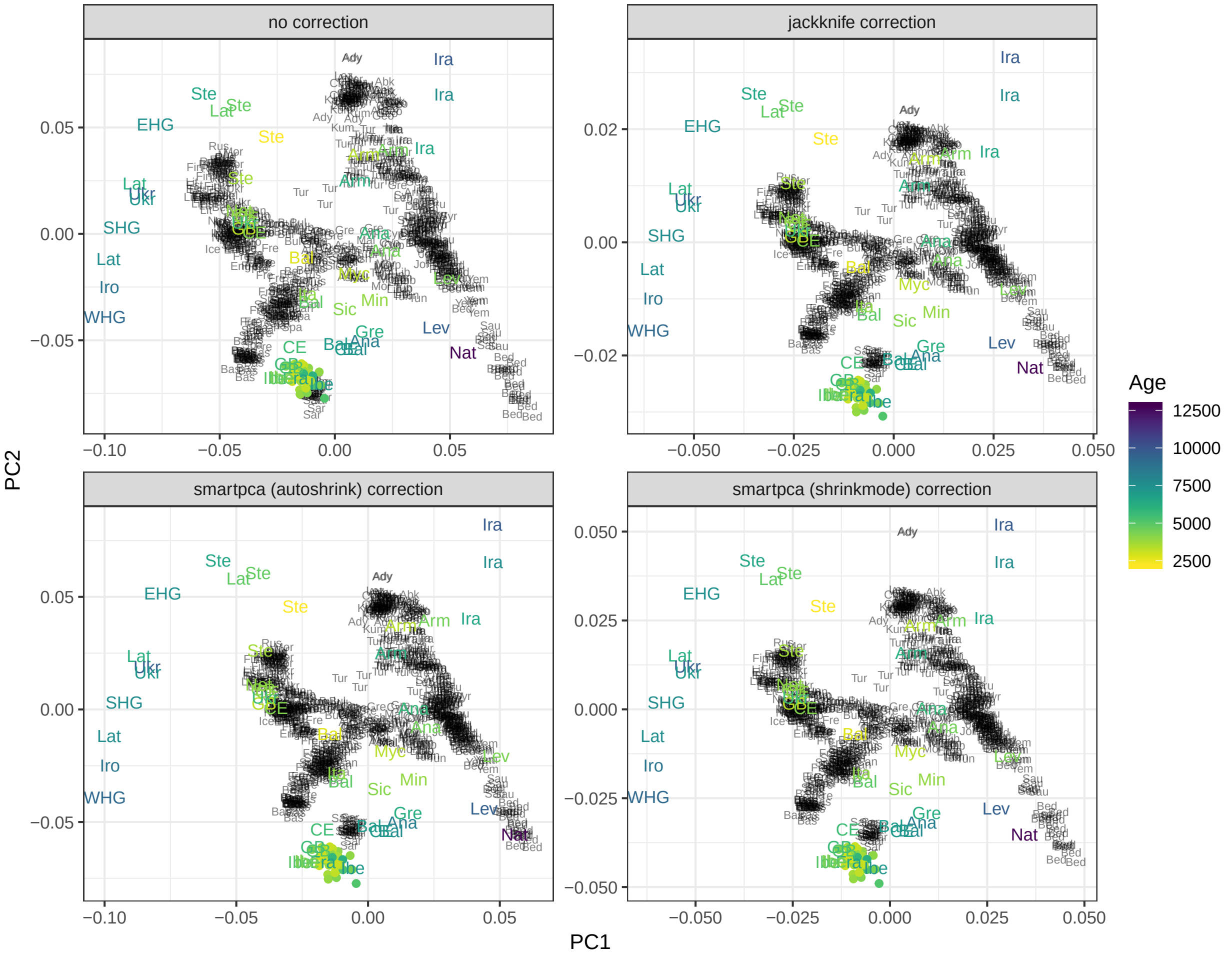


Figure 2.17: **Visualization of the effect of different shrinkage correction approaches on the top two PCs for projected ancient individuals.** All panels show the results an initial PCA on modern Western Eurasian individuals, each of whom are represented as a black three-letter short hand for their assigned population label. We project each ancient individual onto these modern PCs and then represent the median projected PC value of each ancient group as a three-letter short hand colored by the group's median age. Each panel shows a different different correction approach (in the top left showing no correction). We do not observe substantial differences between the three correction approaches especially in the region around the ancient Sardinian individuals from this study.

2.10.5 Supplementary Note 8: Assessing Geographic Substructure in Pre-Nuragic Ancient Sardinia

Joseph Marcus and John Novembre

In our initial analysis we saw no strong signal of sub-population structure in ancient Sardinian individuals until the post-Nurgaic period when we observed heterogeneous shifts in ancestry depending on the sampled archaeological site (main Fig. 2). Here we investigate in more detail if we can detect a signal of geographic substructure before the Nurgaic period.

One approach to investigate signals of fine-scale geographic structure would be to take the genotypes of pre-Nuragic ancient Sardinian individuals and directly visualize their covariance structure, in a low dimensional space, such as typically done with Principal Components Analysis (PCA). Unfortunately, our ancient capture data has high levels of missingness. For instance, the median proportion of missing sites across pairs individuals is 0.862 with the 5th and 95th percentiles being 0.422 and 0.989. These high levels of missingness are problematic for two reasons: 1) There are not many widely used methods that account for missing data, while estimating population structure. 2) Interpreting the resulting structure could be difficult as pairs of individuals have unequal levels of overlapping data, creating a heteroskedastic noise model. As a compromise, we use the projections of ancient Sardinian individuals onto PCs trained on modern Sardinian individuals and tried to see if these PCs were associated with any covariates related to geography or sampling location. As mentioned previously, using the projections onto modern Principal Components is a powerful approach because there is very little missingness in the moderns genotypes and they are not affected by sequencing error modes unique to ancient DNA. This means projecting the ancient genotypes onto modern PCs helps to, in some sense, regularize the estimates of population structure for the ancients. As a drawback, we note that this approach could potentially miss population structure present in the ancients but absent in the moderns.

In linear models individually regressing latitude and longitude against the top 9 PC projections we observe that only the projection onto PC6 of the within-Sardinia variation is significantly associated with longitude (Figure 2.22, Figure 2.23). We also subdivided each ancient individual into broad geographic regions based on the locations of archaeological sites with more than 3 sampled individuals (CentralEast1, CentralEast2, NorthWest1, NorthWest2, SouthWest, and the remainder were put into a group labelled '< 3'). We computed 1 way anovas of each PC projection vs these regional labels and found that only the projection onto PC6 of the within-Sardinia variation was significantly associated with these course geographic labels (Figure 2.24). Finally, we regressed the PC projections against the radio-carbon date age estimate of each individual and found again that only PC6 was significantly associated with age. Because age is confounded with longitude and the course geographic region it is difficult to determine which covariate is driving the association with PC6, although we note the association with age is much more significant than the other covariates. Furthermore, correcting for multiple testing, the associations with PC6 would likely not survive. Figures 2.26 and 2.27 contain plots of each of the PCs for reference.

In summary, at least from the projections on to PCs derived from modern data, we could not detect any strong and significant signals of geographic structure. Higher coverage sequencing data and perhaps haplotype based approaches to reveal signals of geographic structure for individuals with such low levels of differentiation.

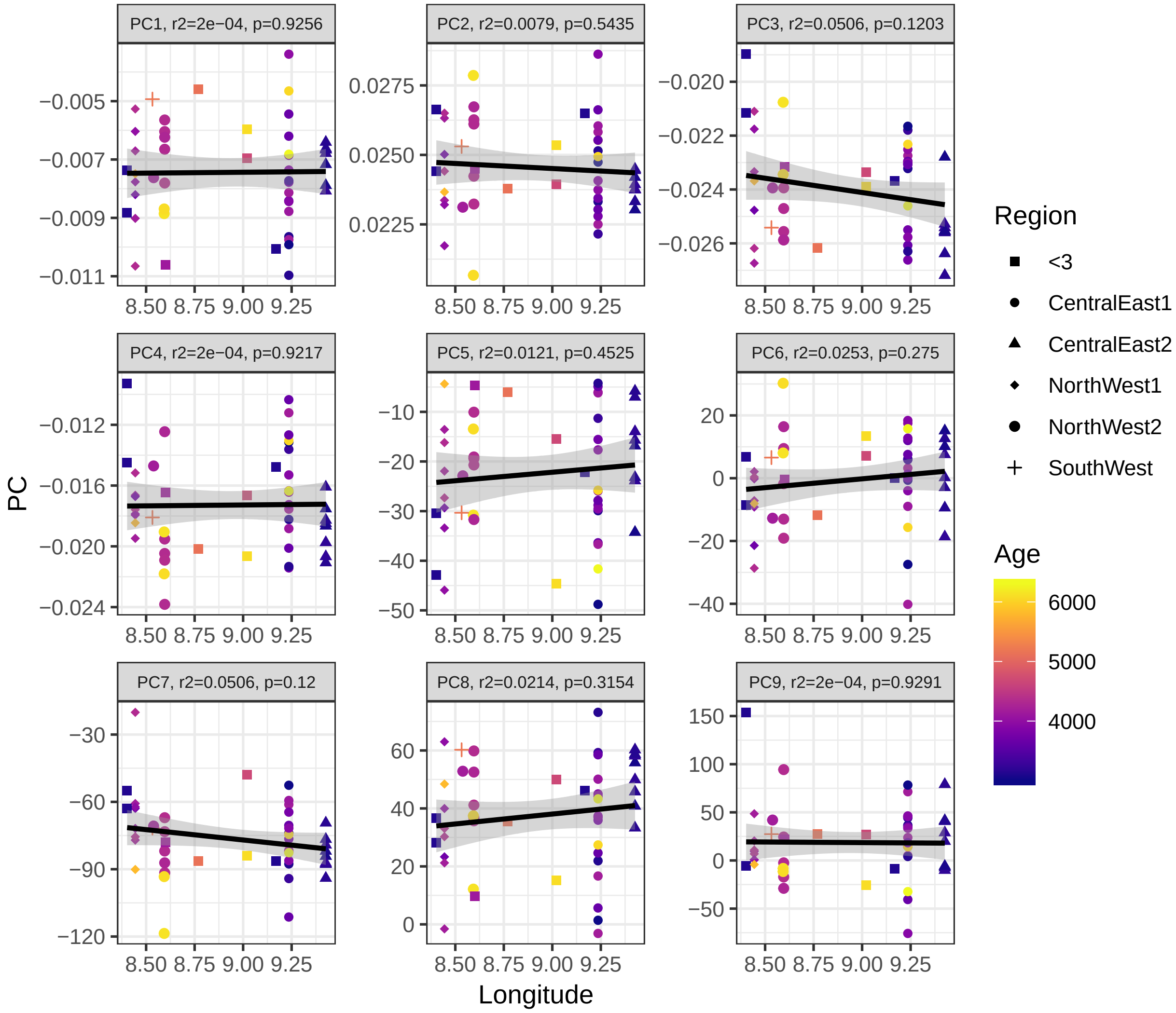


Figure 2.18: **Main Text Figure 2B Principal Component Projections vs. Longitude**: Here each sub-panel displays a visualization of a given PC vs Longitude, with its corresponding fitted linear regression. We show a grid of the top 9 PCs. The text in each sub-panel displays the sample correlation and significance of association between each PC and longitude. PC6 is the only PC that has a significant association.

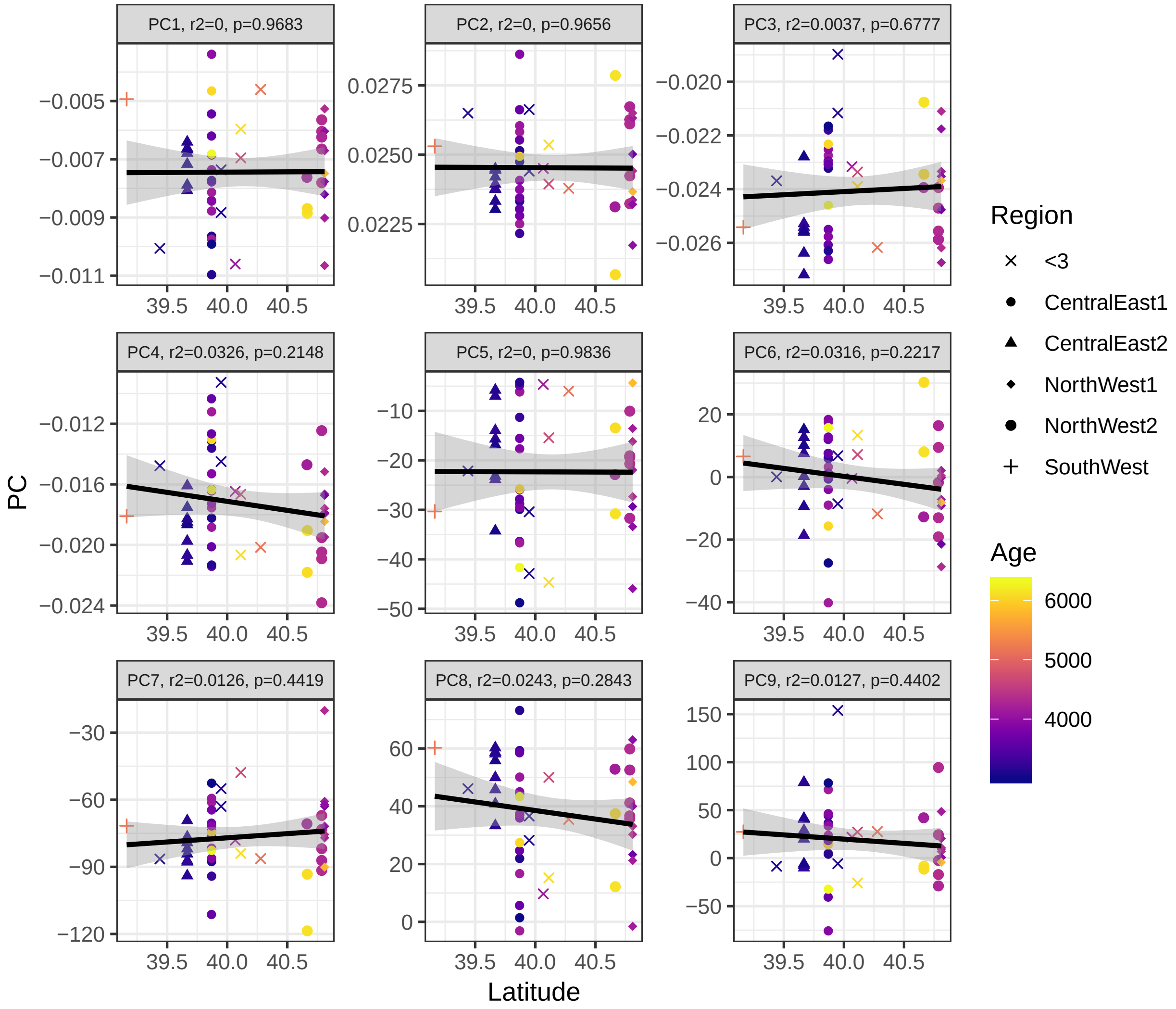


Figure 2.19: **Main Text Figure 2B Principal Component Projections vs. Latitude**: Here each sub-panel displays a visualization of a given PC vs Latitude, with its corresponding fitted linear regression. We show a grid of the top 9 PCs. The text in each sub-panel displays the sample correlation and significance of association between each PC and Latitude.

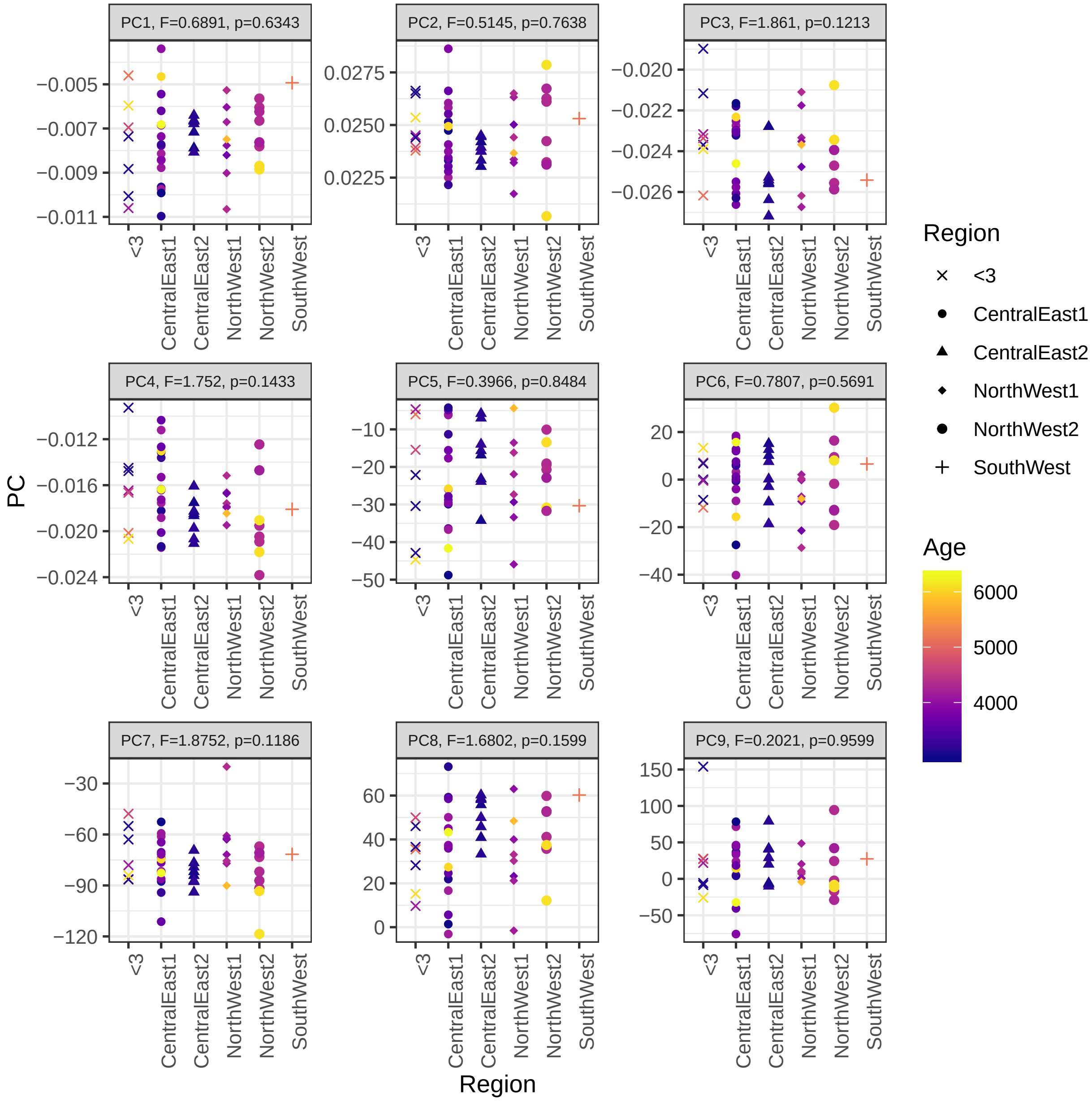


Figure 2.20: **Main Text Figure 2B Principal Component Projections vs. Region**: Here each sub-panel displays a visualization of a given PC vs a course geographic region label. We assigned individuals to the "<3" label if less than three individuals were sampled at the same geographic position. We show a grid of the top 9 PCs. The text in each sub-panel displays the F statistic and p-value output by computing a one-way anova.

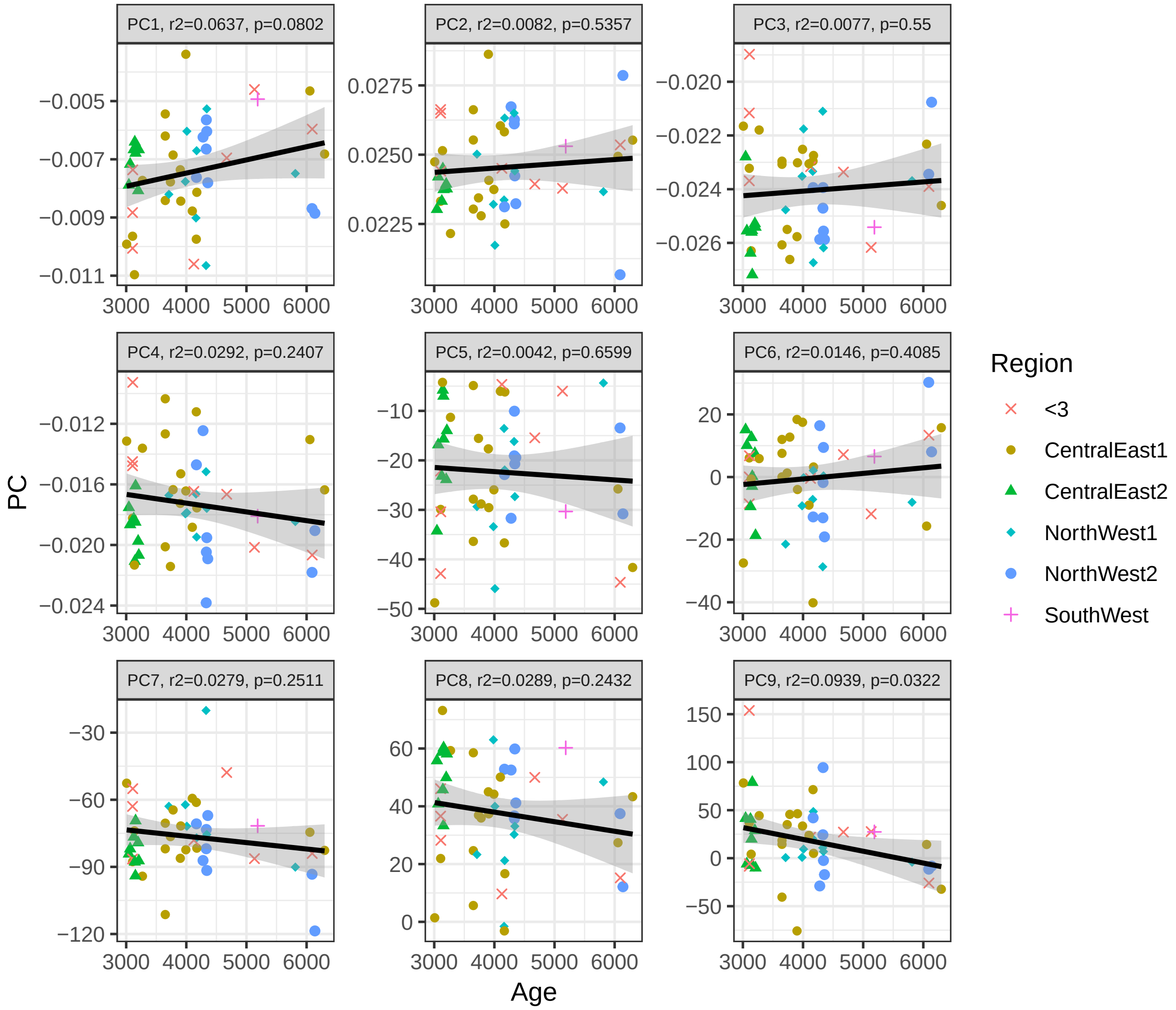


Figure 2.21: **Main Text Figure 2B Principal Component Projections vs. Age**: Here each sub-panel displays a visualization of a given PC vs Age, with its corresponding fitted linear regression. We show a grid of the top 9 PCs. The text in each sub-panel displays the sample correlation and significance of association between each PC and age.

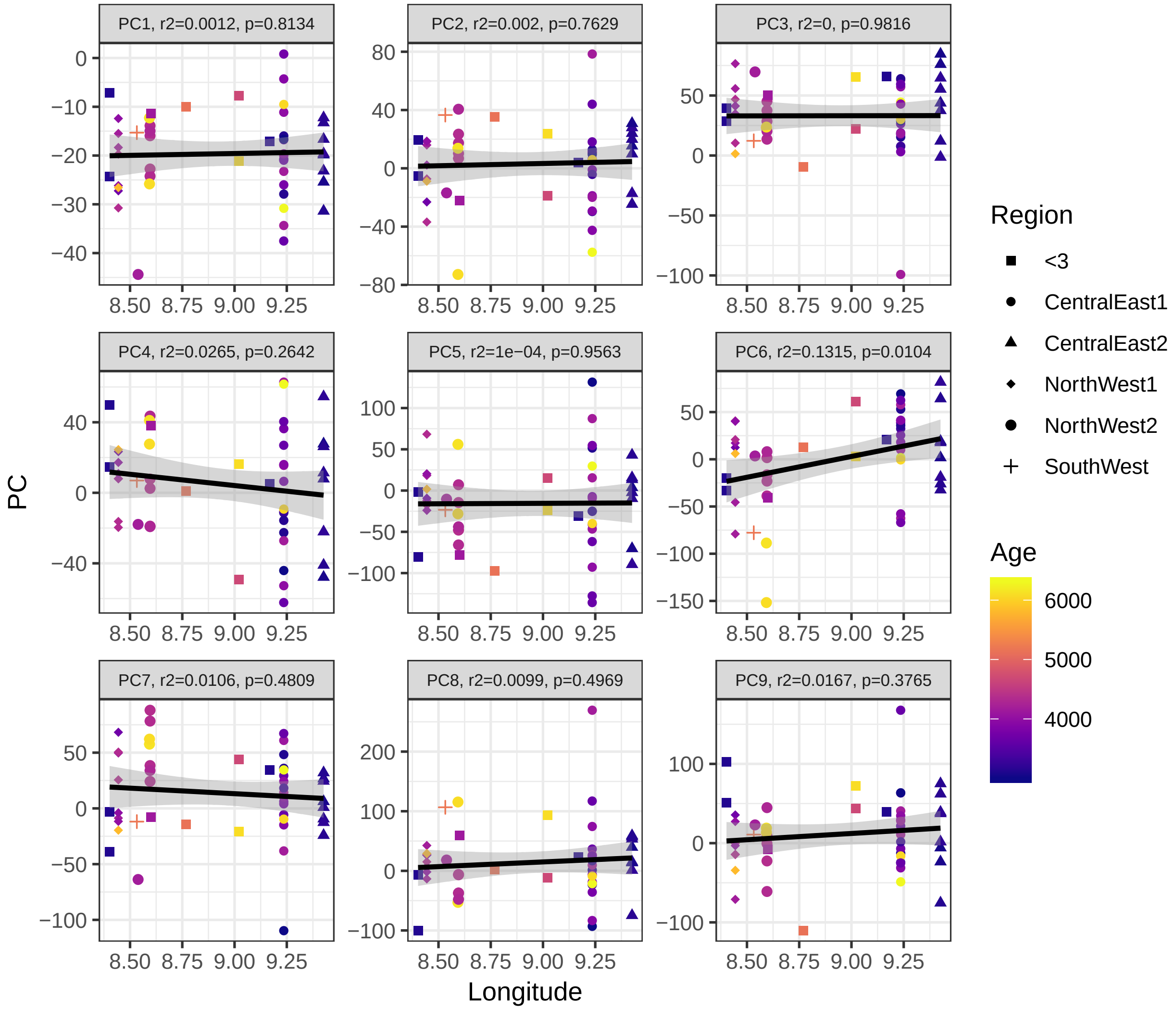


Figure 2.22: **Sardinia Principal Component Projections vs. Longitude**: Here each sub-panel displays a visualization of a given PC vs Longitude, with its corresponding fitted linear regression. We show a grid of the top 9 PCs. The text in each sub-panel displays the sample correlation and significance of association between each PC and longitude. PC6 is the only PC that has a significant association.

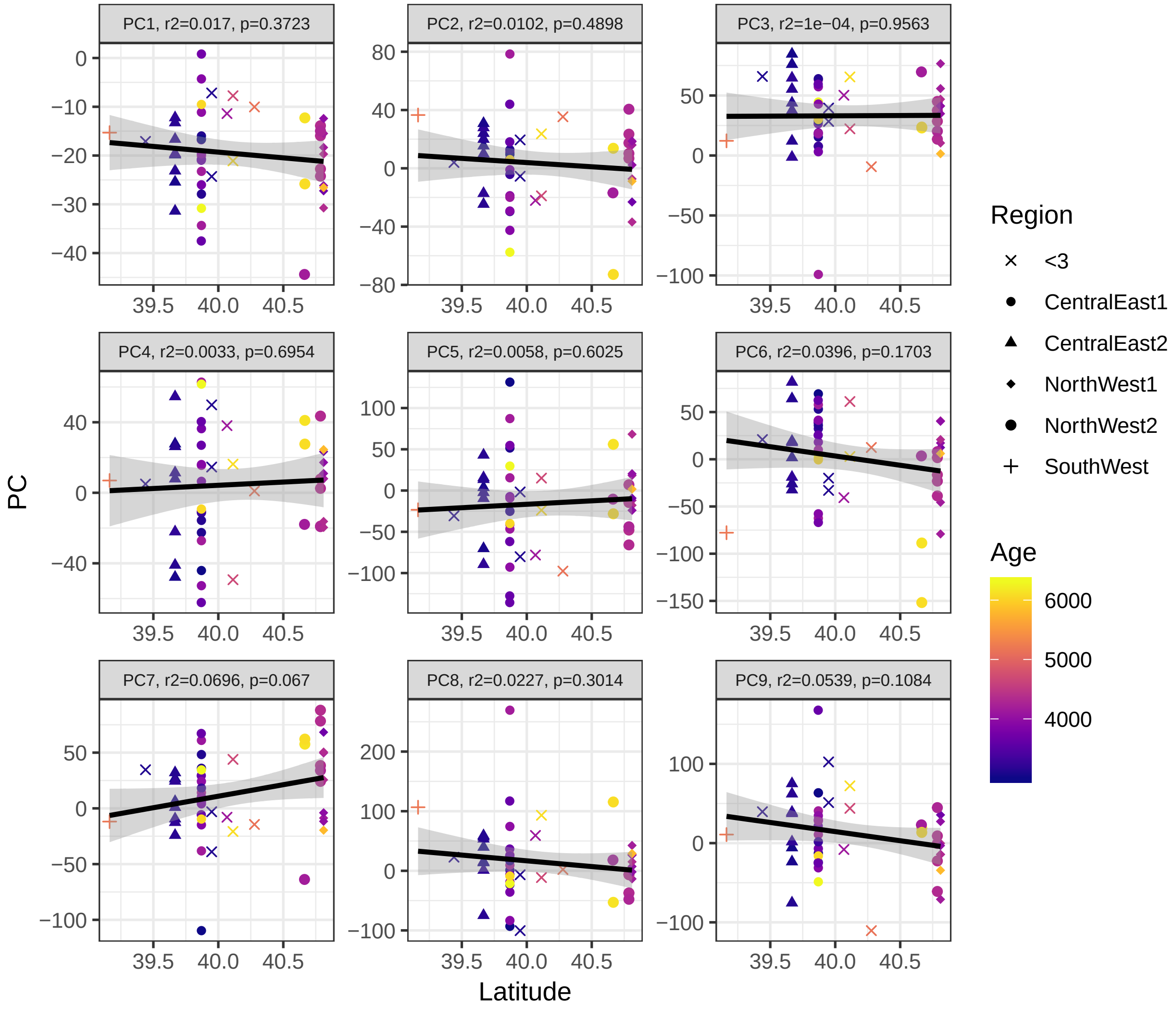


Figure 2.23: **Sardinia Principal Component Projections vs. Latitude**: Here each sub-panel displays a visualization of a given PC vs Latitude, with its corresponding fitted linear regression. We show a grid of the top 9 PCs. The text in each sub-panel displays the sample correlation and significance of association between each PC and Latitude. There are no PCs significantly associated with latitude.

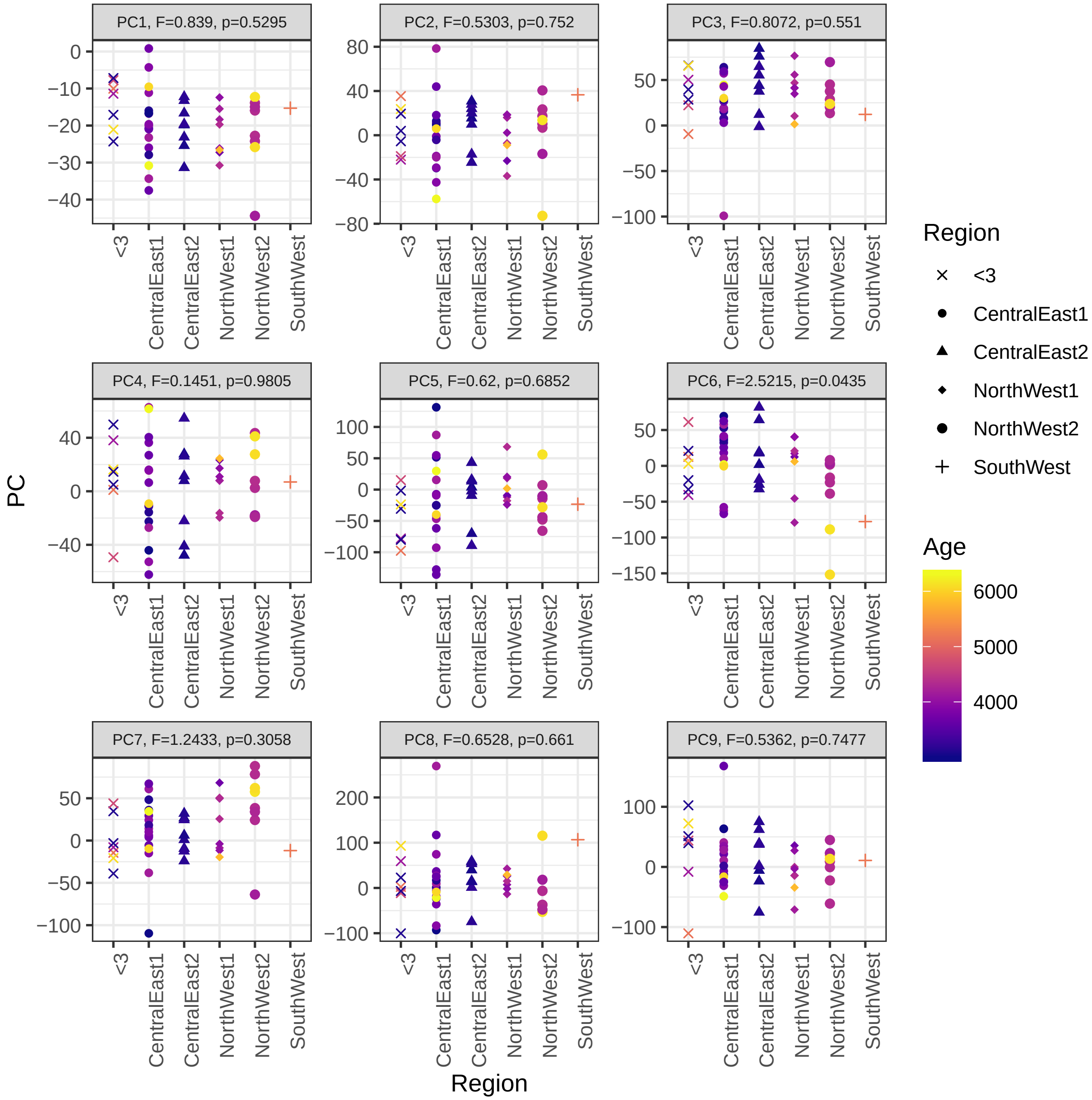


Figure 2.24: **Sardinia Principal Component Projections vs. Region**: Here each sub-panel displays a visualization of a given PC vs a course geographic region label. We assigned individuals to the ”<3” label if less than three individuals were sampled at the same geographic position. We show a grid of the top 9 PCs. The text in each sub-panel displays the F statistic and p-value output by computing a one-way anova. PC6 is the only PC significantly associated with these course geographic labels.

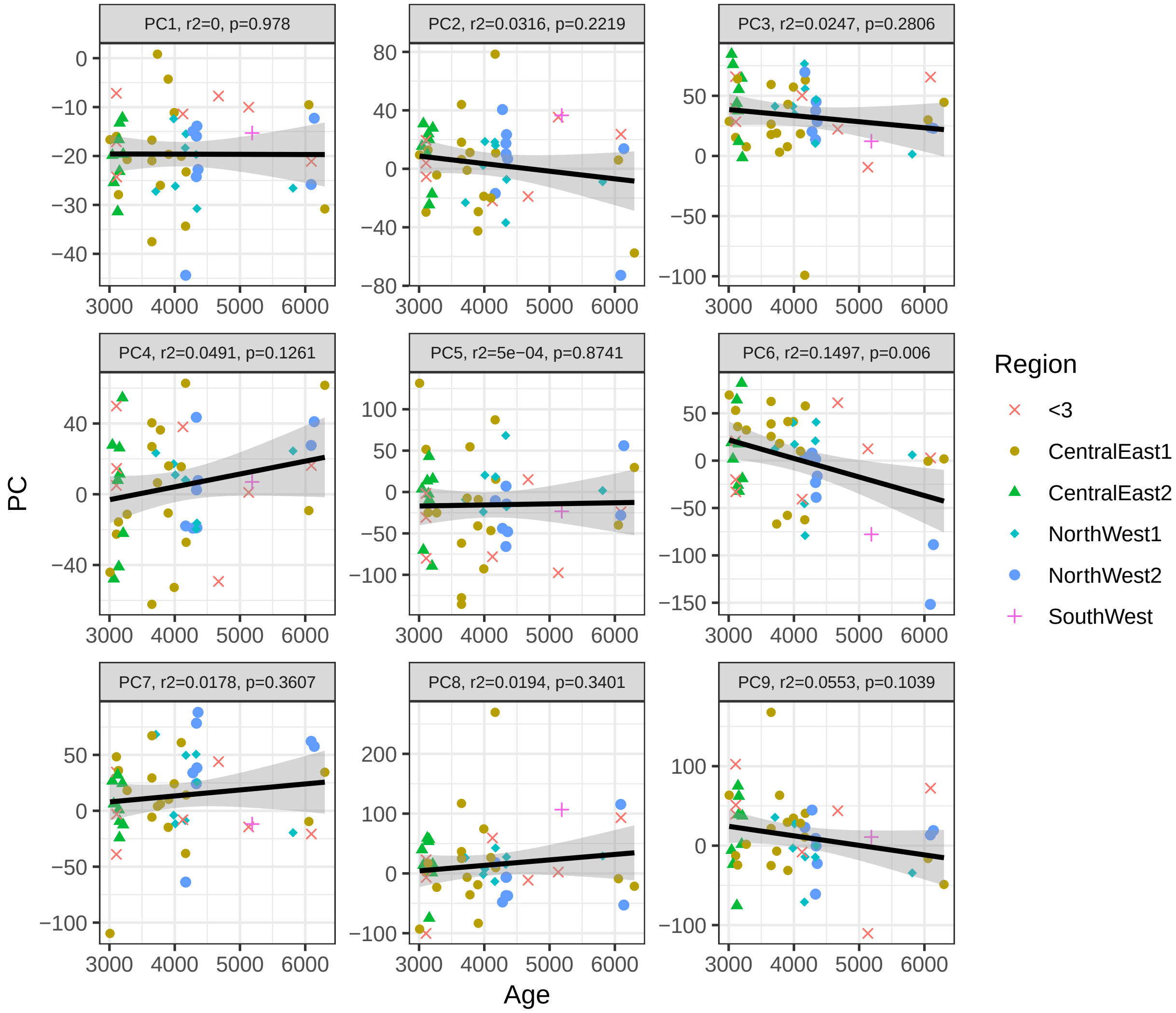


Figure 2.25: **Sardinia Principal Component Projections vs. Age**: Here each sub-panel displays a visualization of a given PC vs Age, with its corresponding fitted linear regression. We show a grid of the top 9 PCs. The text in each sub-panel displays the sample correlation and significance of association between each PC and age. PC6 is the only PC significantly associated with age.

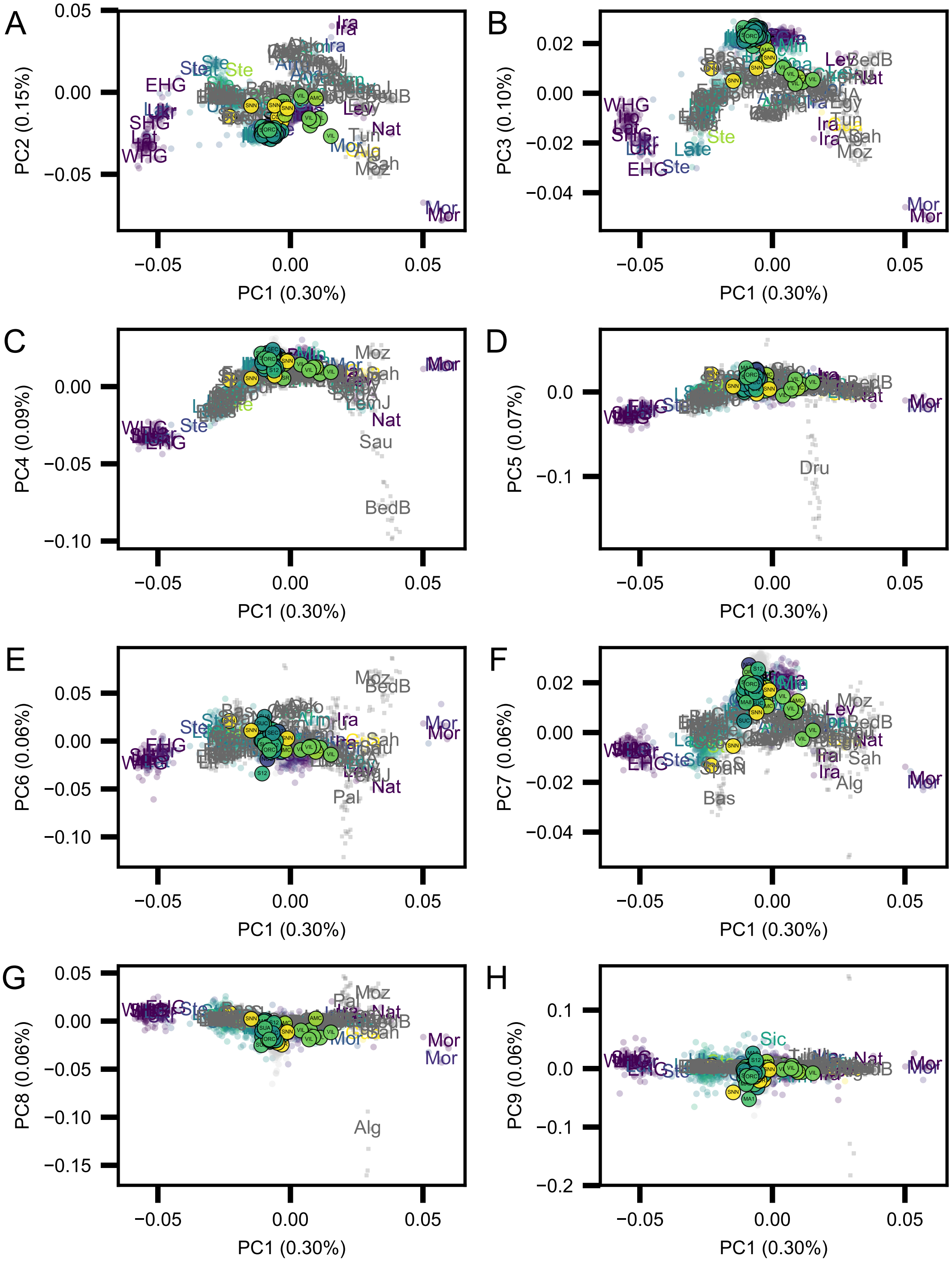


Figure 2.26: **Principal components 1-10 for the PCA shown in Main Text Fig. 2.** Each subplot has PC1 on its x-Axis. Label and color choices are described in detail in the caption of Main Fig. 2.

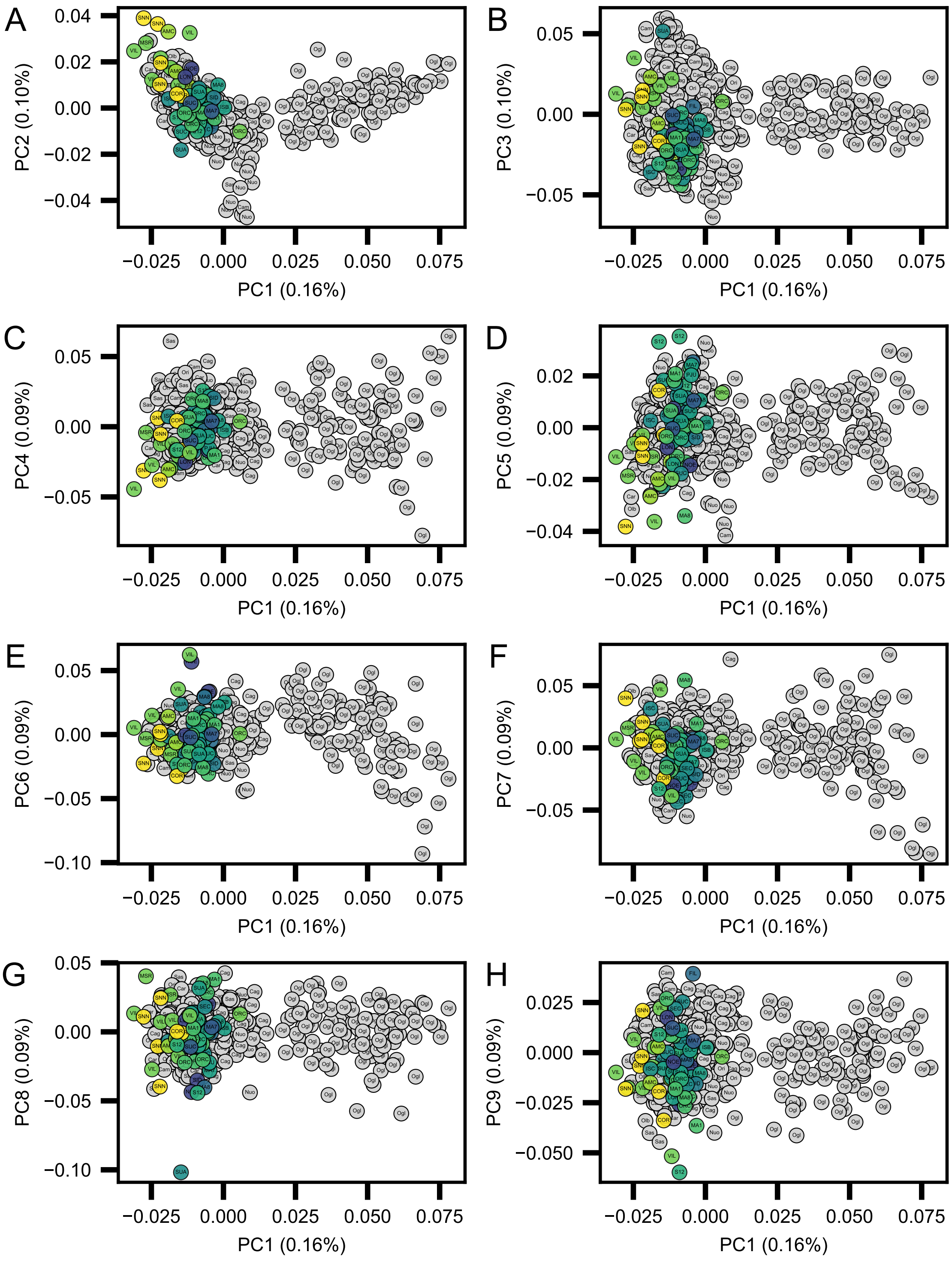


Figure 2.27: **Principal components 1-10 for the PCA shown in Main Fig. 4B.** Each subplot has PC1 on its x-Axis. Label and color choices are described in detail in the caption of Main Fig. 4B.

CHAPTER 3

FAST AND FLEXIBLE ESTIMATION OF EFFECTIVE MIGRATION SURFACES

*Joseph H. Marcus**, *Wooseok Ha**, *Rina Foygel Barber*†, *and John Novembre*†

* *denotes co-first authorship and* † *denotes co-mentorship*

This chapter has been published and can be found here Marcus et al. (2020b).

3.1 Abstract

An important feature in spatial population genetic data is often "isolation-by-distance," where genetic differentiation tends to increase as individuals become more geographically distant. Recently, Petkova et al. (2016) developed a statistical method called Estimating Effective Migration Surfaces (EEMS) for visualizing spatially heterogeneous isolation-by-distance on a geographic map. While EEMS is a powerful tool for depicting spatial population structure, it can suffer from slow runtimes. Here we develop a related method called Fast Estimation of Effective Migration Surfaces (FEEMS). FEEMS uses a Gaussian Markov Random Field in a penalized likelihood framework that allows for efficient optimization and output of effective migration surfaces. Further, the efficient optimization facilitates the inference of migration parameters per edge in the graph, rather than per node (as in EEMS). When tested with coalescent simulations, FEEMS accurately recovers effective migration surfaces with complex gene-flow histories, including those with anisotropy. Applications of FEEMS to population genetic data from North American gray wolves shows it to perform comparably to EEMS, but with solutions obtained orders of magnitude faster. Overall, FEEMS expands the ability of users to quickly visualize and interpret spatial structure in their data.

3.2 Introduction

The relationship between geography and genetics has had enduring importance in evolutionary biology (see Felsenstein, 1982). One fundamental consideration is that individuals who live near one another tend to be more genetically similar than those who live far apart (Wright, 1943, 1946; Malécot, 1948; Kimura, 1953; Kimura and Weiss, 1964). This phenomenon is often referred to as "isolation-by-distance" (IBD) and has been shown to be a pervasive feature in spatial population genetic data across many species (Slatkin, 1985; Dobzhansky and Wright, 1943; Meirmans, 2012). Statistical methods that use both measures of genetic variation and geographic coordinates to understand patterns of IBD have been widely applied (Bradburd and Ralph, 2019; Battey et al., 2020). One major challenge in these approaches is that the relationship between geography and genetics can be complex. Particularly, geographic features can influence migration in localized regions leading to spatially heterogeneous patterns of genetic covariation (Bradburd and Ralph, 2019).

Multiple approaches have been introduced to model non-homogeneous IBD in spatial population genetic data (McRae, 2006; Duforet-Frebourg and Blum, 2014; Hanks and Hooten, 2013; Petkova et al., 2016; Bradburd et al., 2018; Al-Asadi et al., 2019; Safner et al., 2011; Ringbauer et al., 2018). Particularly relevant to our proposed approach is the work of Petkova et al. (2016) and Hanks and Hooten (2013). Both approaches model genetic distance using the "resistance distance" on a weighted graph. This distance metric is inspired by concepts of effective resistance in circuit theory models, or alternatively understood as the commute time of a random walk on a weighted graph or as a Gaussian graphical model (specifically a conditional auto-regressive process) (Chandra et al., 1996; Hanks and Hooten, 2013; Rue and Held, 2005). Additionally, the resistance distance approach is a computationally convenient and accurate approximation to spatial coalescent models (McRae, 2006), though it has limitations in asymmetric migration settings (Lundgren and Ralph, 2019).

Hanks and Hooten (2013) introduced a Bayesian model that uses measured ecological covariates, such as elevation, to help predict genetic distances across sub-populations. Specifically, they use a graph-based model for genotypes observed at different spatial locations. Expected genetic distances across sub-populations in their model are given by resistance distances computed from the edge weights. They parameterize the edge weights of the graph to be a function of known biogeographic covariates, linking local geographic features to genetic variation across the landscape.

Concurrently, the Estimating Effective Migration Surfaces (EEMS) method was developed to help interpret and visualize non-homogeneous gene-flow on a geographic map (Petkova, 2013; Petkova et al., 2016). EEMS uses resistance distances to approximate the between-sub-population component of pairwise coalescent times in a "stepping-stone" model of migration and genetic drift (Kimura, 1953; Kimura and Weiss, 1964). EEMS models the within-sub-population component of pairwise coalescent times, with a node-specific parameter. Instead of using known biogeographic covariates to connect geographic features to genetic variation as in Hanks and Hooten (2013), EEMS infers a set of edge weights (and diversity parameters) that explain the genetic distance data. The inference is based on a hierarchical Bayesian model and a Voronoi-tessellation-based prior to encourage piece-wise constant spatial smoothness in the fitted edge weights.

EEMS uses Markov Chain Monte Carlo (MCMC) and outputs a visualization of the posterior mean for effective migration and a measure of genetic diversity for every spatial position of the focal habitat. Regions with relatively low effective migration can be interpreted to have reduced gene-flow over time whereas regions with relatively high migration can be interpreted as having elevated gene-flow. EEMS has been applied to multiple systems to describe spatial genetic structure, but despite EEMS's advances in computational tractability with respect to the previous work, the MCMC algorithm it uses can be slow to converge, in some cases leading to days of computation time for large datasets (Peter et al., 2018).

These inference problems from spatial population genetics are related to a growing area of interest in the graph signal processing literature referred to as "graph learning" (Dong et al., 2019; Mateos et al., 2019). In graph learning, a noisy signal is measured as a scalar value at a set of nodes from the graph, and the aim is then to infer non-negative edge weights that reflect how spatially "smooth" the signal is with respect to the graph topology (Kalofolias, 2016). In population genetic settings, this scalar could be an allele frequency measured at locations in a discrete spatial habitat with effective migration rates between sub-populations. Like the approach taken by Hanks and Hooten (2013), one widely used representation of smooth graph signals is to associate the smoothness property with a Gaussian graphical model where the precision matrix has the form of a graph Laplacian (Dong et al., 2016; Egilmez et al., 2016). The probabilistic model defined on the graph signal then naturally gives rise to a likelihood for the observed samples, and thus much of the literature in this area focuses on developing specialized algorithms to efficiently solve optimization problems that allow reconstruction of the underlying latent graph. For more information about graph learning and signal processing in general see the excellent survey papers of Dong et al. (2019) and Mateos et al. (2019).

To position the present work in comparison to the "graph learning" literature, our contributions are twofold:

- In population genetics, it is impossible to collect individual genotypes across all the geographic locations and, as a result, we often work with many, often the majority, of nodes having missing data. As far as we are aware, none of the work in graph signal processing considers this scenario and thus their algorithms are not directly applicable to our setting. In addition, if the number of the observed nodes is much smaller than the number of nodes of a graph, one can project the large matrices associated with the graph to the space of observed nodes, therefore allowing for fast and efficient computation.

- On the other hand, highly missing nodes in the observed signals can result in significant degradation of the quality of the reconstructed graph unless it is regularized properly. Motivated by the Voronoi-tessellation-based prior adopted in EEMS (Petkova et al., 2016), we propose regularization that encourages spatial smoothness in the edge weights.

Building on advances in graph learning, we introduce a method, Fast Estimation of Effective Migration Surfaces (FEEMS), that uses optimization rather than MCMC to obtain penalized-likelihood-based estimates of effective migration parameters. In contrast to EEMS which uses a node-specific parameterization of effective migration, we optimize over edge-specific parameters allowing for more flexible migration processes to be fit, such as spatial anisotropy, in which the migration process is not invariant to rotation of the coordinate system (e.g., migration is more extensive along a particular axis). We develop a fast quasi-Newton optimization algorithm (Nocedal and Wright, 2006) and apply it to a dataset of gray wolves from North America. The output is comparable to the results of EEMS but is provided in orders of magnitude less time. With this improvement in speed, FEEMS opens up the ability to perform fast exploratory and iterative data analysis of spatial population structure.

3.3 Results

3.3.1 Overview of FEEMS

Figure 4.1 shows a visual schematic of the FEEMS method. The input data are genotypes and spatial locations (e.g., latitudes and longitudes) for a set of individuals sampled across a geographic region. We construct a dense spatial grid embedded in geographic space where nodes represent sub-populations, and we assign individuals to nodes based on spatial proximity (see Supp. Fig. 3.5 for a visualization of the grid construction and node assignment procedure). The density of the grid is user defined and must be explored to appropriately balance model-mis-specification and computational burden. As the density of the lattice increases, the model is similar to discrete approximations used for continuous spatial processes, but the increased density comes at the cost of computational complexity.

We assume exchangeability of individuals within each sub-population and estimate allele frequencies, $\widehat{f}_j(k)$, for each sub-population, indexed by k, and single nucleotide polymorphism (SNP), indexed by j, under a simple Binomial sampling model. We also use the recorded sample sizes at each node to model the precision of the estimated allele frequency. The use of allele frequencies allows a number of advantages in this context: (1) Allele frequencies can be more easily shared between researchers than individual genotypes due to privacy concerns, which is especially relevant in human population genetic studies; (2) We usually gain large computational savings in memory and speed because in most population genetic studies the number of observed locations, in which allele frequencies are estimated, is smaller than the total number of individuals sampled i.e. many individuals are sampled from the same spatial location.

With the estimated allele frequencies in hand, we model the data at each SNP using an approximate Gaussian model whose covariance is shared across all SNPs, in other words we assume that the observed frequencies at each SNP is an independent realization of the same spatial process after rescaling by SNP-specific variation factors. The latent frequency variables, $f_j(k)$, are modeled as a Gaussian Markov Random Field (GMRF) with a sparse precision matrix determined by the graph Laplacian and a set of residual variances. The graph's weighted edges, denoted by w_{ij} between nodes i and j, represent gene-flow between the sub-populations (Friedman et al., 2008; Hanks and Hooten, 2013; Petkova et al., 2016). We analytically marginalize out the latent frequency variables and use penalized restricted maximum likelihood to estimate the edge weights of the graph after removing the SNP-specific mean allele frequencies by projecting the data onto contrasts (Felsenstein, 1982;

Hanks and Hooten, 2013; Petkova et al., 2016). Our overall goal is to solve the following optimization problem:

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{l} \leqslant \boldsymbol{w} \leqslant \boldsymbol{u}}{\arg\min} \ \ell(\boldsymbol{w}) + \phi_{\lambda,\alpha}(\boldsymbol{w}),$$

where $\boldsymbol{w}$ is a vector that stores all the unique elements of the weighted adjacency matrix, $\boldsymbol{l}$ and $\boldsymbol{u}$ are element-wise non-negative lower and upper bounds for $\boldsymbol{w}$, and $\ell(\boldsymbol{w})$ is the negative log-likelihood function that comes from the GMRF model described above. The penalty, $\phi_{\lambda,\alpha}(\boldsymbol{w})$, controls how constant or smooth the output migration surface will be and is controlled by the hyperparameters λ and α. Specifically, the hyperparameters determine a penalty function based on the squared differences between edge weights for pairs of edges that share a common node,

$$\phi_{\lambda,\alpha}(\boldsymbol{w}) = \frac{\lambda}{2} \|\boldsymbol{\Delta}(\boldsymbol{w} + \alpha \log(\boldsymbol{w})\|_2^2,$$

where $\boldsymbol{\Delta}$ is a signed graph incidence matrix indicating if two edges are connected to the same node. Note that λ controls the overall strength of the penalization placed on the output of migration surface while α controls the relative strength of the penalization on the logarithmic scale. Thus, if the model is highly penalized, the fitted surface will favor a homogeneous spatial process on the graph across orders of magnitude of edge weights and if the penalty is low, more flexible graphs can be fit, but are potentially prone to overfitting. Akin to the choice in admixture models of the number of latent ancestral populations or clusters (K), inspecting the outputs across a series of λ and α values is recommended and demonstrated (below). We use sparse linear algebra routines to efficiently compute the objective function and gradient of our parameters, allowing the use of widely applied quasi-Newton optimization algorithms (Nocedal and Wright, 2006) implemented in standard

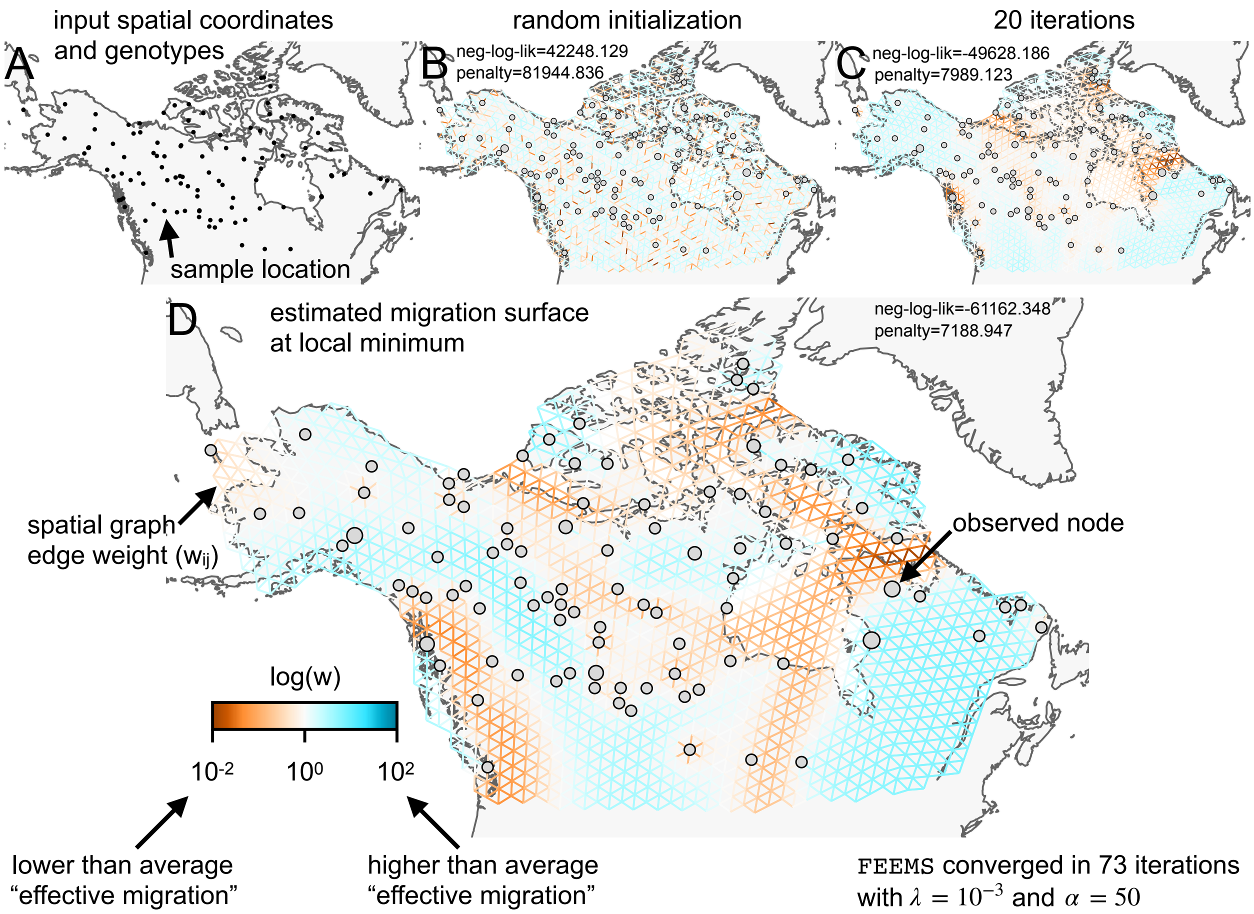


Figure 3.1: **Schematic of the FEEMS model:** The full panel shows a schematic of going from the raw data (spatial coordinates and gentoypes) through optimization of the edge weights, representing effective migration, to convergence of FEEMS to a local optima. (A) Map of sample coordinates (black points) from a dataset of gray wolves from North America (Schweizer et al., 2016). The input to FEEMS are latitude and longitude coordinates as well as genotype data for each sample. (B) The spatial graph edge weights after random initialization uniformly over the graph to begin the optimization algorithm. (C) The edge weights after 20 iterations of running FEEMS, when the algorithm has not converged yet. (D) The final output of FEEMS after the algorithm has fully converged. The output is annotated with important features of the visualization.

numerical computing libraries like `scipy` (Virtanen et al., 2020). See the materials and methods section for a detailed description of the statistical models and algorithms used.

3.3.2 Evaluating FEEMS on "out of model" coalescent simulations

While our statistical model is not directly based on a population genetic process, it is useful to see how it performs on simulated data under the coalescent stepping stone model. In these simulations we know, by construction, the model we fit (FEEMS) is different from the true model we simulate data under (the coalescent), allowing us to assess the robustness of the fit to a controlled form of model mis-specification. In Figure 3.2 we use `msprime` (Kelleher et al., 2016) to recapitulate and extend the results of Petkova et al. (2016), simulating data under the coalescent in three simple migration scenarios with two different spatial sampling designs. Note that in Supp. Fig. 3.6 we display a larger set of simulations with additional sampling configurations. For brevity, here we only show results for $\lambda = .001$ and $\alpha = 50$, based on values that performed well after experimental tuning. In Supp. Fig. 3.7 and Supp. Fig. 3.8, we also show results varying λ and α for two migration scenarios with one particular sampling design.

The first migration scenario (Figure 3.2A-C) is a spatially homogeneous model where all the migration rates are set to be a constant value on the graph, this is equivalent to simulating data under an homogeneous isolation-by-distance model. In the second migration scenario (Figure 3.2D-E) we simulate a non-homogeneous process by representing a geographic barrier to migration, lowering the migration rates by a factor of 10 in the center of the habitat relative to the left and right regions of the graph. Finally, in the third migration scenario (Figure 3.2G-I) we simulate a pattern which corresponds to anisotropic migration with edges that point east/west being assigned to a five-fold higher migration rate than edges pointing north/south. For each migration scenario we simulate two sampling designs. In the first "dense-sampling" sampling design (Figure 3.2B,E,I) we sample individuals for every node of the graph. Next, in the "sparse-samplng" sampling design (Figure 3.2C,F,J) we randomly sample individuals for only 20% of the nodes.

As expected, FEEMS performs best when all the nodes are sampled on the graph,

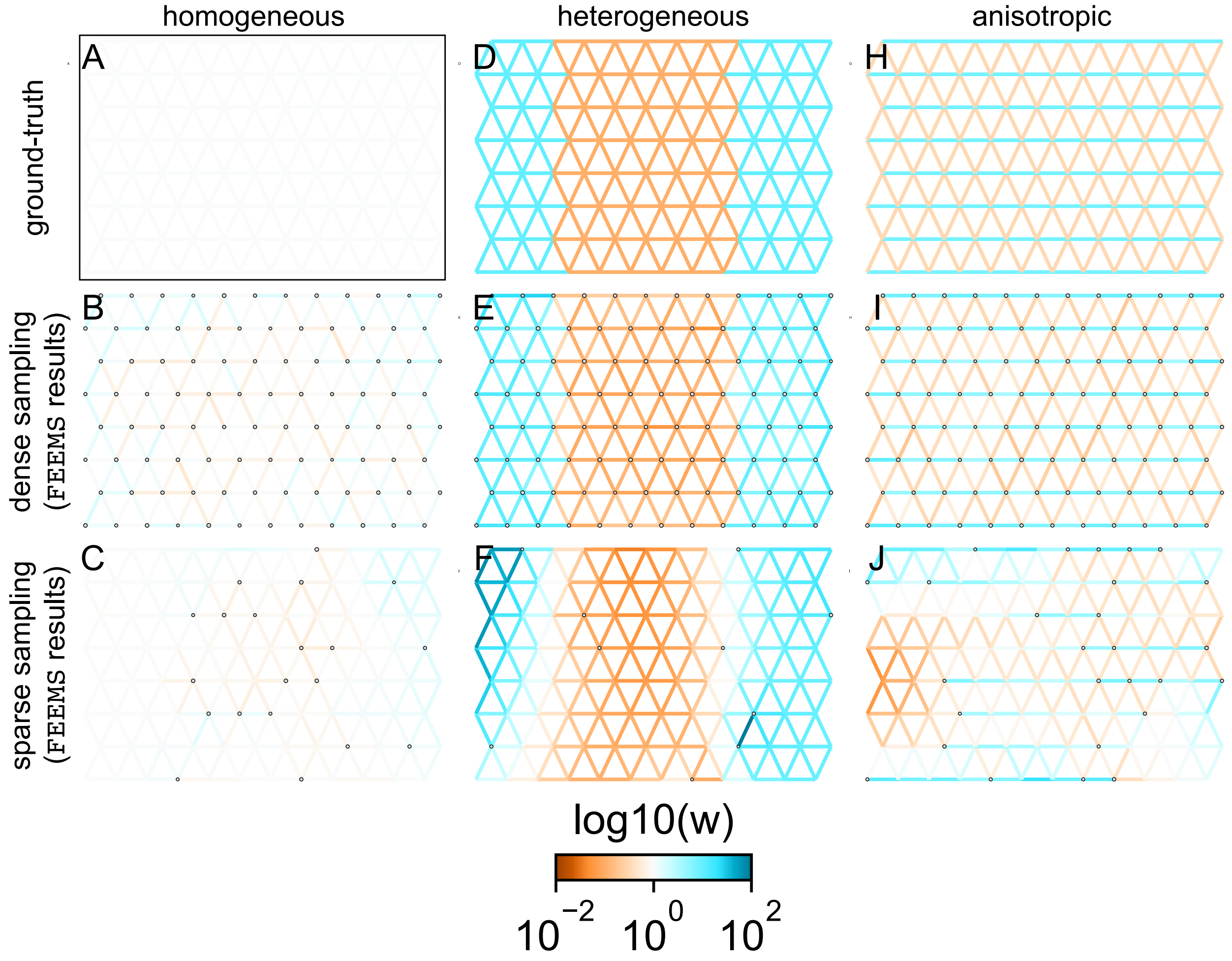


Figure 3.2: **FEEMS fit to coalescent simulations:** We run FEEMS on coalescent simulations, varying the migration history (columns) and sampling design (rows). The first column (A-C) shows the ground-truth and fit of FEEMS to coalescent simulations with a homogeneous migration history i.e. a single migration parameter for all edge weights. Note that the ground-truth simulation figures (A,D,F) display coalescent migration rates, not fitted effective migration rates output by FEEMS. The second column (D-F) shows the ground truth and fit of FEEMS to simulations with a heterogeneous migration history i.e. reduced gene-flow, with 10 fold lower migration, in the center of the habitat. The third column (H-J) shows the ground truth and fit of FEEMS to an anisotropic migration history with edge weights facing east-west having five fold higher migration than north-south. The second row (B,E,H) shows a sampling design with no missing observations on the graph. The final row (C,F,I) shows a sampling design with 80% of nodes missing at random.

i.e. when there is no missing data (Figure 3.2B,E,H). Interestingly, in the simulated scenarios with many missing nodes, FEEMS can still partly recover the migration history, including the presence of anisotropic migration (Figure 3.2F). A sampling scheme with a central gap leads to a slightly narrower barrier in the heterogeneous migration scenario (Supp. Fig. 3.6I) and for the anisotropic scenario, a degree of over-smoothness in the northern and southern regions of the center of the graph (Supp. Fig. 3.6N). For the missing at random sampling design, FEEMS is able to recover the relative edge weights surprisingly well for all scenarios, with the inference being the most challenging when there is anisotropic migration. We emphasize that the potential for FEEMS to recover anisotropic migration is novel relative to EEMS, which was parameterized for fitting non-stationary isotropic migration histories and produces banding patterns perpendicular to the axis of migration when applied to data from anisotropic coalescent simulations (see Petkova et al. (2016) supplementary figure 2; see also Supp. Note "*Edge versus node parameterization*" for a related discussion). Overall, even with sparsely sampled graphs, FEEMS is able to produce visualizations that qualitatively capture the migration history in "out of model" coalescent simulations.

3.3.3 Application of FEEMS to genotype data from North American gray wolves

To assess the performance of FEEMS on real data we used a previously published dataset of 111 gray wolves sampled across North America typed at $17,729$ SNPs (Schweizer et al., 2016), Supp. Fig. 3.9). This dataset has a number of advantageous features that make it a useful test case for evaluating FEEMS: (1) The broad sampling range across North America includes a number of relevant geographic features that, a priori, could conceivably lead to restricted gene-flow averaged throughout the population history. These geographic features include mountain ranges, lakes and island chains. (2) The scale of the data is consistent with many studies for non-model systems whose spatial population structure is of interest. For instance, the relatively sparse sampling leads to a challenging statistical problem where there is the potential for many unobserved nodes (sub-populations), depending the density of the grid chosen. Before applying FEEMS, we confirmed a signature of spatial structure in

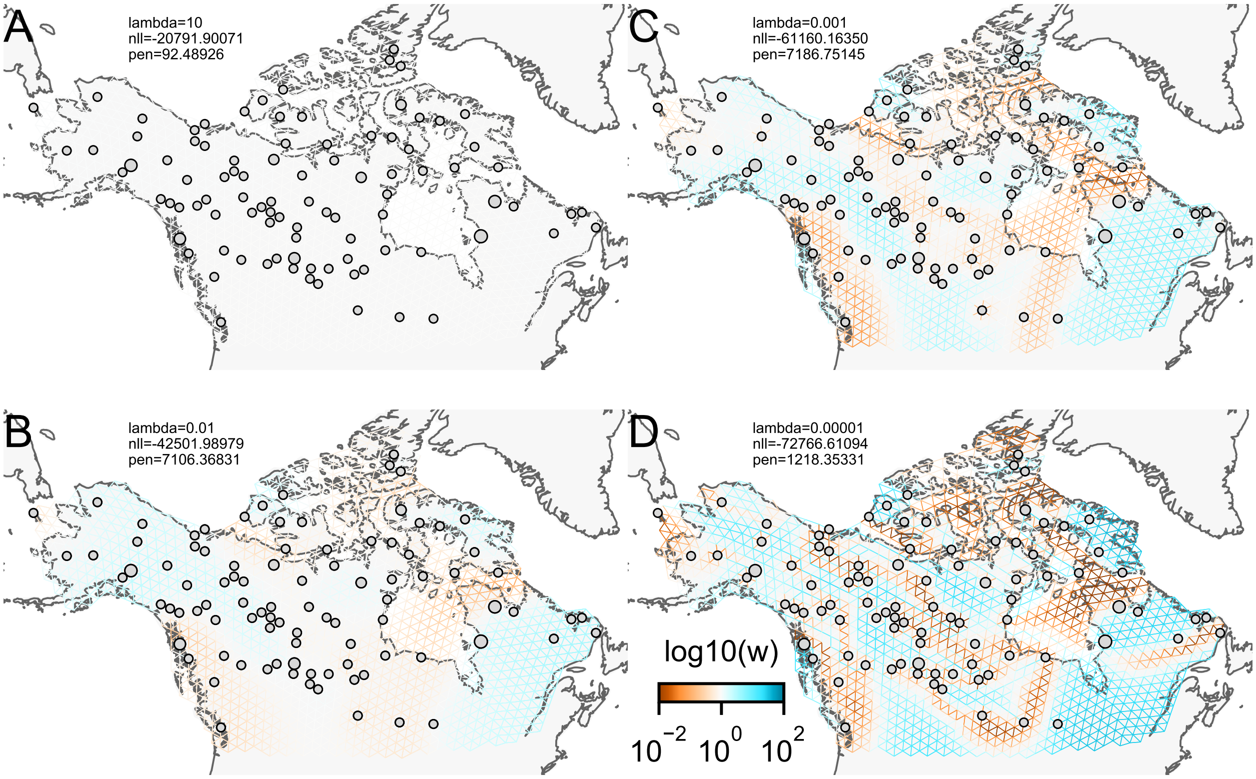


Figure 3.3: **The fit of FEEMS to the North American gray wolf dataset for different choices of the smoothing regularization parameter λ:** (A) $\lambda = 10$, (B) $\lambda = 10^{-2}$, (C) $\lambda = 10^{-3}$, and (D) $\lambda = 10^{-5}$. As expected, when λ decreases from large to small (A-D), the fitted graph becomes less smooth and presumably eventually over-fits to the data, revealing a patchy surface in (D), whereas, earlier in the regularization path FEEMS fits a completely homogeneous surface with all edge weights having the same fitted value, like in (A).

the data through regressing genetic distances on geographic distances and top genetic PCs against geographic coordinates (Supp. Fig. 3.10, 3.11, 3.12, 3.13).

We ran FEEMS with four different values of the smoothness parameter, λ (from large $\lambda = 10$ to small $\lambda = 10^{-5}$), while setting the tuning parameter α to a value that we found that worked for multiple data applications and simulations ($\alpha = 50$, Figure 3.3). One interpretation of our regularization penalty is that it encourages fitting models of homogeneous and isotropic migration. When λ is very large (Figure 3.3A), we see FEEMS fits a model where all of the edge weights on the graph nearly equal the mean value, hence all the edge weights are colored white in the relative log-scale. In this case, FEEMS is fitting a completely homogeneous migration model where all the estimated edge weights get assigned the same value on the graph. Next, as we sequentially lower the penalty parameter and (Figure 3.3B,C,D) the fitted graph begins to appear more complex and heterogeneous as expected (discussed further below).

We also ran multiple replicates of ADMIXTURE for $K = 2$ to $K = 8$, selecting for each K the highest likelihood run among replicates to visualize (Supp. Fig. 3.14). As expected in a spatial genetic dataset, nearby samples have similar admixture proportions and continuous gradients of changing ancestries are spread throughout the map (Bradburd et al., 2018). Whether such gradients in admixture coefficients are due to isolation by distance or specific geographic features that enhance or diminish the levels of genetic differentiation is an interpretive challenge. Explicitly modeling the spatial locations and genetic distance jointly using a method like EEMS or FEEMS is exactly designed to explore and visualize these types of questions in the data (Petkova, 2013; Petkova et al., 2016).

Once we have run FEEMS for a grid of regularization parameters it is helpful to look more closely at particular solutions that find a balance between spatial homogeneity and complexity (Figure 4.5). Spatial features in the FEEMS visualization qualitatively matches the structure plot output from ADMIXTURE using $K = 6$ (Supp. Fig. 3.14). We add labels on the figure to highlight a number of pertinent features: (A) St. Lawrence Island, (B) the coastal islands and mountain ranges in British Columbia, (C) The boundary of Boreal Forest and Tundra eco-regions in the Shield Taiga, (D) Queen Elizabeth Islands, (E) Hudson Bay, and (F) Baffin Island. Many of these features were described in Schweizer et al. (2016) by interpretation of ADMIXTURE, PCA, and F_{ST} statistics. FEEMS is able to succinctly provide an interpretable view of these data in a single visualization. Indeed many of these geographic features plausibly impact gray wolf dispersal and population history (Schweizer et al., 2016).

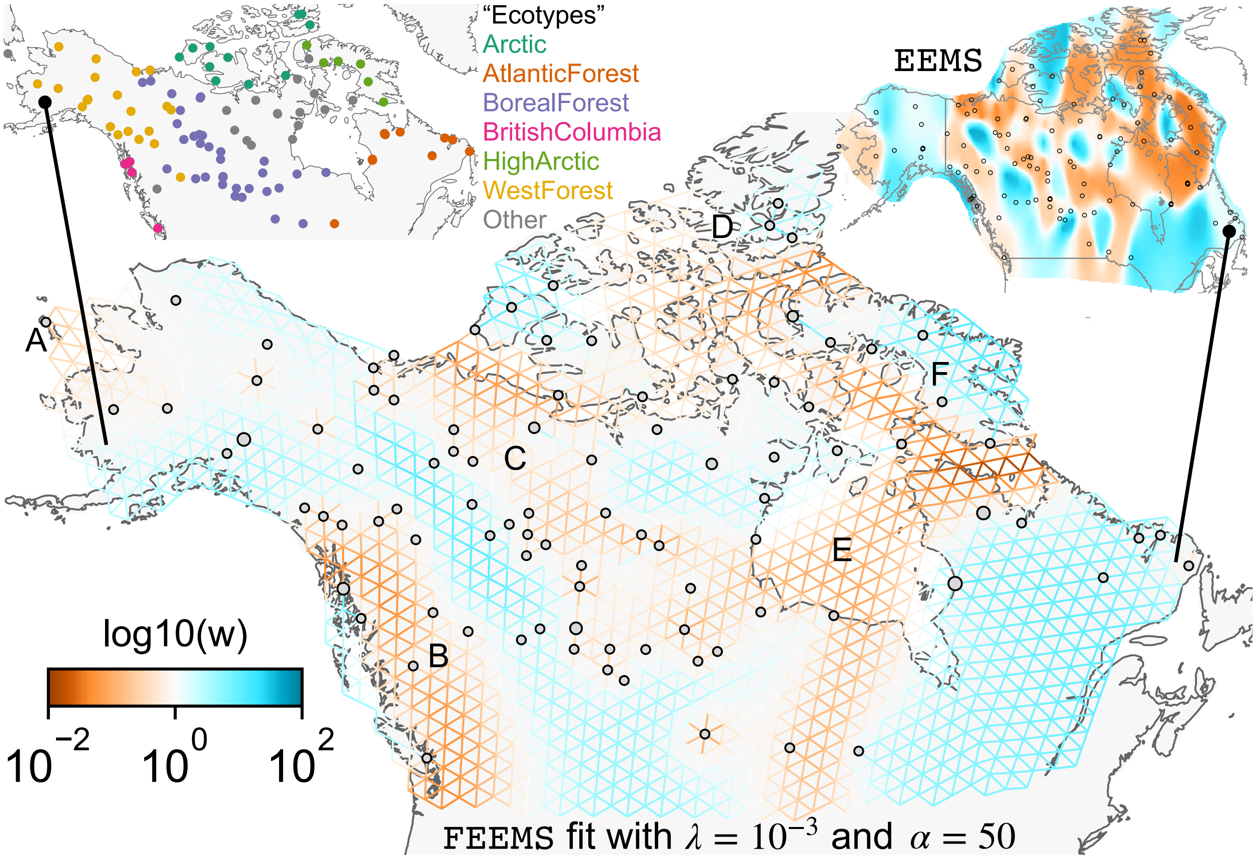


Figure 3.4: **FEEMS applied to a population genetic dataset of North American gray wolves:** We show the fit of FEEMS to a previously published dataset of North American gray wolves. This fit corresponds to a setting of tuning parameters at $\lambda = 10^{-3}, \alpha = 50$. We show the fitted parameters in log-scale with lower effective migration shown in orange and higher effective migration shown in blue. The bold text letters highlights a number of known geographic features that could have plausibly influenced Wolf migration over time: (A) St. Lawerence Island, (B) Coastal mountain ranges in British Columbia, (C) The boundary of Boreal Forest and Tundra eco-regions in the Shield Taiga, (D) Queen Elizabeth Islands, (E) Hudson Bay, and (F) Baffin Island. We also display two insets to help interpret the results and compare them to EEMS. In the top left inset we show a map of sample coordinates colored by an ecotype label provided by Schweizer et al. (2016). These labels were devised using a combination of genetic and ecological information for 94 "un-admixed" gray wolf samples, and the remaining samples were labeled "Other". We can see these ecotype labels align well with the visualization output provided by FEEMS. In the right inset we display a visualization of the posterior mean effective migration rates from EEMS.

3.3.4 Comparison to EEMS

We also ran EEMS on the same gray wolf dataset described throughout this manuscript. We used default parameters provided by EEMS but set the number of burn-in iterations to 20×10^6, MCMC iterations to 50×10^6, and thinning intervals to 2000. We were unable to run EEMS in a reasonable run time ($\leqslant$ 3 days) for the dense spatial grid of 1207 nodes so we ran EEMS and FEEMS on a sparser graph with 307 nodes.

We find that FEEMS is multiple orders of magnitude faster than EEMS, even when running multiple runs of FEEMS for different regularization settings on both the sparse and dense graphs (Table 3.1). The total FEEMS run-times in Table 3.1 also include the time needed to construct relevant graph data structures and initialization. We note that constructing the graph and fitting the model with very low regularization parameters are the most computationally demanding steps in running FEEMS.

We find that many of the same geographic features that have reduced or enhanced gene-flow are concordant between the two methods. The EEMS visualization, qualitatively, best matches solutions of FEEMS with lower regularization penalties (Figure 4.5, Supp. Fig. 3.15); however, based on the ADMIXTURE results and visual inspection in relation to known geographical features, we find these solutions to be less satisfying compared to those with higher penalties and believe the solutions output from lower penalties are likely overfitting the data. Indeed, we only see a small gain in the R^2 when comparing observed and fitted distances computed from the output graphs of Figure 3.3C and Figure 3.3D (Supp. Fig. 3.10). We note that in many of the EEMS runs the MCMC appears to not have converged (based on visual inspection of trace plots) even after a large number of iterations.

3.4 Discussion

FEEMS is a fast approach that provides an interpretable view of spatial population structure in real datasets and simulations. We want to emphasize that beyond being a fast

Method	Sparse Grid (run-time)	Dense Grid (run-time)
EEMS	27.43hrs	N/A
FEEMS (total)	13.02s	3.54min
FEEMS (init)	8.25s	2min 11s
FEEMS ($\lambda = 10$)	604ms	10.7s
FEEMS ($\lambda = 10^{-2}$)	442ms	7.78s
FEEMS ($\lambda = 10^{-3}$)	917ms	9.18s
FEEMS ($\lambda = 10^{-5}$)	2.81s	53.9s

Table 3.1: **Runtimes for FEEMS and EEMS on the North American gray wolf dataset**: We show a table of runtimes for FEEMS and EEMS for two different grid densities, a sparse grid with 307 nodes and a dense grid with 1207 nodes. In the first two rows we show the total runtimes for both EEMS and FEEMS. In the following rows we show the total runtime for FEEMS, broken down into multiple components i.e. initialization time and the time to fit four solutions with different smoothing parameters.

optimization approach for inferring population structure, our parameterization of the likelihood opens up a number of exciting new directions for improving spatial population genetic inference. Notably, one major difference between EEMS and FEEMS is that in FEEMS each edge weight is assigned its own parameter to be estimated whereas, in EEMS, each node is assigned a parameter and each edge is constrained to be the average effective migration between the nodes it connects (see Materials and Methods and Supp. Note "*Edge versus node parameterization*" for details). The node-based parameterization in EEMS makes it difficult to incorporate anisotropy and asymmeteric migration (Lundgren and Ralph, 2019). As we have shown here, FEEMS's simple and novel parameterization already has potential to fit anisotropic migration (as shown in coalescent simulations) and may be extendable to other more complex migration processes (such as long-range migration, see below).

FEEMS estimates one set of graph edge weights for each setting of the tuning parameters λ and α which control the smoothness of the fitted edge-weights. One general challenge, which is not unique to this method, is selecting a particular set of tuning parameters. A natural approach is to use cross-validation, which estimates the out-of-sample fit of FEEMS for a particular model (selection of λ and α). While cross-validation might be useful for assessing the choice of tuning parameters, in preliminary experiments applying cross validation by holding out individuals or observed nodes, and assessing performance via the model-likelihood, we found too much variation across cross-validation folds to reliably tune λ and α (results not shown). In order to reduce the variation across different folds, we also applied cross-validation with standardization (Bradburd et al., 2018), where the model-likelihood is standardized for each fold, and approximate leave-one-out cross-validation Wilson et al. (2020), where the leave-one-out CV likelihood is approximated with a few steps of the quasi-Newton algorithm warm-started from the full training set migration surfaces. Neither of these approaches were promising for reliable model selection. We suspect this poor performance is due to spatial dependency of allele frequencies and the large fraction of unobserved nodes. In unsupervised learning settings like this one, it is not obvious that estimates of out of sample fit will always lead to the most biologically interpretable models and sometimes other metrics can be preferable, such as those based on the stability of the solution to perturbations like variable initialization (Wu et al., 2016). Stability-based approaches for model selection could be a fruitful future direction to develop a formal procedure for tuning. Currently, we recommend fitting FEEMS with several values of the tuning parameters and interpreting the results in an integrative fashion with other analyses.

We find it useful to fit FEEMS to a sequential grid of regularization parameters and to look at what features are consistent and vary across multiple fits. Informally, one can gain an indication of the strongest features in the data by looking at the order they appear in the regularization path i.e. what features overcome the strong penalization of smoothness in the data and that are highly supported by the likelihood. For example, early in the regularization path, we see regions of reduced gene-flow occurring in the west coast of Canada that presumably correspond to Coastal mountain ranges and islands in British Columbia (Figure 3.3B) and this reduced gene-flow appears throughout more flexible fits with lower λ.

Beyond tuning the unknown parameters, we encountered other challenges when solving this difficult optimization problem. Notably, the objective function we optimize is non-convex so any visualization output by FEEMS should be considered a local optimum and, as a result, with different initialization we could get different results. Overall, we found the output visualization was not sensitive to initialization, and thus our default setting is constant initialization fitted under an homogeneous isolation by distance model (See Materials and Methods).

When comparing to EEMS, we found FEEMS to be much faster (Table 3.1). While this is encouraging, care must be taken because the goals and outputs of FEEMS and EEMS have a number of differences. FEEMS fits a sequential grid of solutions for different regularization parameters whereas EEMS infers a posterior distribution and outputs the posterior mean as a point estimate. So in order to compare the results, in principal, one must compare many FEEMS visualizations to a single EEMS visualization. FEEMS is not a Bayesian method and unlike EEMS, which explores the entire landscape of the posterior distribution, FEEMS returns a particular point estimate: a local minimum point of the optimization landscape. Setting the prior hyper-parameters in EEMS act somewhat like a choice of tuning parameters, except that EEMS uses hierarchical priors that in principle allow for exploration of multiple scales of spatial structure in a single run; this arguably results in less sensitivity to user-based settings but requires potentially long computation times for adequate MCMC convergence.

One natural extension to FEEMS, pertinent to a number of biological systems, is incorporating long-range migration (Pickrell and Pritchard, 2012; Bradburd et al., 2016). In this work we have used a triangular lattice embedded in geographic space and enforced smoothness in nearby edge weights through penalizing their squared differences (see Materials and Methods). We could imagine changing the structure of the graph by adding edges to allow for long-range connection; however our current regularization scheme would not be appropriate for this setting. Instead, we could imagine adding an additional penalty to the objective, which would only allow a few long range connections to be tolerated. This could be considered to be a combination of two existing approaches for graph-based inference, graphical lasso (GLASSO) and graph Laplacian smoothing, combining the smoothness assumption for nearby connections and the sparsity assumption for long-range connections (Friedman et al., 2008; Wang et al., 2016). Another potential methodological avenue to incorporate long-range migration is to use a “greedy” approach. We could imagine adding long-range edges one a time, guided by re-fitting the spatial model and taking a data driven approach to select particular long-range edges to include. The proposed greedy approach could be considered to be a spatial graph analog of TreeMix (Pickrell and Pritchard, 2012).

Another interesting extension would be to incorporate asymmetric migration into the framework of resistance distance and Gaussian Markov Random Field based models. Recently, Hanks (2015) developed a promising new framework for deriving the stationary distribution of a continuous time stochastic process with asymmetric migration on a spatial graph. Interestingly, the expected distance of this process has a similar “flavor” to the resistance distance based models, in that it depends on the pseudo-inverse of a function of the graph Laplacian. Hanks (2015) used MCMC to estimate the effect of known covariates on the edge weights of the spatial graph. Future work could adapt this framework into the penalized optimization approach we have considered here, where adjacent edge weights are encouraged to be smooth.

Finally, when interpreted as mechanistic rather than statistical models, both EEMS and FEEMS implicitly assume time-stationarity, so the estimated migration parameters should be considered to be “effective” in the sense of being averaged over time in a reality where migration rates are dynamic and changing (Pickrell and Reich, 2014). The MAPS method is one recent advance that utilizes long stretches of shared haplotypes between pairs of individuals to perform Bayesian inference of time varying migration rates and population sizes (Al-Asadi et al., 2019). With the growing ability to extract high quality DNA from ancient samples, another exciting future direction would be to apply FEEMS to ancient

DNA datasets over different time transects in the same focal geographic region to elucidate changing migration histories (Mathieson et al., 2018). There are a number of technical challenges in ancient DNA data that make this a difficult problem, particularly high levels of missing and low-coverage data. Our modeling approach could be potentially more robust, in that it takes allele frequencies as input, which may be estimable from dozens of ancient samples at the same spatial location, in spite of high degrees of missingness (Korneliussen et al., 2014).

In closing, we look back to a review titled "How Can We Infer Geography and History from Gene Frequencies?" published in 1982 (Felsenstein, 1982). In this review, Felsenstein laid out fundamental open problems in statistical inference in population genetic data, a few of which we restate as they are particularly motivating for our work:

- "For any given covariance matrix, is there a corresponding migration matrix which would be expected to lead to it? If so, how can we find it?"
- "How can we characterize the set of possible migration matrices which are compatible with a given set of observed covariances?"
- "How can we confine our attention to migration patterns which are consistent with the known geometric co-ordinates of the populations?"
- "How can we make valid statistical estimates of parameters of stepping stone models?"

The methods developed here aim to help address these longstanding problems in statistical population genetics and to provide a foundation for future work to elucidate the role of geography and dispersal in ecological and evolutionary processes.

3.5 Methods

3.5.1 Model description

See Supp. Note "*Mathematical notation*" for a detailed description of the notation used to describe the model. To visualize and model spatial patterns in a given population genetic dataset, FEEMS uses an undirected graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = d$, where nodes represent sub-populations and edge weights $(w_{ij})_{(i,j)\in\mathcal{E}}$ represent the level of gene-flow between sub-populations i and j. For computational convenience, we assume $\mathcal{G}$ is a highly sparse graph, specifically a triangular grid that is embedded in geographic space around the sample coordinates. We observe a genotype matrix, $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, with n rows representing individuals and p columns representing SNPs. We imagine diploid individuals are sampled on the nodes of $\mathcal{G}$ so that $y_{ij}(k) \in \{0, 1, 2\}$ records the count of some arbitrarily predefined allele in individual i, SNP j, on node $k \in \mathcal{V}$. We assume a commonly used simple Binomial sampling model for the genotypes:

$$y_{ij}(k)|f_j(k) \sim \text{Binomial}\big(2, f_j(k)\big), \tag{3.1}$$

where conditional on $f_j(k)$ for all j, k, the $y_{ij}(k)$'s are independent. We then estimate an allele frequency at each node and SNP by maximum likelihood:

$$\widehat{f}_j(k) = \frac{\sum_{i=1}^{n_k} y_{ij}(k)}{2n_k},$$

where n_k is the number of individuals sampled at node k. We estimate allele frequencies at o of the observed nodes out of d total nodes on the graph. From (3.1), the estimated frequency in a particular sub-population, conditional on the latent allele frequency, will approximately follow a Gaussian distribution:

$$\widehat{f}_j(k)|f_j(k) \sim \mathcal{N}\left(f_j(k), \frac{f_j(k)\big(1-f_j(k)\big)}{2n_k}\right).$$

Using vector notation, we represent the joint model of estimated allele frequencies as:

$$\widehat{\boldsymbol{f}}_j|\boldsymbol{f}_j \sim \mathcal{N}_o\Big(\boldsymbol{A}\boldsymbol{f}_j, \mathrm{diag}\big(\boldsymbol{d}_{\boldsymbol{f},\boldsymbol{n}}\big)\Big), \tag{3.2}$$

where $\widehat{\boldsymbol{f}}_j$ is a $o \times 1$ vector of estimated allele frequencies at observed nodes, $\boldsymbol{f}_j$ is a $d \times 1$ vector of latent allele frequencies at all the nodes (both observed and unobserved), and $\boldsymbol{A}$ is a $o \times d$ node assignment matrix where $\boldsymbol{A}_{k\ell} = 1$ if the kth estimated allele frequency comes from sub-population ℓ and $\boldsymbol{A}_{k\ell} = 0$ otherwise; and $\mathrm{diag}(\boldsymbol{d}_{\boldsymbol{f},\boldsymbol{n}})$ denotes a $o \times o$ diagonal matrix whose diagonal elements corresponds to the appropriate variance term at observed nodes.

To summarize, we estimate allele frequencies from a subset of nodes on the graph and define latent allele frequencies for all the nodes of the graph. The assignment matrix $\boldsymbol{A}$ maps these latent allele frequencies to our observations. Our summary statistics (the data) are thus $(\widehat{\boldsymbol{F}}, \boldsymbol{n})$ where $\widehat{\boldsymbol{F}}$ is a $o \times p$ matrix of estimated allele frequencies and $\boldsymbol{n}$ is a $o \times 1$ vector of sample sizes for every observed node. We assume the latent allele frequencies come from a Gaussian Markov Random Field:

$$\boldsymbol{f}_j \sim \mathcal{N}_d\Big(\mu_j\boldsymbol{1}, \mu_j(1-\mu_j)\boldsymbol{L}^\dagger\Big), \tag{3.3}$$

where $\boldsymbol{L}$ is the graph Laplacian and μ_j represents the average allele frequency across all of the sub-populations. Note that the multiplication by the SNP-specific factor $\mu_j(1-\mu_j)$ ensures that the variance of the latent allele frequencies vanishes as the average allele frequency approaches to 0 or 1. One interpretation of this model is that the expected squared Euclidean distance between latent allele frequencies on the graph, after being re-scaled by $\mu_j(1-\mu_j)$, is exactly the resistance distance of an electrical circuit (McRae, 2006; Hanks and Hooten, 2013):

$$r_{j,ik} = \frac{\mathbb{E}\Big[\big(f_j(i) - f_j(k)\big)^2\Big]}{\mu_j(1-\mu_j)} = (\boldsymbol{o}_i - \boldsymbol{o}_k)^\top \boldsymbol{L}^\dagger (\boldsymbol{o}_i - \boldsymbol{o}_k) = \boldsymbol{L}^\dagger_{ii} - 2\boldsymbol{L}^\dagger_{ik} + \boldsymbol{L}^\dagger_{kk},$$

where $\boldsymbol{o}_i$ is a one-hot vector (i.e., storing a 1 in element i and zeros elsewhere). It is known that the resistance distance is equivalent to the expected commute time between nodes i and k of a random walker on the weighted graph $\mathcal{G}$ (Chandra et al., 1996). Additionally, the model (3.3) forms a Markov random field, and thus any latent allele frequency $f_j(i)$ is conditionally independent of all other allele frequencies given its neighbors which are encoded by nonzero elements of $\boldsymbol{L}$ (Lauritzen, 1996; Koller and Friedman, 2009).[1]

Using the law of total variance formula, we can derive from (3.2), (3.3) an analytic form for the marginal likelihood. Before proceeding, however, we further approximate the model by assuming $\frac{1}{2} f_j(k)(1 - f_j(k)) \approx \sigma^2 \mu_j(1-\mu_j)$ for all j and k. This assumption is mainly for computational purposes and may be a coarse approximation in general. On the other hand, the assumption is not too strong if we exclude SNPs with extremely rare allele frequencies, and more importantly, we find it leads to a good empirical performance, both statistically and computationally. With this approximation the residual variance parameter σ^2 is still unknown and needs to be estimated.

With the above considerations, we arrive at the following marginal likelihood:[2]

$$\widehat{\boldsymbol{f}}_j \sim \sqrt{\mu_j(1-\mu_j)} \cdot \mathcal{N}_o\Big(\mu_j \boldsymbol{1}, \boldsymbol{A}\boldsymbol{L}^\dagger\boldsymbol{A}^\top + \sigma^2 \mathrm{diag}\big(\boldsymbol{n}^{-1}\big)\Big), \tag{3.4}$$

where $\mathrm{diag}(\boldsymbol{n}^{-1})$ is a $o \times o$ diagonal matrix computed from the sample sizes at observed

[1] Specifically, since we use a triangular grid embedded in geographic space to define the graph $\mathcal{G}$, the pattern of nonzero elements is prefixed by the structure of the sparse traingular grid.

[2] To be more precise, under (3.2), (3.3), the law of total variance formula leads to specific formulas for the mean and variance structure as given in (3.4), whereas the marginal distribution of $\widehat{f}_j$ is not necessarily a Gaussian distribution. We simply chose the Gaussian distribution here to enable easy calculation for the data likelihood. We believe the specific choice of the likelihood is not that critical as long as the first two moments of the distribution can be matched closely.

nodes. To remove the SNP means we transform the estimated frequencies by a contrast matrix, $\boldsymbol{C} \in \mathbb{R}^{(o-1)\times o}$, that is orthogonal to the one-vector:

$$\boldsymbol{C}\widehat{\boldsymbol{f}}_j \sim \sqrt{\mu_j(1-\mu_j)} \cdot \mathcal{N}_{o-1}\Bigg(\mathbf{0}, \boldsymbol{CAL}^{\dagger}\boldsymbol{A}^{\top}\boldsymbol{C}^{\top} + \sigma^2\boldsymbol{C}\mathrm{diag}\big(\boldsymbol{n}^{-1}\big)\boldsymbol{C}^{\top}\Bigg). \tag{3.5}$$

Letting $\widehat{\boldsymbol{\Sigma}} = \frac{1}{p}\widehat{\boldsymbol{F}_{\mathrm{S}}}\widehat{\boldsymbol{F}_{\mathrm{S}}}^{\top}$ be the $o \times o$ sample covariance matrix of estimated allele frequencies after rescaling, i.e. $\widehat{\boldsymbol{F}_{\mathrm{S}}}$ is a matrix formed by rescaling the columns of $\widehat{\boldsymbol{F}}$ by $\sqrt{\widehat{\mu}_j(1-\widehat{\mu}_j)}$, where $\widehat{\mu}_j$ is an estimate of the average allele frequency (see above). We can then express the model in terms of the transformed sample covariance matrix:

$$p \cdot \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top} \sim \mathcal{W}_{o-1}\Bigg(\boldsymbol{CAL}^{\dagger}\boldsymbol{A}^{\top}\boldsymbol{C}^{\top} + \sigma^2\boldsymbol{C}\mathrm{diag}\big(\boldsymbol{n}^{-1}\big)\boldsymbol{C}^{\top}, p\Bigg), \tag{3.6}$$

where $\mathcal{W}_p$ denotes a Wishart distribution with p degrees of freedom.[3] Note we can equivalently use the sample squared Euclidean distance (often refereed to as a genetic distance) as a summary statistic: letting $\widehat{\boldsymbol{D}}$ be the genetic distance matrix with $\boldsymbol{D}_{ik} = \sum_{j=1}^{p}(\widehat{f}_j(i) - \widehat{f}_j(k))^2/p \cdot \widehat{\mu}_j(1-\widehat{\mu}_j)$, we have

$$\widehat{\boldsymbol{D}} = \mathbf{1}\mathrm{diag}(\widehat{\boldsymbol{\Sigma}})^{\top} + \mathrm{diag}(\widehat{\boldsymbol{\Sigma}})\mathbf{1}^{\top} - 2\widehat{\boldsymbol{\Sigma}},$$

and so

$$\boldsymbol{C}\widehat{\boldsymbol{D}}\boldsymbol{C}^{\top} = -2\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top},$$

using the fact that the contrast matrix $\boldsymbol{C}$ is orthogonal to the one-vector. Thus we can use the same spatial covariance model implied by the allele frequencies once we project the

[3] Our model (3.6) says that the p SNPs are independent. This assumption is unlikely to hold when SNPs are in close chromosomal proximity are analyzed due to linkage disequilibrium. In (Petkova et al., 2016), they introduce the effective degree of freedom $\nu \in [o-1, p]$ to account for such dependency and instead consider the model $\nu \cdot \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top} \sim \mathcal{W}_{o-1}(\boldsymbol{CAL}^{\dagger}\boldsymbol{A}^{\top}\boldsymbol{C}^{\top} + \sigma^2\boldsymbol{C}\mathrm{diag}\big(\boldsymbol{n}^{-1}\big)\boldsymbol{C}^{\top}, \nu)$ with ν being estimated alongside other model parameters. In FEEMS, we note that the degree of freedom parameter does not affect the point estimate produced by our algorithm.

distances on to the space of contrasts:[4]

$$-\frac{p}{2}\cdot \boldsymbol{C}\widehat{\boldsymbol{D}}\boldsymbol{C}^\top \sim \mathcal{W}_{o-1}\left(\boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^\dagger\boldsymbol{A}^\top\boldsymbol{C}^\top + \sigma^2\boldsymbol{C}\mathrm{diag}\big(\boldsymbol{n}^{-1}\big)\boldsymbol{C}^\top, p\right).$$

Overall, the negative log-likelihood function implied by our spatial model is (ignoring constant terms):

$$\begin{aligned}\ell(\boldsymbol{w},\sigma^2;\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) = p\cdot \mathrm{tr}\left(\left(\boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^\dagger\boldsymbol{A}^\top\boldsymbol{C}^\top + \sigma^2\boldsymbol{C}\mathrm{diag}(\boldsymbol{n}^{-1})\boldsymbol{C}^\top\right)^{-1}\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top\right) \\ - p\cdot\log\det\left(\boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^\dagger\boldsymbol{A}^\top\boldsymbol{C}^\top + \sigma^2\boldsymbol{C}\mathrm{diag}(\boldsymbol{n}^{-1})\boldsymbol{C}^\top\right)^{-1}, \quad (3.7)\end{aligned}$$

where $\boldsymbol{w} \in \mathbb{R}^m$ is a vectorized form of the non-zero lower-triangular entries of the weighted adjacency matrix $\boldsymbol{W}$ (recall that the graph Laplacian is completely defined by the edge weights $\boldsymbol{L} = \mathrm{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W}$ so there is an implicit dependency here). Since the graph is a triangular lattice, we only need to consider the non-zero entries to save computational time, i.e. not all sub-populations are connected to each other.

One key difference between EEMS (Petkova et al., 2016) and FEEMS is how the edge weights are parameterized. In EEMS, each node is given an effective migration parameter m_i for node $i \in \mathcal{V}$ and the edge weight is paramertized as the average between the nodes it connects, i.e. $w_{ij} = (m_i + m_j)/2$ for $(i,j) \in \mathcal{E}$. FEEMS, on the other hand, assigns a parameter to every nonzero edge-weight. The former has fewer parameters, with the specific consequence that it only allows isotropy and imposes an additional degree of similarity among edge weights; instead, in the latter, the edge weights are free to vary apart from the

[4] We remark that besides the effective degree of freedom and the SNP-specific re-scaling by $\mu_j(1-\mu_j)$, the EEMS (Petkova et al., 2016) and FEEMS likelihoods are equivalent up to constant factors, as long as only one individual is observed per node and the residual variance σ^2 is allowed to vary across nodes—See Supp. Note "*Jointly estimating the residual variance and edge weights*" for details. In addition, constant factors are effectively absorbed into the unknown model parameters $\boldsymbol{L}$ and σ^2 and therefore it does not affect the estimation of effective migration rates, up to constant factors.

regularization imposed by the penalty. See Supp. Note "*Edge versus node parameterization*" and Supp. Fig. 3.20 for more details.

3.5.2 Penalty description

As mentioned previously we would like to encourage that nearby edge weights on the graph have similar values to each other. This can be performed by penalizing the squared differences between all edges connected to the same node, i.e. spatially adjacent edges:

$$\phi_{\lambda,\alpha}(\boldsymbol{w}) = \frac{\lambda}{2} \sum_{i \in \mathcal{V}} \sum_{k,\ell \in \mathcal{E}(i)} \left(\Big(w_{ik} + \alpha \log(w_{ik}) \Big) - \Big(w_{i\ell} + \alpha \log(w_{i\ell}) \Big) \right)^2,$$

where $\phi_{\lambda,\alpha}$ is our penalty function that represents the total amount of smoothness on the graph and $\mathcal{E}(i)$ denotes the set of edges that connected to node i. Here we penalize a weighted combination of the edge weights on the original scale and logarithmic-scale where α, a tuning parameter, controls how strong the penalization is placed on the logarithmic scale—in the special case that $\alpha = 0$, it reduces to the commonly used Laplacian smoothing-type penalty. Adding a logarithmic scale leads to smooth graphs for small edge values and thus allow for an additional degree of flexibility across orders of magnitude of edge weights. The smoothness parameter, λ, controls the overall contribution of the penalty to the objective function. It is convenient to write the penalty in matrix-vector form which we will use throughout:

$$\phi_{\lambda,\alpha}(\boldsymbol{w}) = \frac{\lambda}{2} \|\boldsymbol{\Delta}(\boldsymbol{w} + \alpha \log(\boldsymbol{w})\|_2^2, \tag{3.8}$$

where $\boldsymbol{\Delta}$ is a signed graph incidence matrix derived from a unweighted graph denoting if pairs of edges are connected to the same node. This penalty function (3.8) is also scale invariant, in the sense that for any $c > 0$, $\phi_{\lambda,\alpha}(\boldsymbol{w}) = \phi_{c^{-2}\lambda,c\alpha}(c\boldsymbol{w})$.

One might wonder whether it is possible to use the ℓ_1 norm in the penalty form (3.8) in place of the ℓ_2 norm. While it is known that the ℓ_1 norm might increase local adaptivity and better capture the sharp changes of the underlying structure of the latent allele frequencies, (e.g. Wang et al., 2016), in our case, we found an inferior performance when using the ℓ_1 norm over the ℓ_2 norm—in particular, our primary application of interest is the regime of highly missing nodes, i.e. $o \ll d$, in which case the global smoothing seems somewhat necessary to encourage stable recovery of the edge weights at regions with sparsely observed nodes (see Supp. Note "*Smooth penalty with ℓ_1 norm*"). In addition, adding the penalty $\phi_{\lambda,\alpha}(\boldsymbol{w})$ allows us to implement faster algorithms to solve the optimization problem due to the differentiability of the ℓ_2 norm, and as a result, it leads to better overall computational savings and a simpler implementation.

3.5.3 Optimization

Putting (3.7) and (3.8) together, we infer the migration edge weights $\widehat{\boldsymbol{w}}$ by minimizing the following penalized negative log-likelihood function:

$$\begin{aligned}\widehat{\boldsymbol{w}} &= \underset{\boldsymbol{l} \leqslant \boldsymbol{w} \leqslant \boldsymbol{u}}{\arg\min}\ \ell(\boldsymbol{w}, \sigma^2; \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) + \phi_{\lambda,\alpha}(\boldsymbol{w}) \\ &= \underset{\boldsymbol{l} \leqslant \boldsymbol{w} \leqslant \boldsymbol{u}}{\arg\min}\ \Bigg[p \cdot \mathrm{tr}\Bigg(\Big(\boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^\dagger \boldsymbol{A}^\top \boldsymbol{C}^\top + \sigma^2 \boldsymbol{C}\mathrm{diag}(\boldsymbol{n}^{-1})\boldsymbol{C}^\top \Big)^{-1} \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top \Bigg) \qquad (3.9) \\ &\qquad - p \cdot \log\det \Bigg(\boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^\dagger \boldsymbol{A}^\top \boldsymbol{C}^\top + \sigma^2 \boldsymbol{C}\mathrm{diag}(\boldsymbol{n}^{-1})\boldsymbol{C}^\top \Bigg)^{-1} + \frac{\lambda}{2} \|\boldsymbol{\Delta}(\boldsymbol{w} + \alpha \log(\boldsymbol{w})\|_2^2 \Bigg],\end{aligned}$$

where $\boldsymbol{l}, \boldsymbol{u} \in \mathbb{R}_+^m$ represent respectively the entrywise lower- and upper bounds on $\boldsymbol{w}$, i.e. we constrain the lower- and upper bound of the edge weights to $\boldsymbol{l}$ and $\boldsymbol{u}$ throughout the optimization. When no prior information is available on the range of the edge weights, we often set $\boldsymbol{l} = \boldsymbol{0}$ and $\boldsymbol{u} = +\infty$.

One advantage of the formulation of (3.9) is the use of the vector form parameterization $\boldsymbol{w} \in \mathbb{R}_+^m$ of the symmetric weighted adjacency matrix $\boldsymbol{W} \in \mathbb{R}_+^{d \times d}$. In our triangular graph

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the number of non-zero lower-triangular entries is $m = \mathcal{O}(d) \ll d^2$, so working directly on the space of vector parameterization saves computational cost. In addition, this avoids the symmetry constraint imposed on the adjacency matrix $\boldsymbol{W}$, hence making optimization easier (Kalofolias, 2016).

We solve the optimization problem using a constrained quasi-Newton optimization algorithm, specifically L-BFGS implemented in `scipy` (Byrd et al., 1995; Virtanen et al., 2020).[5] Since our objective (3.9) is non-convex, the L-BFGS algorithm is guaranteed to converge only to a local minimum. Even so, we empirically observe that local minima starting from different initial points are qualitatively similar to each other across many datasets. The L-BFGS algorithm requires gradient and objective values as inputs. Note the naive computation of the objective (3.9) is computationally prohibitive since inverting the graph Laplacian has complexity $\mathcal{O}(d^3)$. We take advantage of the sparsity of the graph and specific structure of the problem to efficiently compute gradient and objective values. In theory, our implementation has computational complexity of $\mathcal{O}(do + o^3)$ per iteration which, in the setting of $o \ll d$, is substantially smaller than $\mathcal{O}(d^3)$.[6]

3.5.4 Estimating the residual variance and edge weights under the null model

For estimating the residual variance parameter σ^2, we first estimate it via maximum likelihood assuming homogeneous isolation by distance. This corresponds to the scenario where every edge-weight in the graph is given the exact same unknown parameter value w_0. Under this model we only have two unknown parameters w_0 and the residual variance σ^2. We

[5] We solve using linearized ADMM when the penalty function is ℓ_1 norm, i.e. $\lambda \|\boldsymbol{\Delta}((w) + \alpha \log((w))\|_1$ (Boyd et al., 2011).

[6] More precisely, it is possible to achieve $\mathcal{O}(do + o^3)$ per-iteration complexity if one employs a solver that is specially designed for sparse Laplacian system. In our work we use sparse Cholesky factorization which may slightly slow down the per-iteration complexity. See Supp. Material for the details of the gradient and objective computation.

estimate these two parameters by jointly optimizing the marginal likelihood using a Nelder-Mead algorithm implemented in `scipy` (Virtanen et al., 2020). This requires only likelihood computations which are efficient due to the sparse nature of the graph. This optimization routine outputs an estimate of the residual variance $\widehat{\sigma}^2$ and the null edge weight $\hat{w}_0$, which can be used to construct $\boldsymbol{W}(\widehat{w}_0)$ and in turn $\boldsymbol{L}(\widehat{w}_0)$.

One strategy we found effective is to fit the model of homogeneous isolation by distance and then fix the estimated residual variance $\widehat{\sigma}^2$ throughout later fits of the more flexible penalized models—See Supp. Note "*Jointly estimating the residual variance and edge weights*". Additionally we find that initializing the edge weights to $\widehat{w}_0$ to be a useful and intuitive strategy to set the initial values for the entries of $\boldsymbol{w}$ to the correct scale.

3.5.5 Data description and quality control

We analyzed a population genetic dataset of North American gray wolves previously published in Schweizer et al. (2016). For this, we downloaded plink formatted files and spatial coordinates from `https://doi.org/10.5061/dryad.c9b25`. We removed all SNPs with minor allele frequency less than 5% and with missingness greater then 10% resulting in a final set of 111 individuals and $17,729$ SNPs.

3.5.6 Population structure analyses

We fit the Pritchard, Donnelly, and Stephens model (PSD) and ran principal components analysis on the genotype matrix of North American gray wolves (Price et al., 2006; Pritchard et al., 2000). For the PSD model we used the ADMIXTURE software on the un-normalized genotypes, running 5 replicates per choice of K, from $K = 2$ to $K = 8$ (Alexander et al., 2009). For each K we choose the one that achieved the highest likelihood to visualize. For PCA, we centered and scaled the genotype matrix and then ran `sklearn` implementation of PCA, truncated to compute 50 eigenvectors.

3.5.7 Grid construction

To create a dense triangular lattice around the sample locations, we first define an outer boundary polygon. As a default, we construct the lattice by creating a convex hull around the sample points and manually trimming the polygon to adhere to the geography of the study organism and balancing the sample point range with the extent of local geography using the following website `https://www.keene.edu/campus/maps/tool/`. We often do not exclude internal "holes" in the habitat (e.g. water features for terrestrial animals), and let the model instead fit effective migration rates for those features to the extent they lead to elevated differentiation. We also emphasize the importance of defining the lattice for FEEMS as well as EEMS and suggest this should be carefully curated with prior biological knowledge about the system.

To ensure edges cover an equal area over the entire region we downloaded and intersected a uniform grid defined on the spherical shape of earth (Sahr et al., 2003). These defined grids are pre-computed at a number of different resolutions, allowing a user to test FEEMS at different grid densities which is an important feature to explore.

3.6 Code Availability

The code to reproduce the results of this paper and more can be found in `https://github.com/jhmarcus/feems-paper`. A `python` package implementing the method can be found in `https://github.com/jhmarcus/feems` with documentation found in `http://jhmarcus.com/feems/`.

3.7 Data Availability

We included a processed version of the dataset used in this manuscript in the `feems` package found here: `https://github.com/jhmarcus/feems`. An example tutorial on how to access the data the can be found here: `http://jhmarcus.com/feems/notebooks/getting-started.html`.

3.8 Acknowledgements

We thank Rena Schweizer for helping us download and process the gray wolf dataset used in the paper, Ben Peter for providing feedback and code for helping to construct the discrete global grids and preparing the human genetic dataset, and Hussein Al-Asadi, Peter Carbonetto, Dan Rice for helpful conversations about the optimization and modeling approach. We also acknowledge helpful feedback from Arjun Biddanda, Anna Di Rienzo, Matthew Stephens, the Stephens Lab, the Novembre Lab, and the University of Chicago 4th floor computational biology group.

This study was supported in part by the National Science Foundation via fellowship DGE-1746045 and the National Institute of General Medical Sciences via training grant T32GM007197 to J.H.M and R01GM132383 to J.N. W.H. was partially supported by the NSF via the TRIPODS program and by the Berkeley Institute for Data Science. R.F.B. was supported by the National Science Foundation via grant DMS–1654076, and by the Office of Naval Research via grant N00014-20-1-2337.

3.9 Author Contributions

J.H.M and J.N. conceived of the project. J.H.M. and W.H. developed the statistical methodology with guidance from R.F.B. (lead) and J.N. (supporting). J.H.M, W.H. carried out method testing and application with guidance from J.N. (lead) and R.F.B. (supporting). J.H.M. and W.H. developed the software. J.H.M. and W.H. wrote the paper with edits from R.F.B. and J.N.

3.10 Supplementary Information

3.10.1 Mathematical notation

We denote matrices using bold capital letters $\boldsymbol{A}$. Bold lowercase letters are vectors $\boldsymbol{a}$, and non-bold lowercase letters are scalars a. We denote by $\boldsymbol{A}^{-1}$ and $\boldsymbol{A}^{\dagger}$ the inverse and (Moore-Penrose) pseudo-inverse of $\boldsymbol{A}$ respectively. We use $\boldsymbol{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to express that the random vector $\boldsymbol{y}$ is modeled as a p-dimensional multivariate Gaussian distribution with fixed parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and use the conditional notation $\boldsymbol{y}|\boldsymbol{\mu} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if $\boldsymbol{\mu}$ is random.

A graph is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes a set of nodes or vertices and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes a set of edges. Throughout we assume the graph $\mathcal{G}$ is undirected, weighted, and contains no self loops, i.e. $(i,j) \in \mathcal{E} \iff (j,i) \in \mathcal{E}$ and $= (i,i) \notin \mathcal{E}$ and each edge $(i,j) \in \mathcal{E}$ is given a weight $w_{ij} = w_{ji} > 0$. We write $\boldsymbol{W}$ to indicate the symmetric weighted adjacency matrix, i.e.

$$\boldsymbol{W}_{ij} = \begin{cases} w_{ij}, & \text{if } (i,j) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

$\boldsymbol{w} \in \mathbb{R}^m$ is a vectorized form of the non-zero lower-triangular entries of $\boldsymbol{W}$ where $m = |\mathcal{E}|/2$ is the number of non-zero lower triangular elements. We denote by $\boldsymbol{L} = \operatorname{diag}(\boldsymbol{W}\boldsymbol{1}) - \boldsymbol{W}$ the graph Laplacian.

3.10.2 Gradient computation

In practice, we make a change of variable from $\boldsymbol{w} \in \mathbb{R}_+^m$ to $\boldsymbol{z} = \log(\boldsymbol{w}) \in \mathbb{R}^m$ and the algorithm is applied to the transformed objective function:

$$\ell(\exp(\boldsymbol{z}), \sigma^2; \boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) + \phi_{\lambda,\alpha}(\exp(\boldsymbol{z})) = \tilde{\ell}(\boldsymbol{z}, \sigma^2; \boldsymbol{C}\hat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) + \tilde{\phi}_{\lambda,\alpha}(\boldsymbol{z}).$$

After the change of variable, the objective value remains the same whereas it follows from the chain rule that $\nabla(\widetilde{\ell}(\boldsymbol{z})+\widetilde{\phi}_{\lambda,\alpha}(\boldsymbol{z})) = \nabla(\ell(\boldsymbol{w})+\phi_{\lambda,\alpha}(\boldsymbol{w}))\odot\boldsymbol{w}$ where $\odot$ indicates the Hadamard product or elementwise product—for notational convenience, we drop the dependency of ℓ on the quantities σ^2 and $\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top$. Furthermore, the computation of $\nabla\phi_{\lambda,\alpha}(\boldsymbol{w})$ is relatively straightforward, so in the rest of this section, we discuss only the computation of the gradient of the negative log-likelihood function with respect to $\boldsymbol{w}$, i.e. $\nabla\ell(\boldsymbol{w})$.

Recall, by definition, the graph Laplacian $\boldsymbol{L}$ implicitly depends on the variable $\boldsymbol{w}$ through $\boldsymbol{L} = \operatorname{diag}(\boldsymbol{W}\boldsymbol{1})-\boldsymbol{W}$. Throughout we assume the first o rows and columns of $\boldsymbol{L}$ correspond to the observed nodes. With this assumption, our node assignment matrix has block structure $\boldsymbol{A} = [\mathbf{I}_{o\times o} \mid \mathbf{0}_{o\times(d-o)}]$. To simplify some of the equations appearing later, we introduce the notation: we define

$$\boldsymbol{L}_{\text{full}} := \boldsymbol{L} + \frac{\boldsymbol{1}\boldsymbol{1}^\top}{d}, \;\; \boldsymbol{\Sigma} := \boldsymbol{A}\boldsymbol{L}_{\text{full}}^{-1}\boldsymbol{A}^\top + \sigma^2\operatorname{diag}(\boldsymbol{n}^{-1}), \tag{3.10}$$

and

$$\boldsymbol{M} := \boldsymbol{C}^\top\left((\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C})^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C})(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C})^{-1} - (\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C})^{-1}\right)\boldsymbol{C}.$$

Applying the chain rule and matrix derivatives, we can calculate:

$$\nabla\ell(\boldsymbol{w}) = \frac{\partial\ell(\boldsymbol{w})}{\partial\operatorname{vec}(\boldsymbol{L})}\cdot\frac{\partial\operatorname{vec}(\boldsymbol{L})}{\partial\boldsymbol{w}^\top},$$

where vec is the vectorization operator and $\partial\ell/\partial\operatorname{vec}(\boldsymbol{L})$ and $\partial\operatorname{vec}(\boldsymbol{L})/\partial\boldsymbol{w}^\top$ are $1\times d^2$ vector and $d^2\times d$ matrix, respectively, given by

$$\frac{\partial\ell(\boldsymbol{w})}{\partial\operatorname{vec}(\boldsymbol{L})} = p\cdot\operatorname{vec}\left(\boldsymbol{L}_{\text{full}}^{-1}\boldsymbol{A}^\top\boldsymbol{M}\boldsymbol{A}\boldsymbol{L}_{\text{full}}^{-1,\top}\right), \;\; \frac{\partial\operatorname{vec}(\boldsymbol{L})}{\partial\boldsymbol{w}^\top} = \boldsymbol{S}-\boldsymbol{T}. \tag{3.11}$$

Here $\boldsymbol{S}$ and $\boldsymbol{T}$ are linear operators that satisfy $\boldsymbol{S}\boldsymbol{w} = \operatorname{diag}(\boldsymbol{W}\boldsymbol{1})$ and $\boldsymbol{T}\boldsymbol{w} = \boldsymbol{W}$. Note $\boldsymbol{S}$ and

$\boldsymbol{T}$ both have $\mathcal{O}(d)$ many nonzero entries, so we can perform sparse matrix multiplication to efficiently compute the matrix-vector multiplication $\partial\ell/\partial\text{vec}(\boldsymbol{L}) \cdot (\boldsymbol{S} - \boldsymbol{T})$. On the other hand, the computation of $\partial\ell/\partial\text{vec}(\boldsymbol{L})$ is more challenging as it requires inverting the full $d \times d$ matrix $\boldsymbol{L}_{\text{full}}$. Next we develop a procedure that efficiently computes $\partial\ell/\partial\text{vec}(\boldsymbol{L})$. We proceed by dividing the task into multiple steps.

1. Computing $\boldsymbol{\Sigma}^{-1}$ Recalling the block structure $\boldsymbol{A} = [\mathbf{I}_{o\times o} \mid \mathbf{0}_{o\times(d-o)}]$ of the node assignment matrix, we can write $\boldsymbol{\Sigma}$ as:

$$\boldsymbol{\Sigma} = \left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o} + \sigma^2\text{diag}(\boldsymbol{n}^{-1}),$$

where $\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}$ denotes the $o \times o$ upper-left block of $\boldsymbol{L}_{\text{full}}^{-1}$. Following Petkova et al. (2016), the inverse $\boldsymbol{\Sigma}^{-1}$ has the form

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{X} + \sigma^{-2}\text{diag}(\boldsymbol{n}), \tag{3.12}$$

for some matrix $\boldsymbol{X} \in \mathbb{R}^{o\times o}$. Equating $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \mathbf{I}$, it follows that

$$\begin{aligned}
&\left[\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o} + \sigma^2\text{diag}(\boldsymbol{n}^{-1})\right]\left(\boldsymbol{X} + \sigma^{-2}\text{diag}(\boldsymbol{n})\right) = \mathbf{I} \\
\iff &\left[\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o} + \sigma^2\text{diag}(\boldsymbol{n}^{-1})\right]\boldsymbol{X} = -\sigma^{-2}\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}\text{diag}(\boldsymbol{n}).
\end{aligned} \tag{3.13}$$

Therefore, $\boldsymbol{\Sigma}^{-1}$ can be obtained by solving the $o \times o$ linear system (3.13) and plugging the solution into (3.12). The challenge here is to compute $\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}$ without matrix inversion of the full-dimensional $\boldsymbol{L}_{\text{full}}$.

2. Computing $\left(\boldsymbol{L}_{\textbf{full}}^{-1}\right)_{o\times o}$ Let $\boldsymbol{L}_{\text{full},o\times o}$ be the $o \times o$ block matrix corresponding to the observed nodes of $\boldsymbol{L}_{\text{full}}$, and similarly let $\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}$ and $\boldsymbol{L}_{\text{full},o\times(d-o)} = \boldsymbol{L}_{\text{full},(d-o)\times o}^{\top}$ be the corresponding block matrices of $\boldsymbol{L}_{\text{full}}$ respectively. The inverse of $\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}$ is then given by the Schur complement of $\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}$ in $\boldsymbol{L}$:

$$\left[\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}\right]^{-1} = \boldsymbol{L}_{\text{full},o\times o} - \boldsymbol{L}_{\text{full},o\times(d-o)}\left(\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}\right)^{-1}\boldsymbol{L}_{\text{full},(d-o)\times o}. \tag{3.14}$$

See also Hanks and Hooten (2013); Petkova et al. (2016). Since every term in (3.14) has sparse + rank-1 structure, the matrix multiplications can be performed fast. In addition, for the term $\left(\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}\right)^{-1}$, we can use the Sherman-Morrison formula so that the inverse is given explicitly by

$$\begin{aligned}\left(\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}\right)^{-1} &= \left(\boldsymbol{L}_{(d-o)\times(d-o)} + \frac{\boldsymbol{1}\boldsymbol{1}^\top}{d}\right)^{-1} \\ &= \boldsymbol{L}_{(d-o)\times(d-o)}^{-1} - \frac{1}{d + \boldsymbol{1}^\top\boldsymbol{L}_{(d-o)\times(d-o)}^{-1}\boldsymbol{1}}\boldsymbol{L}_{(d-o)\times(d-o)}^{-1}\boldsymbol{1}\boldsymbol{1}^\top\boldsymbol{L}_{(d-o)\times(d-o)}^{-1}.\end{aligned}$$

Hence, in order to compute $\left(\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}\right)^{-1}\boldsymbol{L}_{\text{full},(d-o)\times o}$, we need to solve two systems of linear equations:

$$\boldsymbol{L}_{(d-o)\times(d-o)}\boldsymbol{U} = \boldsymbol{L}_{\text{full},(d-o)\times o} \text{ and } \boldsymbol{L}_{(d-o)\times(d-o)}\boldsymbol{u} = \boldsymbol{1}.$$

Note that the matrix $\boldsymbol{L}_{(d-o)\times(d-o)}$ is sparse, so both systems can be solved efficiently by performing sparse Cholesky factorization on $\boldsymbol{L}_{(d-o)\times(d-o)}$ (Hanks and Hooten, 2013). Alternatively, one can implement fast Laplacian solvers (Vishnoi et al., 2013) that solve the Laplacian system in time nearly linear in the dimension $\mathcal{O}(d)$. After we obtain $\left[\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}\right]^{-1}$ via sparse + rank-1 matrix multiplication and sparse Cholesky factorization, we can invert the $o \times o$ matrix to get $\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}$.

3. Computing $\left(\boldsymbol{L}_{\textbf{full}}^{-1}\right)_{d\times o}$ Write

$$\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{d\times o} = \begin{bmatrix} \left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o} \\ \left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{(d-o)\times o} \end{bmatrix}.$$

Using the inversion of the matrix in a block form, the $(d-o)\times o$ block component is given by

$$\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{(d-o)\times o} = -\underbrace{\left(\boldsymbol{L}_{\text{full},(d-o)\times(d-o)}\right)^{-1} \boldsymbol{L}_{\text{full},(d-o)\times o}}_{\text{(A)}} \underbrace{\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{o\times o}}_{\text{(B)}}. \tag{3.15}$$

Since each of the two terms (A) and (B) has been already computed in the previous step, there is no need to recompute them. In total, it requires a $(d-o)\times o$ matrix and $o\times o$ matrix multiplication.

4. Computing the full gradient Going back to the expression of $\nabla\ell(\boldsymbol{w})$ in (3.11), and noting the block structure of the assignment matrix $\boldsymbol{A}$, we have:

$$\frac{\partial\ell(\boldsymbol{w})}{\partial\text{vec}(\boldsymbol{L})} = p\cdot\text{vec}\left(\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{d\times o}\boldsymbol{M}\left(\boldsymbol{L}_{\text{full}}^{-1}\right)_{d\times o}^{\top}\right).$$

Let $\Pi_{\mathbf{1}} = \mathbf{1}\left(\mathbf{1}^{\top}\boldsymbol{\Sigma}^{-1}\mathbf{1}\right)^{-1}\mathbf{1}^{\top}\boldsymbol{\Sigma}^{-1}$ be projection to the space of constant vectors with respect to the inner product $\langle\boldsymbol{x},\boldsymbol{y}\rangle = \boldsymbol{x}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$. Using the identity $\mathbf{I}-\boldsymbol{\Pi}_{\mathbf{1}} = \boldsymbol{\Sigma}\boldsymbol{C}^{\top}(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{\top})^{-1}\boldsymbol{C}$ (McCullagh, 2009), then we can write $\boldsymbol{M}$ in terms of $\boldsymbol{\Pi}_{\mathbf{1}}$:

$$\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}\left(\mathbf{I}-\boldsymbol{\Pi}_{\mathbf{1}}\right)\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{I}-\boldsymbol{\Pi}_{\mathbf{1}}\right) - \boldsymbol{\Sigma}^{-1}\left(\mathbf{I}-\boldsymbol{\Pi}_{\mathbf{1}}\right). \tag{3.16}$$

Since $\Pi_{\mathbf{1}}$ is a rank-1 matrix, this expression of $\boldsymbol{M}$ allows easier computation. Finally we can put together (3.12), (3.13), (3.15), and (3.16), to compute the gradient of the negative log-likelihood function with respect to the graph Laplacian.

3.10.3 Objective computation

The graph Laplacian $\boldsymbol{L}$ is orthogonal to the one vector $\mathbf{1}$, so using the notation introduced in (3.10), we can express our objective function as

$$\ell(\boldsymbol{w})+\phi_{\lambda,\alpha}(\boldsymbol{w}) = p\cdot\mathrm{tr}\left(\left(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top\right)^{-1}\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top\right) - p\cdot\log\det\left(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}\right)^{-1} + \frac{\lambda}{2}\|\boldsymbol{\Delta}(\boldsymbol{w}+\alpha\log(\boldsymbol{w})\|_2^2.$$

With the identity $\mathbf{I}-\boldsymbol{\Pi}_{\mathbf{1}} = \boldsymbol{\Sigma}\boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}$, the trace term is:

$$\mathrm{tr}\left(\left(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top\right)^{-1}\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top\right) = \mathrm{tr}\left(\boldsymbol{C}^\top\left(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top\right)^{-1}\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\right) = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}(\mathbf{I}-\boldsymbol{\Pi}_{\mathbf{1}})\widehat{\boldsymbol{\Sigma}}\right).$$

The matrix inside the trace has been constructed in the gradient computation, see equation (3.16). In terms of the determinant, we use the same approach considered in Petkova et al. (2016)—in particular, concatenating $\boldsymbol{C}^\top$ and $\mathbf{1}$, the matrix $[\boldsymbol{C}^\top \mid \mathbf{1}]$ is orthogonal, so it can be shown that

$$\det(\boldsymbol{\Sigma}) = \frac{\det(\mathbf{1}^\top\mathbf{1})\det(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top)}{\det(\boldsymbol{C}\boldsymbol{C}^\top)\det(\mathbf{1}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1})}.$$

Rearranging terms and using the fact $\det(\boldsymbol{U}^{-1}) = \det(\boldsymbol{U})^{-1}$ for any matrix $\boldsymbol{U}$, we obtain:

$$\det(\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^\top)^{-1} = \frac{\det(\mathbf{1}^\top\mathbf{1})\det(\boldsymbol{\Sigma}^{-1})}{\det(\boldsymbol{C}\boldsymbol{C}^\top)\det(\mathbf{1}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1})} = \frac{o}{\mathbf{1}^\top\boldsymbol{\Sigma}^{-1}\mathbf{1}}\det(\boldsymbol{\Sigma}^{-1}).$$

We have computed $\boldsymbol{\Sigma}^{-1}$ in equation (3.12), so each of the terms above can be computed without any additional matrix multiplications. Finally, the signed graph incidence matrix $\boldsymbol{\Delta}$ defined on the edges of the graph is, by construction, highly sparse with $\mathcal{O}(d)$ many nonzero entries. Hence we implement sparse matrix multiplication to evaluate the penalty function $\phi_{\lambda,\alpha}(\boldsymbol{w})$ while avoiding the full-dimensional matrix-vector product.

3.10.4 Estimating the edge weights under the exact likelihood model

Recall that, when describing our data model, we employed the approximation $\frac{1}{2}f_j(k)(1-f_j(k)) \approx \sigma^2\mu_j(1-\mu_j)$ for all SNPs j and nodes k (see equation (3.4)) and estimated the residual variance σ^2 under the homogeneous isolation by distance model. Here we examine whether this approximation results in a significant difference with respect to the estimation quality of the edge weights of the graph.

Without approximation, we can calculate the exact analytical form for the marginal likelihood of the estimated frequency as follows (after removing the SNP means):

$$\boldsymbol{C}\widehat{\boldsymbol{f}}_j \sim \sqrt{\mu_j(1-\mu_j)}\cdot\mathcal{N}_{o-1}\left(\boldsymbol{0}, \boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^{\dagger}\boldsymbol{A}^{\top}\boldsymbol{C}^{\top}+\boldsymbol{C}\mathrm{diag}(\boldsymbol{n}^{-1})\boldsymbol{A}\mathrm{diag}\left(\left\{\frac{1-L^{\dagger}_{kk}}{2}\right\}_{k=1}^{d}\right)\boldsymbol{A}^{\top}\boldsymbol{C}^{\top}\right), \tag{3.17}$$

where $\{a_k\}_{k=1}^{d}$ represents the vector $\boldsymbol{a} = (a_1, \ldots, a_d)$. We then consider estimating the edge weights with the likelihood based on (3.17) and without relying on approximating the residual variance. In particular, comparing to the model (3.5), this formulation does not introduce the unknown residual variance parameter σ^2 but rather it is given implicitly by $(1-L^{\dagger}_{kk})/2$. This means that the model (3.17) is well-defined only when $L^{\dagger}_{kk} \leqslant 1$ for all nodes k, hence leading to the following constrained optimization problem:

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{l}\leqslant\boldsymbol{w}\leqslant\boldsymbol{u}}{\arg\min}\left\{\ell_{\mathsf{exact}}(\boldsymbol{w}; \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top}) + \phi_{\lambda,\alpha}(\boldsymbol{w}) : L^{\dagger}_{kk} \leqslant 1 \text{ for all } k \in \mathcal{V}\right\}, \tag{3.18}$$

where ℓ_{exact} is the negative log-likelihood function implied by the model (3.17) and $\phi_{\lambda,\alpha}$ is our smooth penalty function. The main difficulty of solving (3.18) is that enforcing the constraint $L^{\dagger}_{kk} \leqslant 1$ for all nodes $k \in \mathcal{V}$, requires full computation of the pseudo-inverse of a $d \times d$ matrix $\boldsymbol{L}$ whereas in order to evaluate the likelihood, we only need to calculate $\boldsymbol{L}^{\dagger}$ on the observed nodes. To overcome this computational challenge, we may relax the constraint and consider the following form as a proxy for optimization (3.18):

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{l} \leqslant \boldsymbol{w} \leqslant \boldsymbol{u}}{\arg\min} \left\{ \ell_{\mathsf{exact}}(\boldsymbol{w}; \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) + \phi_{\lambda,\alpha}(\boldsymbol{w}) : L^\dagger_{kk} \leqslant 1 \text{ for all observed nodes } k \right\}. \qquad (3.19)$$

We can solve this problem efficiently using a gradient-based algorithm where the gradient of ℓ_{exact} with respect to $\boldsymbol{L}$ is given by

$$\frac{\partial \ell_{\mathsf{exact}}(\boldsymbol{w})}{\partial \mathrm{vec}(\boldsymbol{L})} = p \cdot \mathrm{vec}\left(\boldsymbol{L}^{-1}_{\mathrm{full}} \boldsymbol{A}^\top \boldsymbol{M} \boldsymbol{A} \boldsymbol{L}^{-1,\top}_{\mathrm{full}} \right) - p \cdot \mathrm{diag}(\boldsymbol{M})^\top \mathrm{diag}((2\boldsymbol{n})^{-1}) \boldsymbol{N},$$

where $\boldsymbol{M}$ is a $o \times o$ matrix defined in (3.16), while $\boldsymbol{N}$ is a $o \times d^2$ matrix whose rows correspond to the observed subsets of the rows of the $d^2 \times d^2$ matrix $\boldsymbol{L}^{-1}_{\mathrm{full}} \otimes \boldsymbol{L}^{-1}_{\mathrm{full}}$.

Overall, when we implement the penalized restricted maximum likelihood procedure in (3.19), we find that it does not make much of a difference and output qualitatively comparable results to FEEMS—for example, Supp. Fig. 3.16 shows one such fit with a setting of $\lambda = 10^{-3}$ and $\alpha = 50$. Unfortunately, this approach has a drawback that after the algorithm reaches the solution, the term $1 - L^\dagger_{kk}$ is not guaranteed to be positive for the unobserved nodes, since, due to the computational efficiency, the constraints $L^\dagger_{kk} \leqslant 1$ are only placed on the observed nodes. This, in principle, results in an ill-defined model if we would like interpretable results at unobserved as well as observed nodes, and therefore we replace the calculation (3.17) with the approximation (3.5) to avoid this issue. In addition, by decoupling the residual variance parameter σ^2 from the graph-related weighted edges $\boldsymbol{w}$, the model (3.6) has more resemblance to spatial coalescent model used in EEMS (Petkova et al., 2016).

3.10.5 Jointly estimating the residual variance and edge weights

One simple strategy we have used throughout the paper was to fit σ^2 first under a model of homogeneous isolation by distance and prefix the estimated residual variance to the resulting $\widehat{\sigma}^2$ for later fits of the effective migration rates. Alternatively, one might come up with a strategy to estimate the unknown residual variance jointly with the edge weights, instead of prefixing it from the estimation of the null model—the hope here is to simultaneously correct the model misspecification and allow for improving model fit to the data.

As it turns out, given such a small fraction of sampled spatial locations in the data, the strategy of jointly optimizing the marginal likelihood with respect to both variables has the tendency to overfit to the data unless it is properly regularized. Specifically, we can consider the model that generalizes (3.6), namely

$$p \cdot \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top \sim \mathcal{W}_{o-1}\Bigg(\boldsymbol{C}\boldsymbol{A}\boldsymbol{L}^\dagger\boldsymbol{A}^\top\boldsymbol{C}^\top + \boldsymbol{C}\text{diag}(\boldsymbol{n}^{-1})\boldsymbol{A}\text{diag}\big(\boldsymbol{\sigma}^2\big)\boldsymbol{A}^\top\boldsymbol{C}^\top, p\Bigg),$$

where $\boldsymbol{\sigma}^2$ is a $d \times 1$ vector of node specific residual variances, i.e. each deme has its own residual parameter σ_k for all nodes k. If the node specific parameters σ_k's are assumed to be same across all nodes, this reduces to the model (3.6). Supp. Fig. 3.17 shows the results of different strategies of estimating the residual variances. As expected, when the model has a single residual variance σ^2, either prefixing it from the null model (Figure 4.5) or estimating it jointly with the edge weights (Supp. Fig. 3.17A) lead to similar and comparable outputs. The major difference is the high migration edge forming long path appearing in Supp. Fig. 3.17A to separate the reduced gene-flows in the middle, which tends to disappear as α increases. Whereas, if the residual variances are allowed to be node specific, the fitted σ_k^2's are highly variable and as a result the estimated graph misses some geographic features present in the data, such as reduced effective migration around St. Lawerence Island (Supp. Fig. 3.17B). Presumably this is attributed to overfitting, due to the absence of data in many unobserved demes. In EEMS, in order to estimate the genetic diversity parameters for every spatial position, which play a similar role as the residual variance in FEEMS, a Voronoi-tessellation prior is placed to encourage sharing of information across adjacent nodes and prevent over-fitting. While we can similarly estimate the node specific residual variances on every node of the graph with our penalty function ($\phi_{\lambda,\alpha}$ defined on the variable $\boldsymbol{\sigma}^2$), we do not find it substantially improves the extent to which the model suits the data. Thus, we take the approach of fitting the single residual variance σ^2 under the null model and prefixing it as a simple but effective strategy with apparent good empirical performance.

3.10.6 Edge versus node parameterization

One of the novel features of FEEMS is its ability to directly find the edge weights of the graph that best suit the data. This direct edge parameterization may increase the risk of model's overfitting, but also allows for more flexible estimation of migration histories. Furthermore, as seen in Figure 3.2 and Supp. Fig. 3.6, it has potential to recover anisotropic migration processes. This is in contrast to EEMS wherein every spatial node is assigned an effective migration parameter m_k and a migration rate on each edge joining nodes k and k' is given by the average effective migration $w_{kk'} = (m_k + m_{k'})/2$. Not surprisingly, parameterization via node-specific parameters induces implicit regularization by substantially constraining the feasible set of graph's edge weights. In some cases, this has the desirable property of imposing an additional degree of similarity among edge weights, but it often restricts the model's capacity to capture a richer set of structure present in the data, (e.g. Petkova et al., 2016, supplementary figure 2). To be concrete, Supp. Fig. 3.19 displays two different fits of FEEMS based on edge parameterization (Supp. Fig. 3.19A) and node parameterization (Supp. Fig. 3.19B), run on a previously published dataset of human genetic variation from Africa (see Peter et al. (2018) for details on the description of the dataset). Running FEEMS with a node-based parameterization is straightforward in our framework—all we have to do is to reparameterize the edge weights by the average effective migration and solve the corresponding optimization problem (3.9) with respect to $\boldsymbol{m}$. It is evident from the results that FEEMS with edge parameterization exhibits subtle correlations that exist between the annotated demes in the figure whereas node parameterization fails to recover them. We also compare the model fit of FEEMS to the observed genetic covariance (Supp. Fig. 3.20) and find that edge-based parameterization provides a better fit to the African dataset. Supp. Fig. 3.21 further demonstrates that in the coalescent simulations with anisotropic migration, the node parameterization is unable to recover the ground truth of the underlying migration rates even when the nodes are fully observed.

3.10.7 Smooth penalty with ℓ_1 norm

FEEMS's primary optimization objective (see equation (3.9)) is:

$$\underset{\boldsymbol{l} \leqslant \boldsymbol{w} \leqslant \boldsymbol{u}}{\text{Minimize}}\ \ell(\boldsymbol{w}, \sigma^2; \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) + \phi_{\lambda,\alpha}(\boldsymbol{w}),$$

where the spatial smoothness penalty is given by $\phi_{\lambda,\alpha}(\boldsymbol{w}) = \frac{\lambda}{2}\|\boldsymbol{\Delta}(\boldsymbol{w} + \alpha\log(\boldsymbol{w}))\|_2^2$. It is widely known that ℓ_1-based method leads to better local adaptive fitting and structural recovery than ℓ_2-based methods (Wang et al., 2016), but at the cost of handling non-smooth objective functions that are often computationally more challenging and demanding. In a spatial genetic dataset, one major challenge is to deal with the relatively sparse sampling design where there are many unobserved nodes on the graph. In this challenging statistical setting, our finding is that an ℓ_2-based method enables more accurate and reliable estimation of the geographic features.

Specifically, writing $\phi^{\ell_1}_{\lambda,\alpha}(\boldsymbol{w}) = \lambda\|\boldsymbol{\Delta}(\boldsymbol{w} + \alpha\log(\boldsymbol{w}))\|_1$, we considered the alternate following composite objective function:

$$\ell(\boldsymbol{w}, \sigma^2; \boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^\top) + \phi^{\ell_1}_{\lambda,\alpha}(\boldsymbol{w}). \tag{3.20}$$

To solve (3.20), we apply linearized alternating direction method of multipliers (ADMM) (Boyd et al., 2011), a variant of the standard ADMM algorithm, that iteratively optimizes the augmented Lagrangian over the primal and dual variables. The derivation of the algorithm is a standard calculation so we omit the detailed description of the algorithm. As opposed to the common belief about the effectiveness of the ℓ_1 norm for structural recovery, the recovered graph of FEEMS using ℓ_1-based smooth penalty shows less accurate reconstruction of the migration patterns, particularly when the sampling design has many locations with missing data on the graph (Supp. Fig. 3.22A, Supp. Fig. 3.23H). It appears that the ℓ_1-based penalty function is not capable of accurately estimating edge weights at regions with little data, partially due to its local adaptation, in contrast to the ℓ_2-based method that considers regularization more globally. This suggests that in order to use the ℓ_1 penalty $\phi^{\ell_1}_{\lambda,\alpha}(\boldsymbol{w})$ in the presence of many missing nodes, one needs an additional regularization term that promotes global smoothness of the graph's edge weights, e.g., a combination of $\phi^{\ell_1}_{\lambda,\alpha}(\boldsymbol{w})$ and $\phi_{\lambda,\alpha}(\boldsymbol{w})$ (same spirit as elastic net (Zou and Hastie, 2005)), or $\phi^{\ell_1}_{\lambda,\alpha}(\boldsymbol{w})$ on top of node-based parameterization (Supp. Fig. 3.22B).

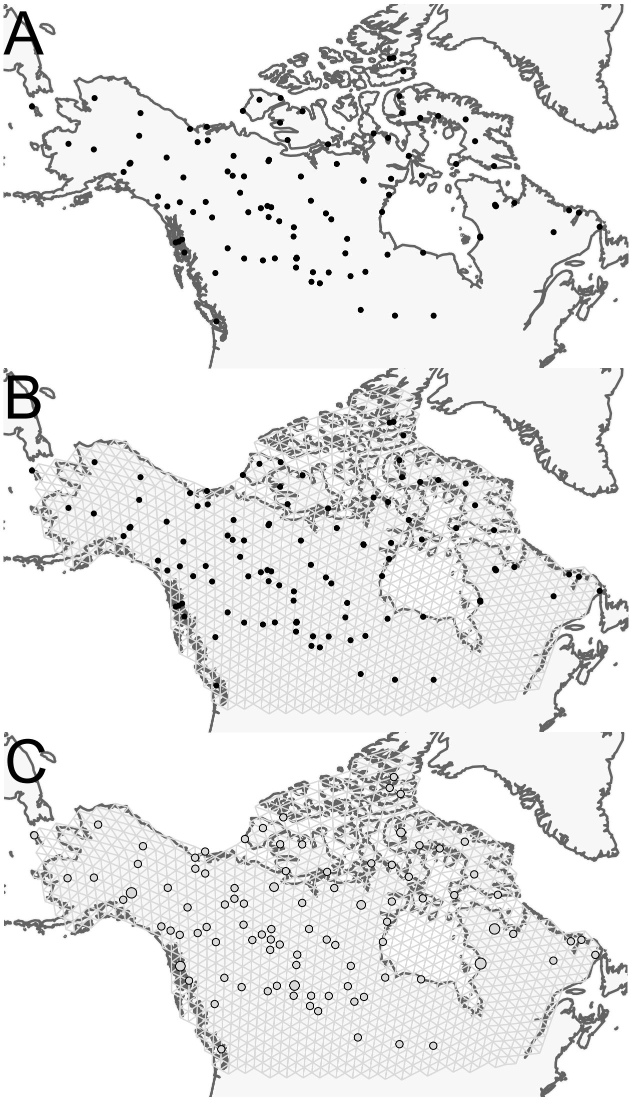


Figure 3.5: **Visualization of grid construction and node assignment:** (A) Map of sample coordinates (black points) from a dataset of gray wolves from North America. The input to FEEMS are latitude and longitude coordinates as well as genotype data for each sample. (B) Map of sample coordinates with an example dense spatial grid. The nodes of the grid represent sub-populations and the edges represent local gene-flow between adjacent sub-populations. (C) Individuals are assigned to nearby nodes (sub-populations) and summary statistics (e.g., allele frequencies) are computed for each observed location.

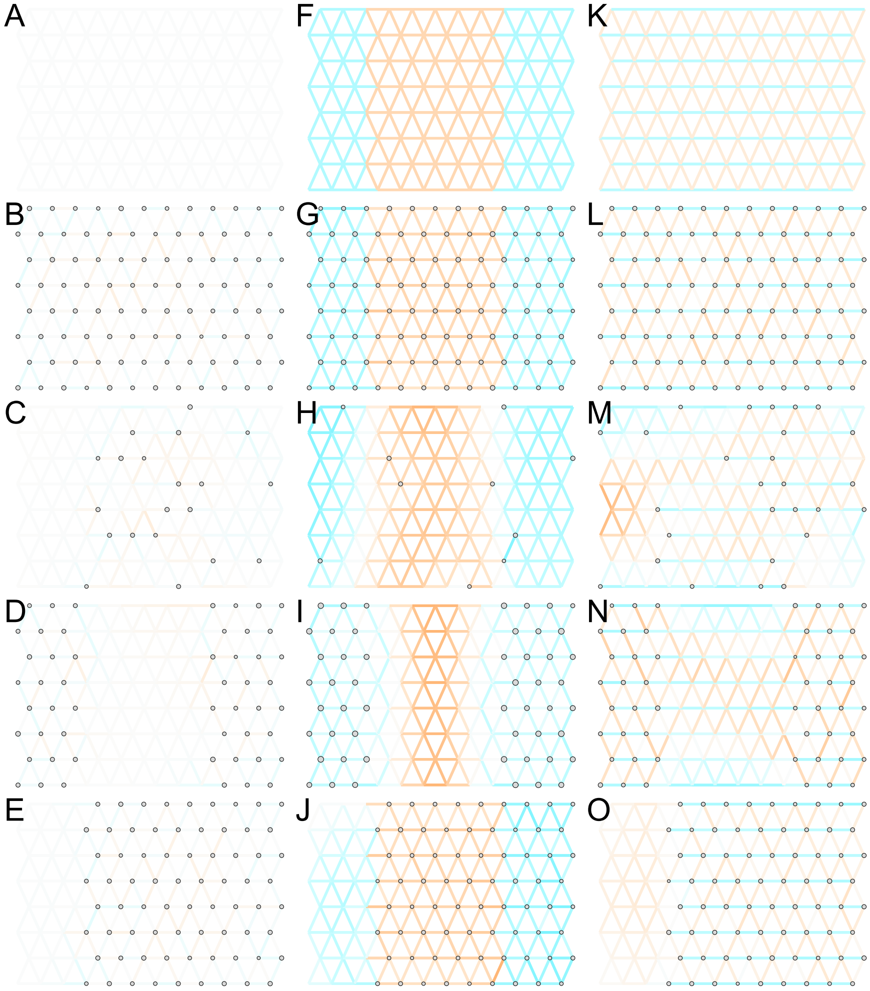


Figure 3.6: **Application of FEEMS to an extended set of coalescent simulations:** We display an extended set of coalescent simulations with multiple migration scenarios and sampling designs. The sample sizes across the grid are represented by the size of the grey dots at each node. The migration rates are obtained by solving FEEMS objective function (3.9) where the regularization parameters are specified at $\lambda = 10^{-2}, \alpha = 30$ (I), $\lambda = 10^{-4}, \alpha = 30$ (N), and $\lambda = 10^{-3}, \alpha = 30$ for the rest. (A, F, K) display the ground truth of the underlying migration rates. (B, G, L) Shows simulations where there is no missing data on the graph. (C, H, M) Shows simulations with sparse observations and nodes missing at random. (D, I, N) Shows simulations of biased sampling where there are no samples from the center of the simulated habitat. (E, J, O) Shows simulations of biased sampling where there are only samples on the right side of the habitat.

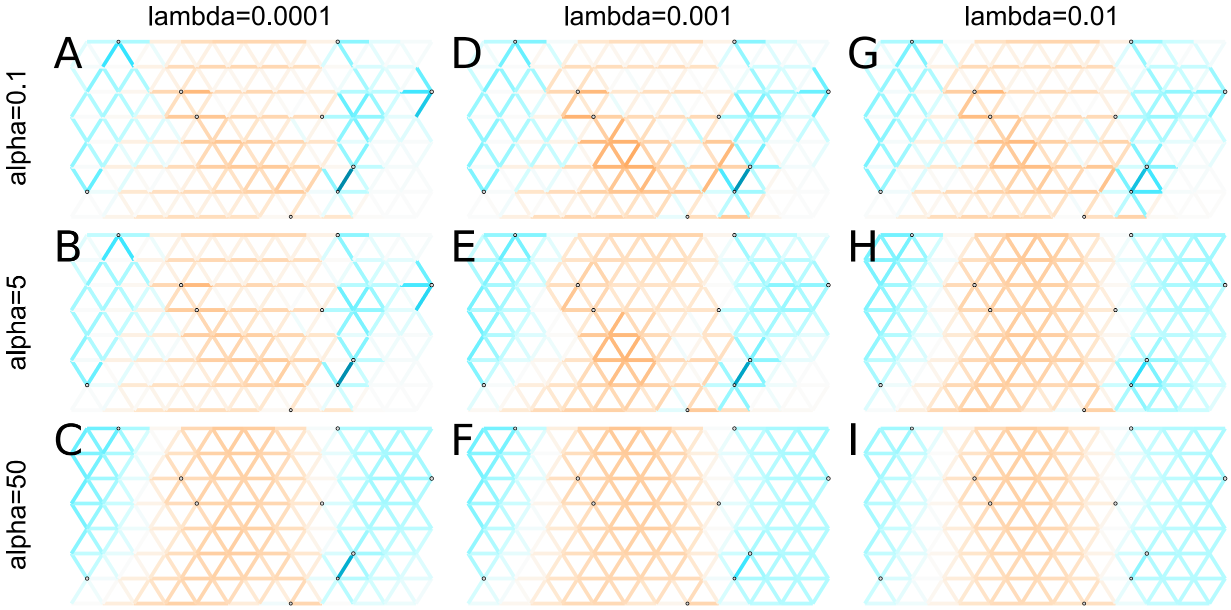


Figure 3.7: **Application of FEEMS to a heterogeneous migration scenario with a "missing at random" sampling design:** We run FEEMS on coalescent simulation with a non-homogeneous process while varying hyperparameters λ (rows) and α (columns). We randomly sample individuals for 20% of nodes. When λ grows, the fitted graph becomes overall smoother, whereas α effectively controls the degree of similarity among low migration rates.

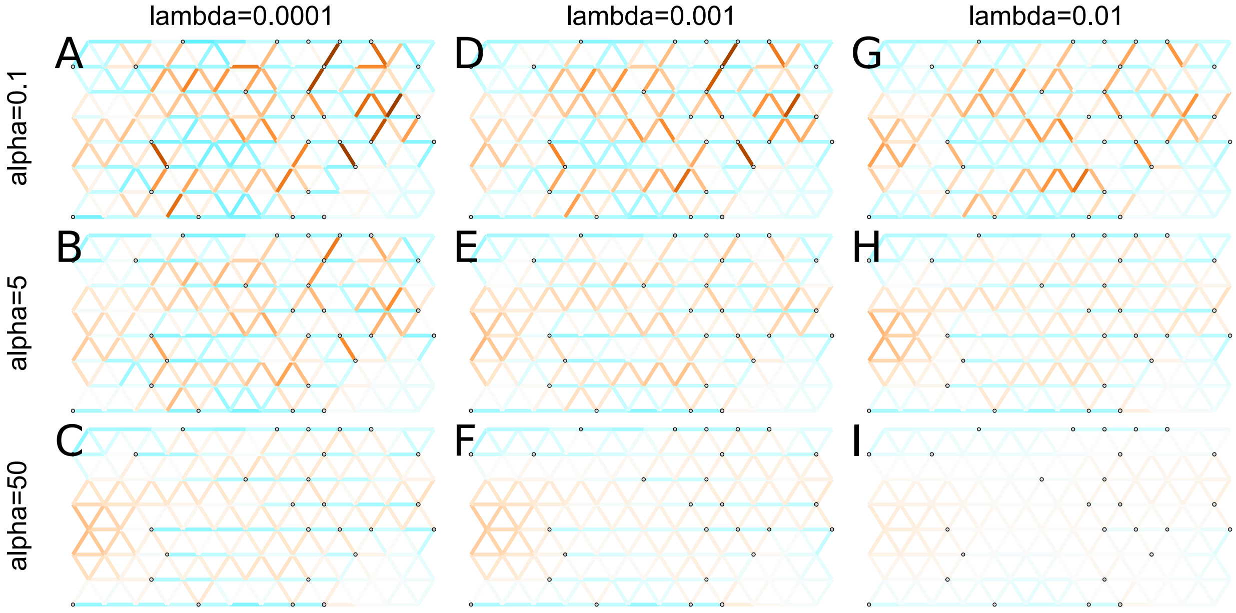


Figure 3.8: **Application of FEEMS to an anisotropic migration scenario with a “missing at random” sampling design:** We run FEEMS on coalescent simulation with an anisotropic process while varying hyperparameters λ (rows) and α (columns). We randomly sample individuals for 20% of nodes. When λ grows, the fitted graph becomes overall smoother, whereas α effectively controls the degree of similarity among low migration rates.

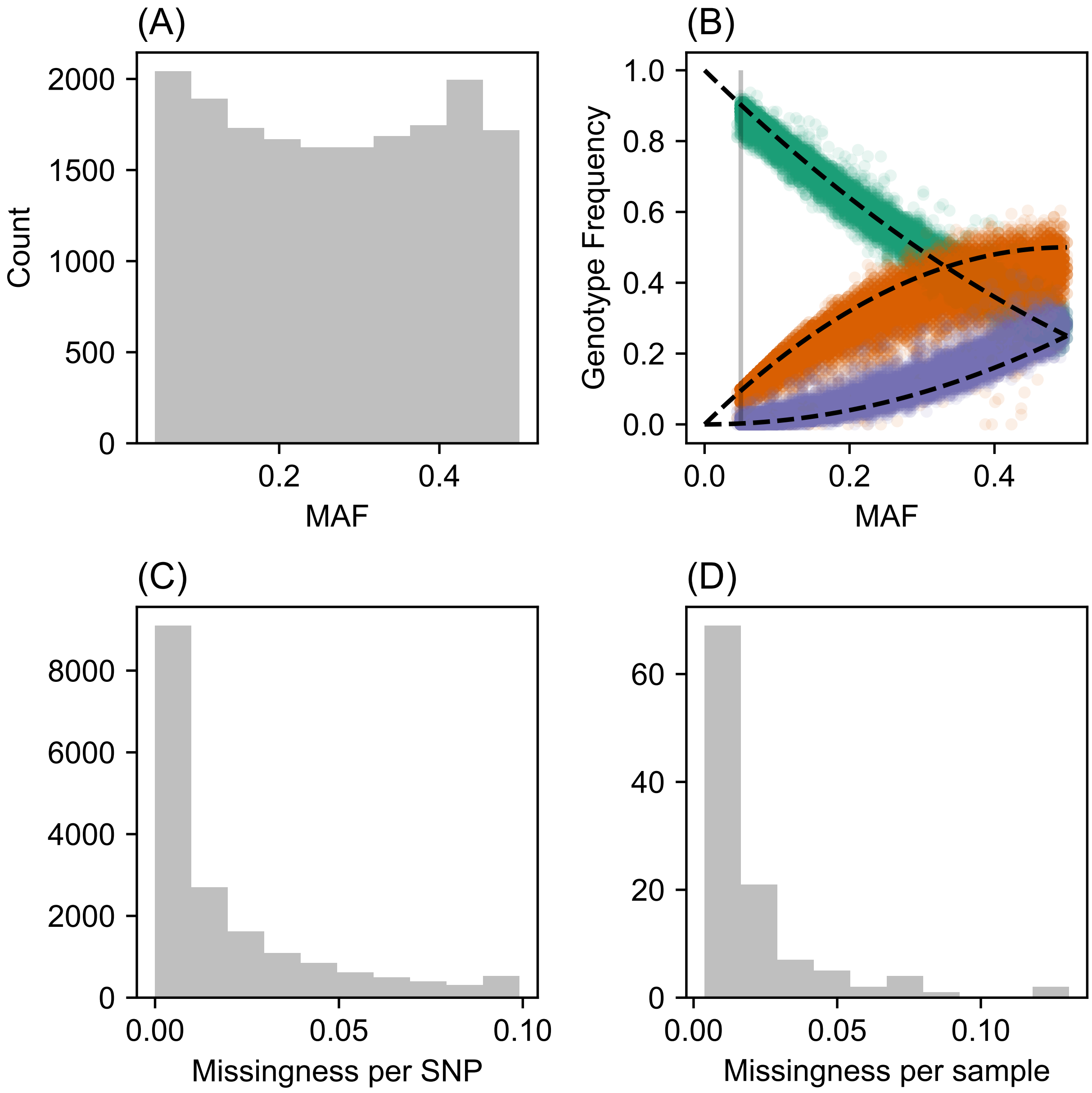


Figure 3.9: **SNP and individual quality control:** (A) Displays a visualization of the sample site frequency spectrum. Specifically, we display a histogram of minor allele frequencies across all SNPs. We see a relatively uniform histogram which reflects the ascertainment of common SNPs on the array that was designed to genotype gray wolf samples. (B) Visualization of allele frequencies plotted against genotype frequencies. Each point represents a different SNP and the colors represent the 3 possible genotype values. The black dashed lines display the expectation as predicted from a simple binomial sampling model i.e. Hardy-Weinberg equilibrium. (C) Displays a histogram of the missingness fraction per SNP. We observe the missingness tends to be relatively low for each SNP. (D) Displays a histogram of the missingness fraction per sample. Generally, the missingness tends to be low for each sample.

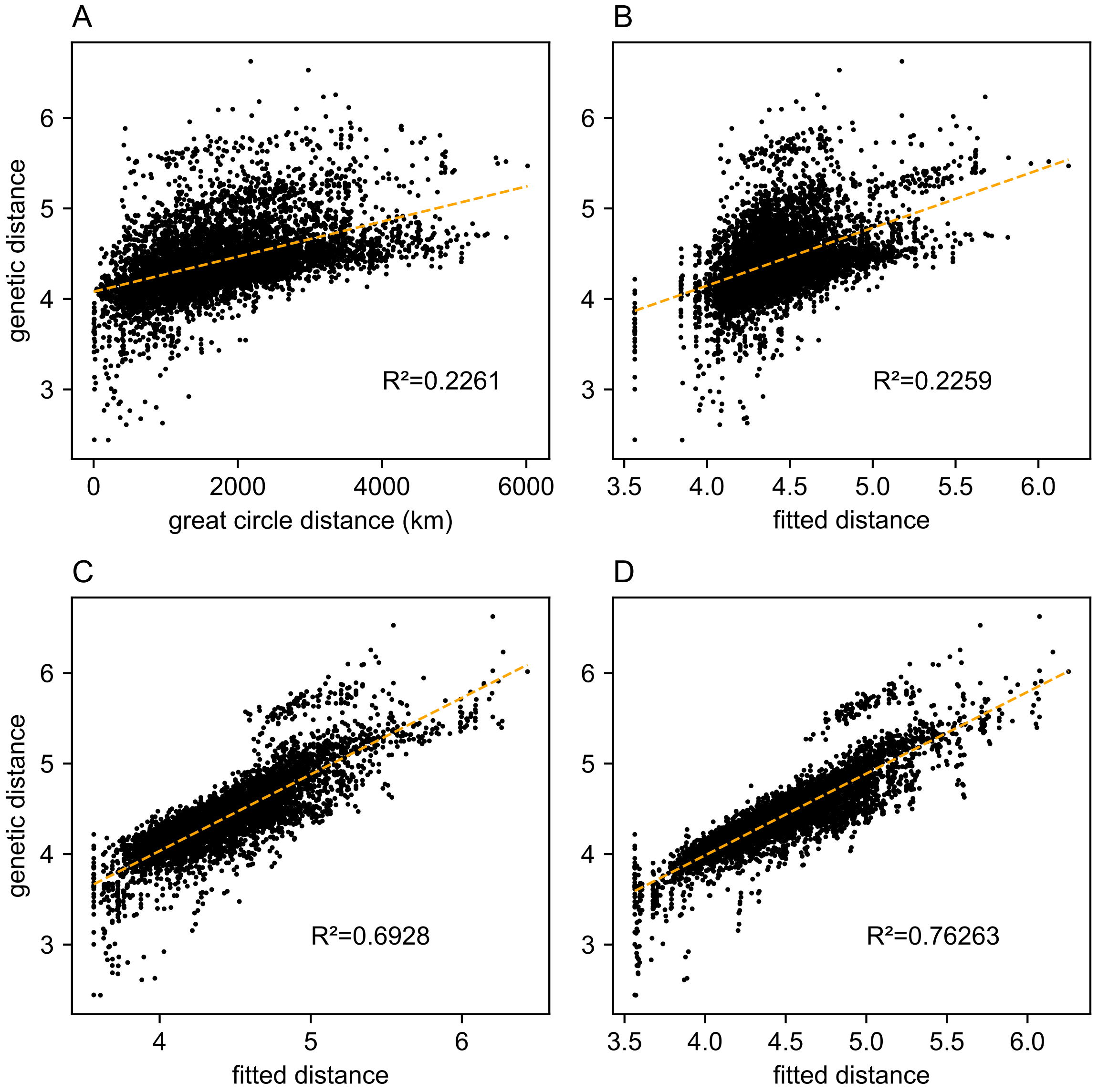


Figure 3.10: **Comparing predictions of observed genetic distances:** We display different predictions of observed genetic distances using geographic distance or the fitted genetic distance output by FEEMS. (A) The x-axis displays the geographic distance between two individuals, as measured by the great circle distance (haversine distance). The y-axis displays the squared Euclidean distance between two individuals averaged over all SNPs. (B-D) The x-axis displays the fitted genetic distance as predicted by the FEEMS model and y-axis displays the squared Euclidean distance between two individuals averaged over all SNPs. For (B-D) we display the fit of λ getting subsequently smaller $(10, 10^{-3}, 10^{-5})$ and as expected the fit becomes better because we tolerate more complex surfaces and we are not evaluating the fit on out-of-sample data.

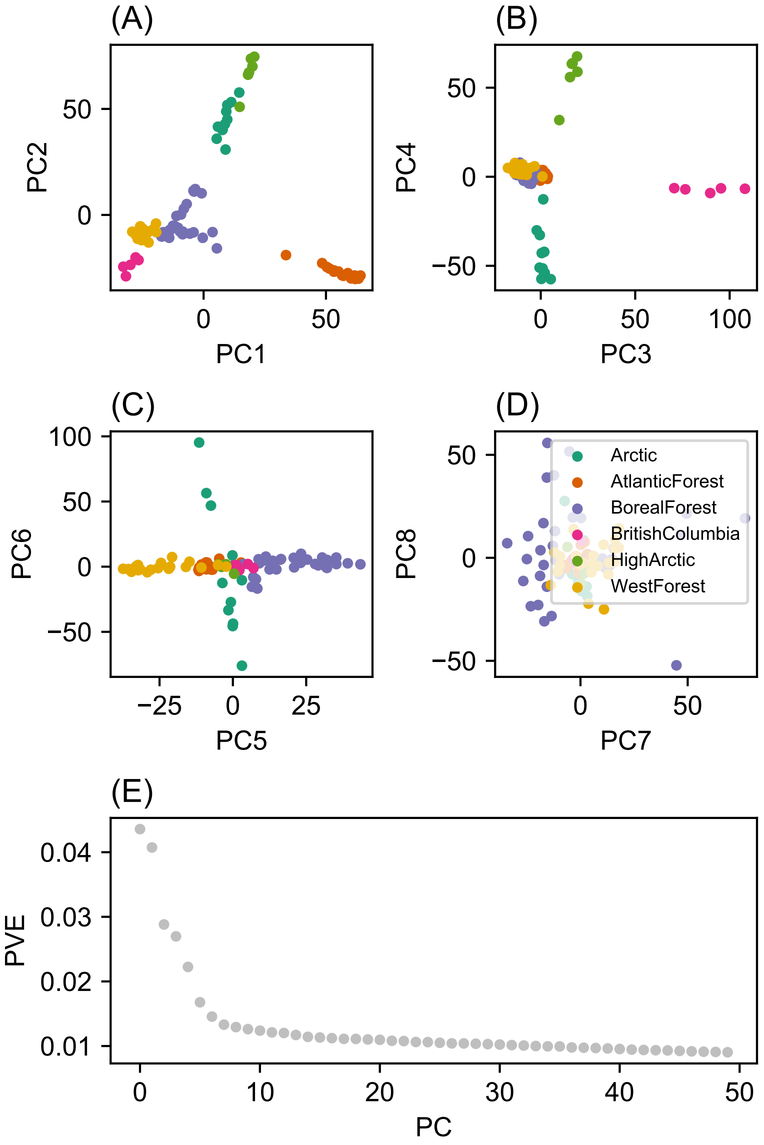


Figure 3.11: **Summary of top axes of genotypic variation:** We display a visual summary of Principal Components Analysis (PCA) applied to the normalized genotype matrix from the North American gray wolf dataset. (A-D) Displays PC bi-plots of the top seven PCs plotted against each other. The colors represent predefined ecotypes defined in (Schweizer et al., 2016). We can see that the top PCs delineate these predefined ecotypes. (E) Shows a "scree" plot with the proportion of variance explained for each of the top 50 PCs. As expected by genetic data (Patterson et al., 2006), the eigen-values of the genotype matrix tend to be spread over many PCs.

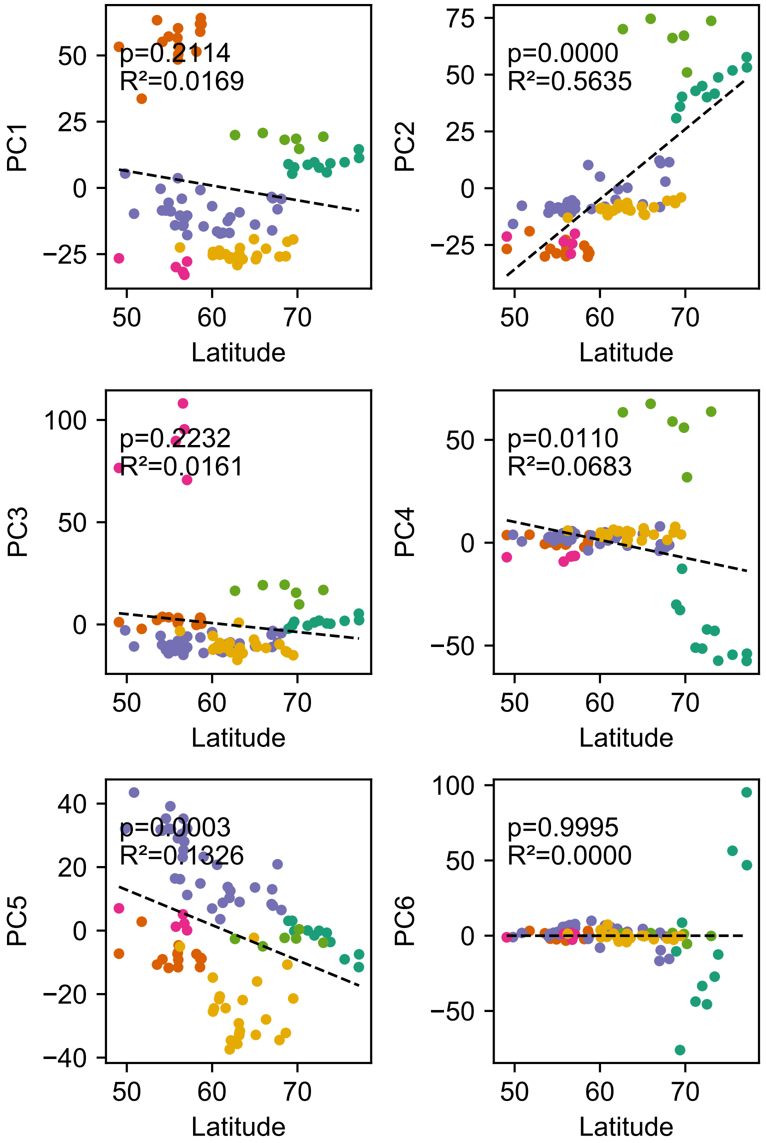


Figure 3.12: **Relationship between top axes of genetic variation and latitude:** In each sub-panel we plot the PC value against latitude for each sample in gray the wolf dataset. We see many of the top PCs are significantly correlated with latitude as tested by linear regression.

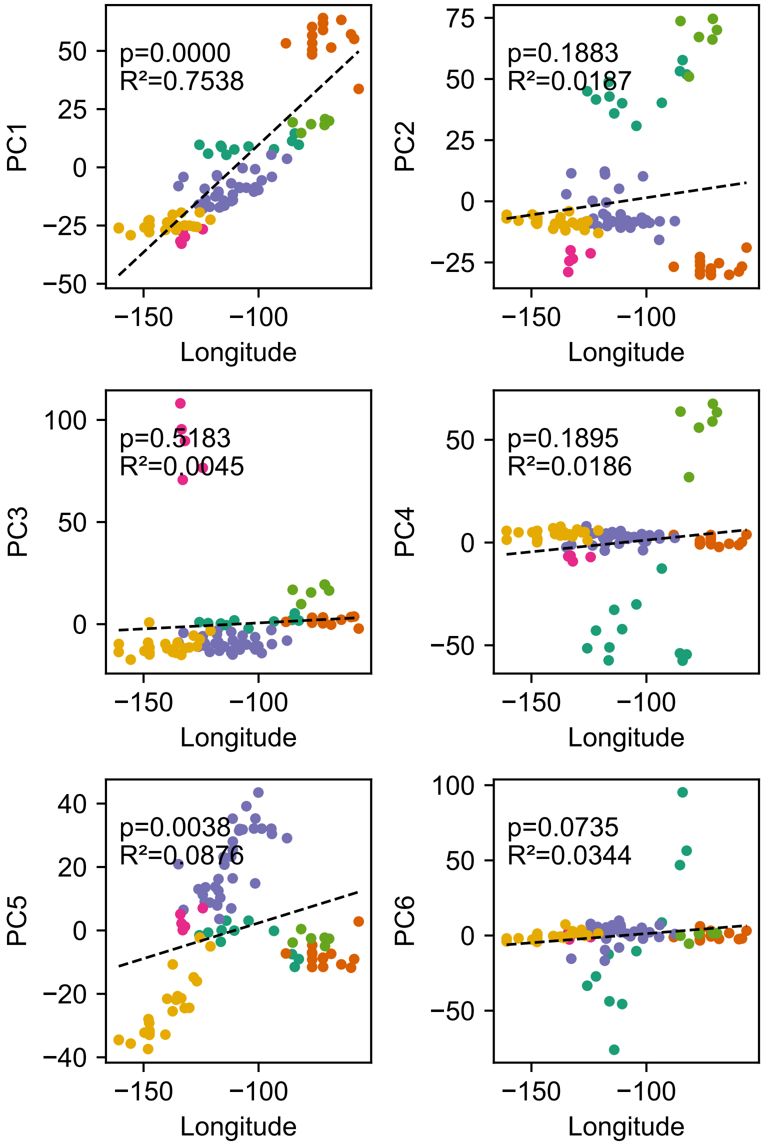


Figure 3.13: **Relationship between top axes of genetic variation and longitude:** In each sub-panel we plot the PC value against longitude for each sample in the gray wolf dataset. We see many of the top PCs are significantly correlated with longitude as tested by linear regression.

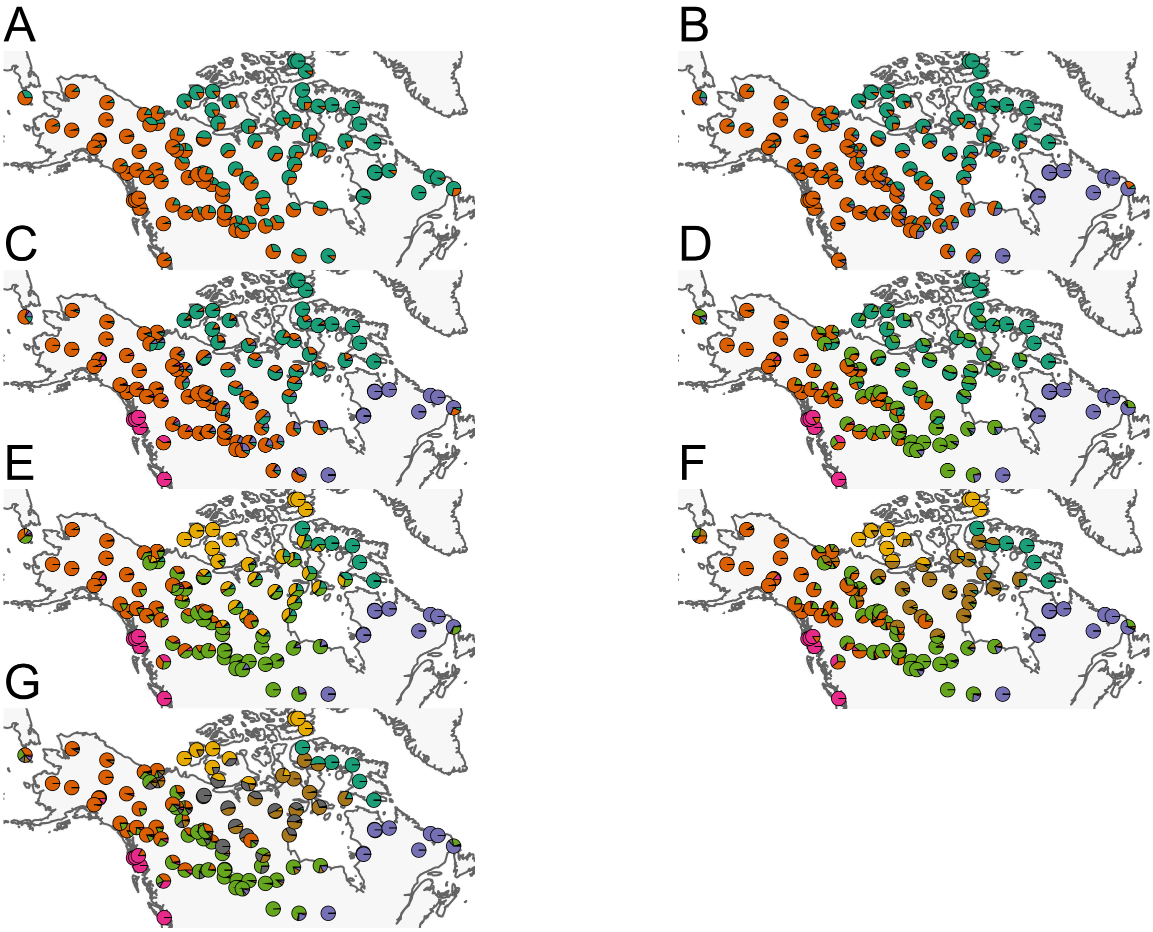


Figure 3.14: **Summary of ADMIXTURE results:** (A-G) Visualization of ADMIXTURE results for $K = 2$ to $K = 8$. We display admixture fractions for each sample as colored slices of the pie chart on the map. For each K we ran 5 replicate runs of ADMIXTURE and in this visualization we display the solution that achieves the highest likelihood amongst the replicates. The ADMIXTURE results qualitatively reveal a spatial signal in the data as admixture fractions tend to be spatially clustered.

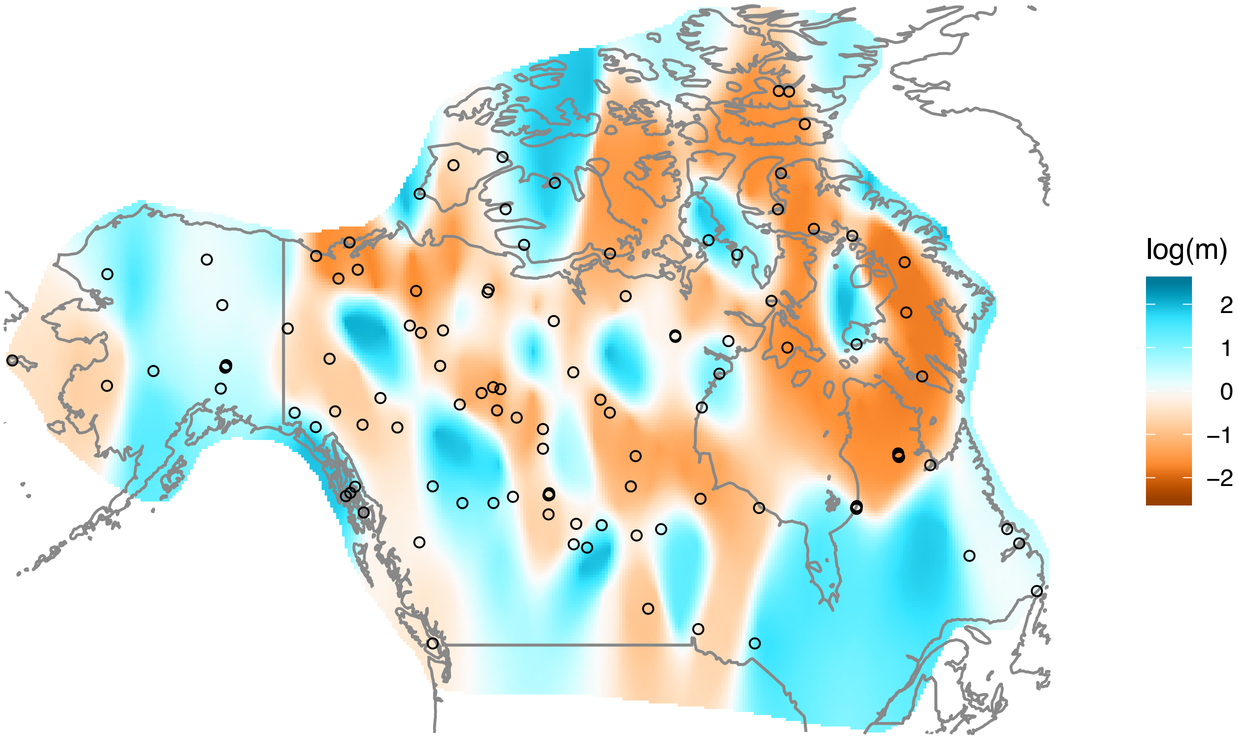


Figure 3.15: **Application of EEMS to the North American gray wolf dataset:** We display a visualization of EEMS applied to the North American gray wolf dataset. The more orange colors represent lower than average effective migration on the log-scale and the more blue colors represent higher than average effective migration on the log-scale. The results of EEMS are qualitatively similar to FEEMS.

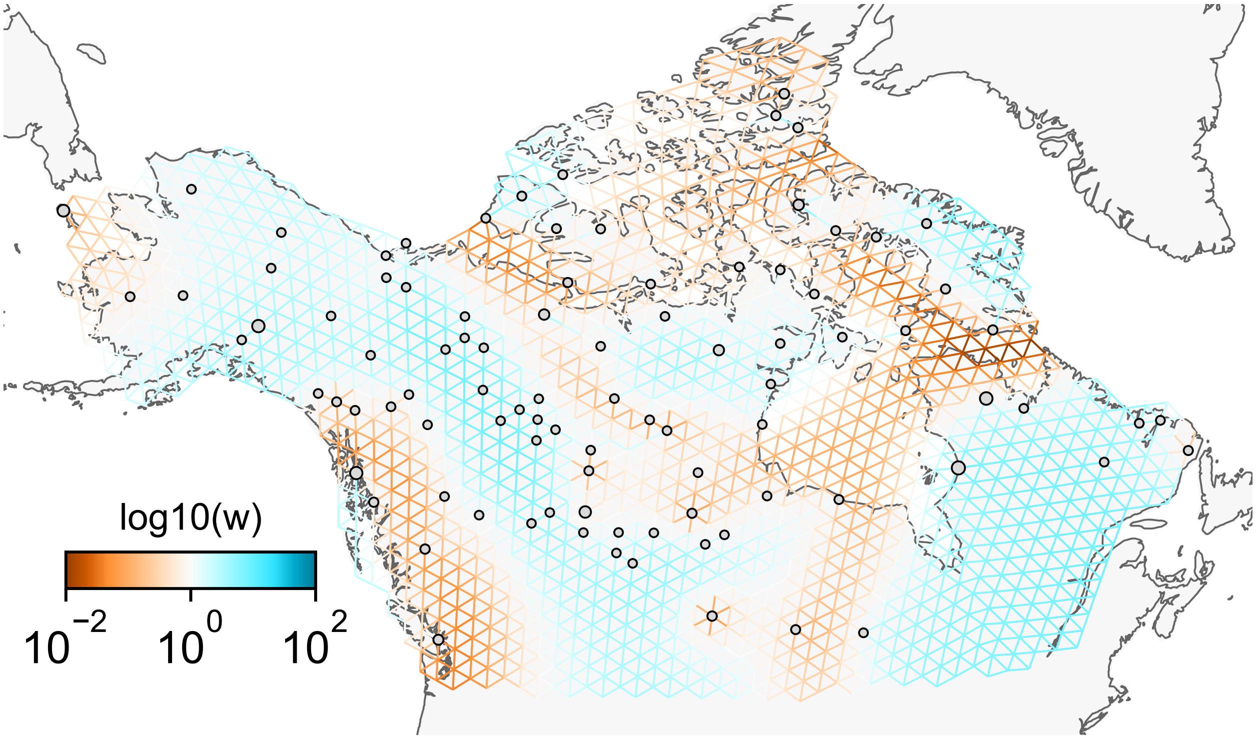


Figure 3.16: **Application of FEEMS on the North American gray wolf dataset with an exact likelihood model:** We display the fit of FEEMS based in the formulation (3.19) to the North American gray wolf dataset. This fit corresponds to a setting of tuning parameters at $\lambda = 10^{-3}, \alpha = 50$. Additionally we set the lower bound of the edge weights to $\boldsymbol{l} = 0.01$, to ensure that the diagonal elements of $\boldsymbol{L}$ does not become too small—this has an implicit effect on $L^{\dagger}_{kk}$, preventing it from blowing up at unobserved nodes. The more orange colors represent lower than average effective migration on the log-scale and the more blue colors represent higher than average effective migration on the log-scale. Visually the result is comparable to that of FEEMS fit (Figure 4.5) based in the formulation (3.9).

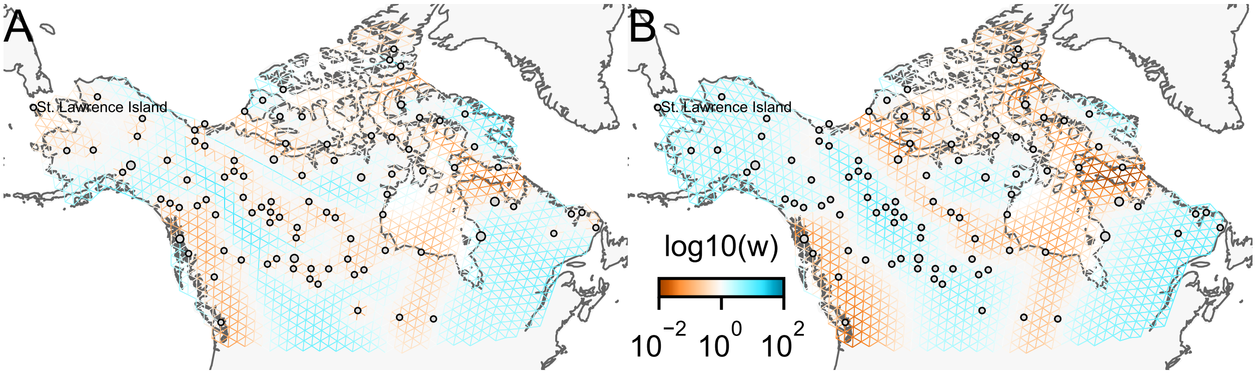


Figure 3.17: **Application of FEEMS on the North American gray wolf dataset with joint estimation of the residual variance and graph's edge weights:** We show visualizations of fits of FEEMS to the North American gray wolf dataset when the residual variance and edge weights of the graph are jointly estimated. Both fits correspond to a setting of tuning parameters at $\lambda = 10^{-3}, \alpha = 50$. (A) Displays the estimated effective migration surfaces where every deme shares a single residual parameter σ. The result is similar to the procedure that prefixes σ from the homogeneous isolation by distance model (Figure 4.5), except the high migration edge forming long path in (A) which disappears with higher values of α. (B) Displays the estimated effective migration surfaces where each node has its own residual parameter σ_k for all nodes k. These node specific residual parameters allow more flexible graphs, but at the cost of over-fitting to the data. In particular, without adding smooth regularization term on the residual variances, it fails to recover some geographic features like St. Lawerence Island.

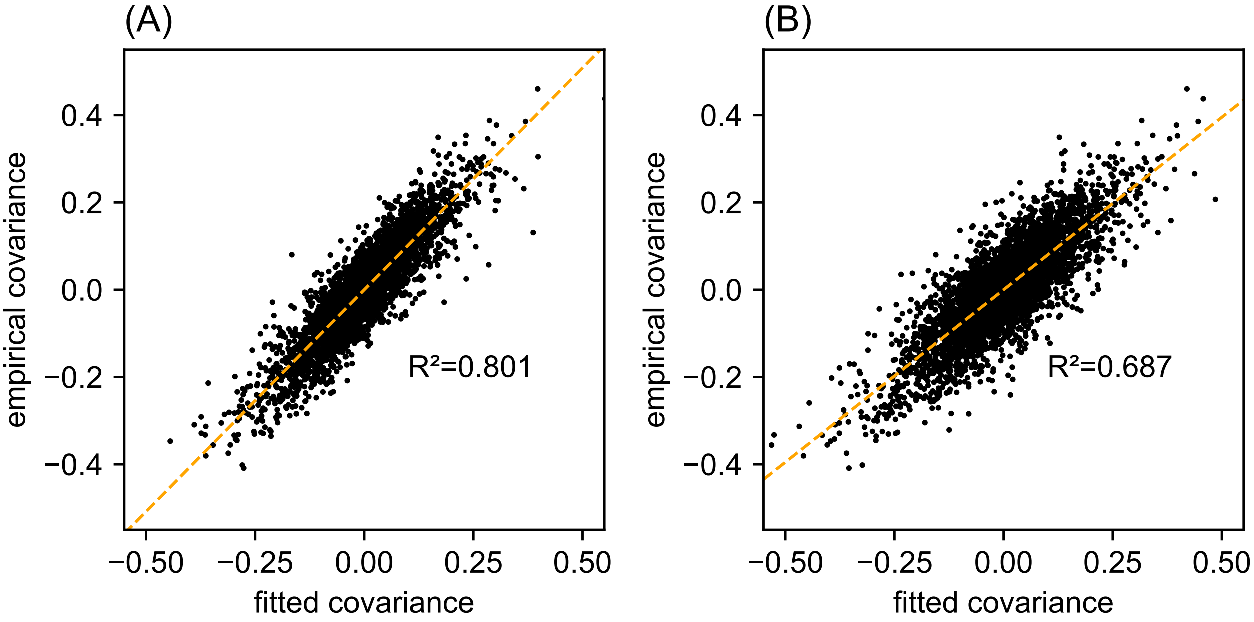


Figure 3.18: **Relationship between fitted and empirical covariance on the North American gray wolf dataset:** We display scatter plots of empirical genetic covariances versus fitted covariances from FEEMS fits on the gray wolf dataset. (A) Corresponds to the result shown in Figure 4.5. (B) Corresponds to the result shown in Supp. Fig. 3.17B. The x-axis represents the transformed fitted covariance matrix, i.e. $\boldsymbol{C}\boldsymbol{A}\widehat{\boldsymbol{L}}^{\dagger}\boldsymbol{A}^{\top}\boldsymbol{C}^{\top}+\widehat{\sigma}^{2}\boldsymbol{C}\mathrm{diag}\left(\boldsymbol{n}^{-1}\right)\boldsymbol{C}^{\top}$ (see equation (3.6)). The y-axis represents the transformed sample covariance matrix, i.e. $\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top}$. The simple linear regression fit is shown in orange dashed lines and R^2 is given.

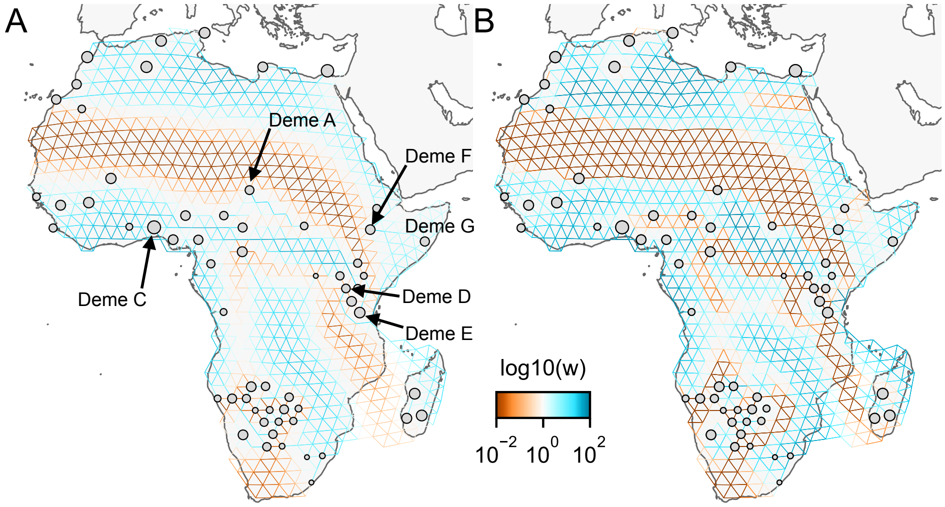


Figure 3.19: **Application of FEEMS to a dataset of human genetic variation from Africa with different parameterization:** We display visualizations of FEEMS to a dataset of human genetic variation from Africa with different parameterization of the graph's edge weights. See (Peter et al., 2018) for the description of the dataset. (A) Displays the recovered graph under the edge parameterization. (B) Displays the recovered graph under the node parameterization. Both parameterization have their own regularization parameters λ and α, but these parameters are not on the same scale. We set $\lambda = 2 \cdot 10^{-4}, \alpha = 10$ for the node parameterization which is seen to yield similar results to those in (Peter et al., 2018). For the edge parameterization, we keep the same λ value while we set $\alpha = 60$ so that the resulting graph reveals similar geographic structure to the node parameterization. We also set the lower bound $\boldsymbol{l} = 0.01$. From the plots, it is worth noting two important distinctions: (1) We see the migration surfaces shown in (B) recover sharper edge features while the migration surfaces in (A) are overall smoother. This is attributed to the fact that node parameterization has its own additional regularization effect on the edge weights, and in order to achieve similar degree of regularization strength for the edge parameterization, it needs a higher regularization parameters, which results in more blurring edges than the node parameterization. (2) When measuring correlation of the estimated allele frequencies among nodes, we find that Deme B is the node with the second highest correlation to Deme A, whereas Deme C (and nearby demes) is not as much correlated to Deme A compared to Deme B. Panel (A) reflects this feature by exhibiting a corridor between Deme A and Deme B and reduced gene-flow beneath that corridor. This reduced gene-flow disappears in (B), even if the regularization parameters are varied over a range of values. Additionally, Deme D is most highly correlated to Deme E, F, and G, and this is implicated by a long-range corridor connecting those demes appearing in Panel (A) while not shown in (B). These results point a conclusion that the form of the node parameterization is perhaps too strong and in this case it limits model's ability to capture desirable geographic features that are subtle to detect.

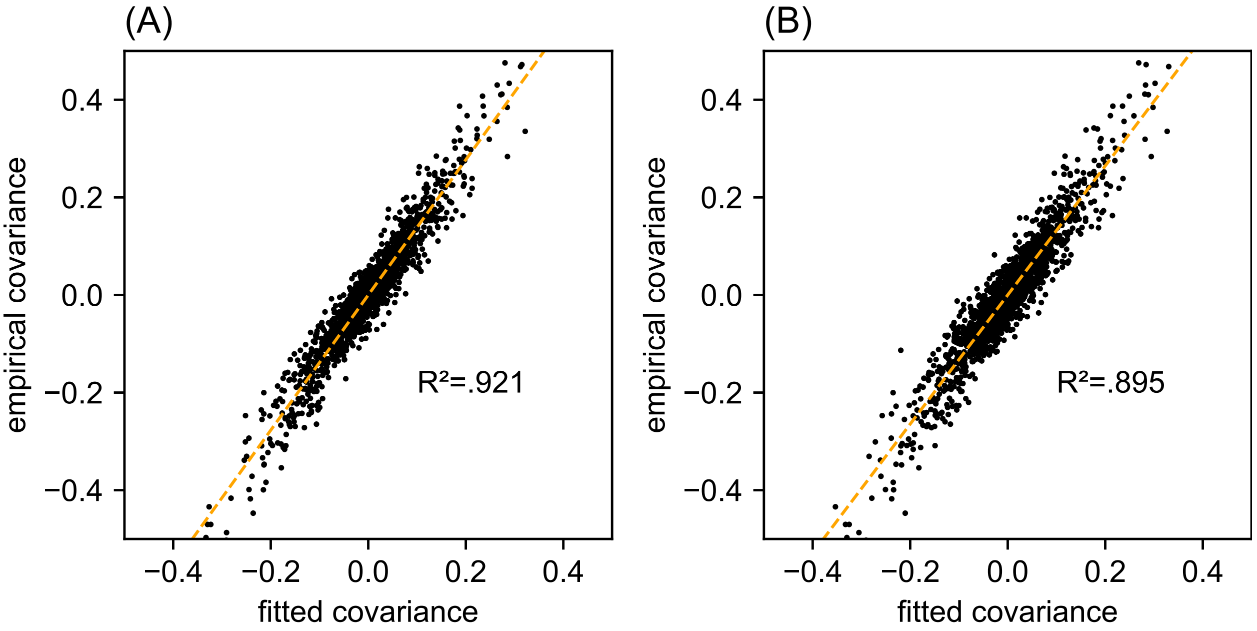


Figure 3.20: **Relationship between fitted and empirical covariance on a dataset of human genetic variation from Africa:** We display scatter plots of empirical genetic covariance versus fitted covariance from FEEMS fits on the African dataset. (A) Corresponds to the result shown in Supp. Fig. 3.19A. (B) Corresponds to the result shown in Supp. Fig. 3.19B. The x-axis represents the transformed fitted covariance matrix, i.e. $\boldsymbol{C}\boldsymbol{A}\widehat{\boldsymbol{L}}^{\dagger}\boldsymbol{A}^{\top}\boldsymbol{C}^{\top} + \widehat{\sigma}^2\boldsymbol{C}\mathrm{diag}\left(\boldsymbol{n}^{-1}\right)\boldsymbol{C}^{\top}$ (see equation (3.6)). The y-axis represents the transformed sample covariance matrix, i.e. $\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}\boldsymbol{C}^{\top}$. The simple linear regression fit is shown in orange dashed lines and R^2 is given.

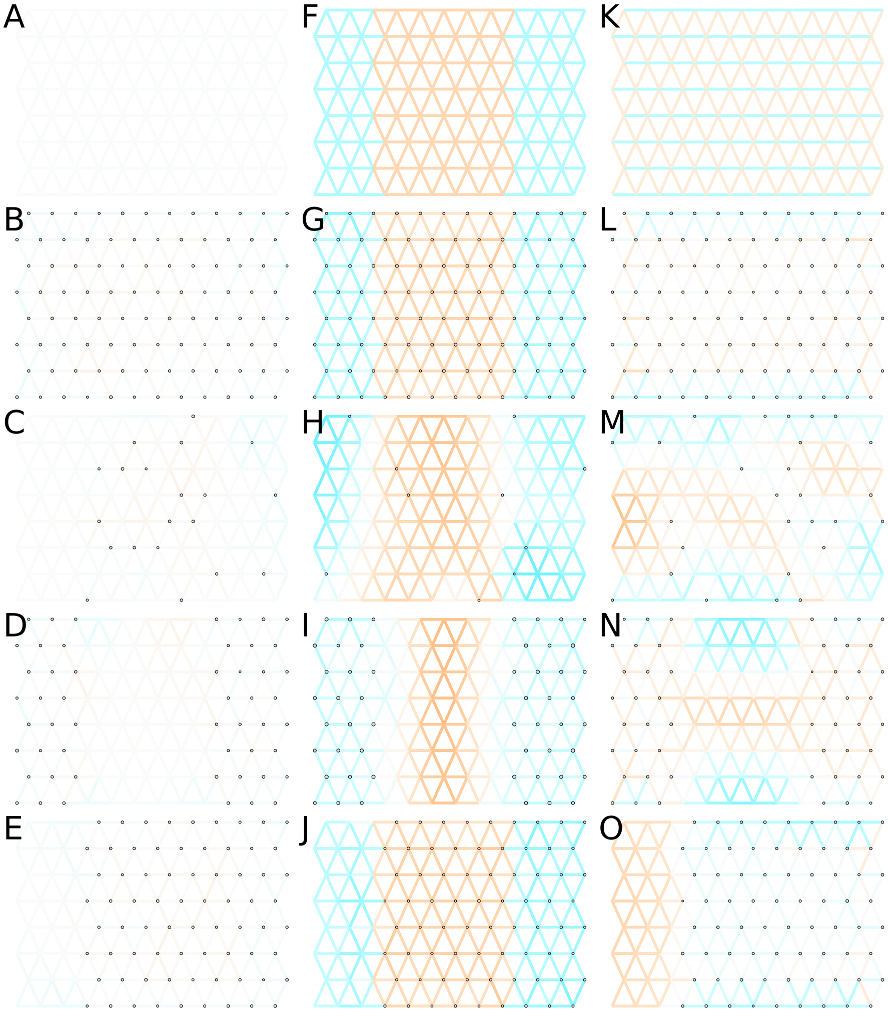


Figure 3.21: **Application of FEEMS based on node parameterization to an extended set of coalescent simulations:** We display an extended set of coalescent simulations with the same migration scenarios and sampling designs as Supp. Fig. 3.6. The sample sizes across the grid are represented by the size of the grey dots at each node. The migration rates are obtained by solving the FEEMS objective function (3.9) with node parameterization where the regularization parameters are specified at $\lambda = 10^{-3}, \alpha = 50$. (A, F, K) display the ground truth of the underlying migration rates. (B, G, L) Shows simulations where there is no missing data on the graph. (C, H, M) Shows simulations with sparse observations and nodes missing at random. (D, I, N) Shows simulations of biased sampling where there are no samples from the center of the simulated habitat. (E, J, O) Shows simulations of biased sampling where there are only samples on the right side of the habitat.

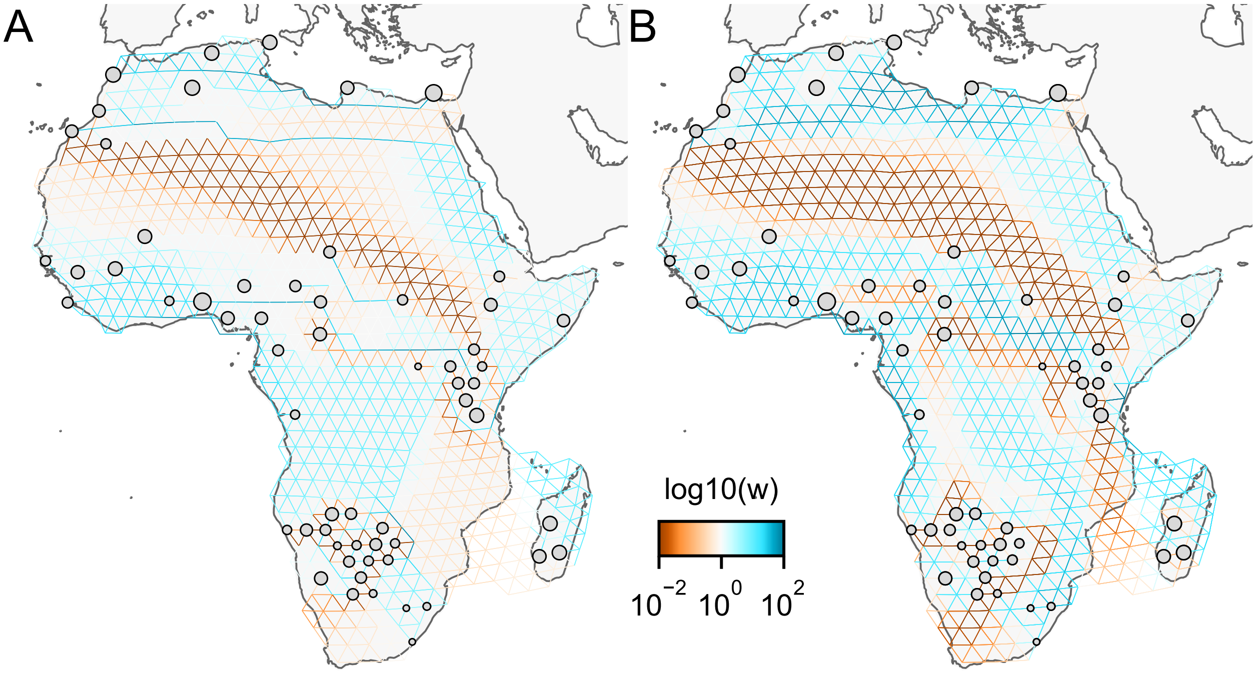


Figure 3.22: **Application of ℓ_1-norm-based FEEMS to a dataset of human genetic variation from Africa:** We display visualizations of FEEMS to a dataset of human genetic variation from Africa with the ℓ_1-based penalty function. See (Peter et al., 2018) for the description of the dataset. (A) Displays the recovered graph under the edge parameterization with ℓ_1 norm based penalty where the regularization parameters are specified at $\lambda = 4 \cdot 10^{-2}, \alpha = 30$. (B) Displays the recovered graph under the node parameterization with ℓ_1 norm based penalty where the regularization parameters are specified at $\lambda = 4 \cdot 10^{-2}, \alpha = 1$. To minimize the objective (3.20), linearized ADMM is applied with $20,000$ number of iterations. The lower bound is set to be $\boldsymbol{l} = 0.01$ for both parameterizations. Note that due to the high degrees of missingness, the estimated effective migration surfaces using solely ℓ_1-based penalty exhibit many likely artifacts (e.g., high migration edges forming long paths, seen in A) unless an additional penalty term is added to promote global smoothness of the edge weights such as a combination of ℓ_1 norm penalty function and node parameterization as shown in (B).

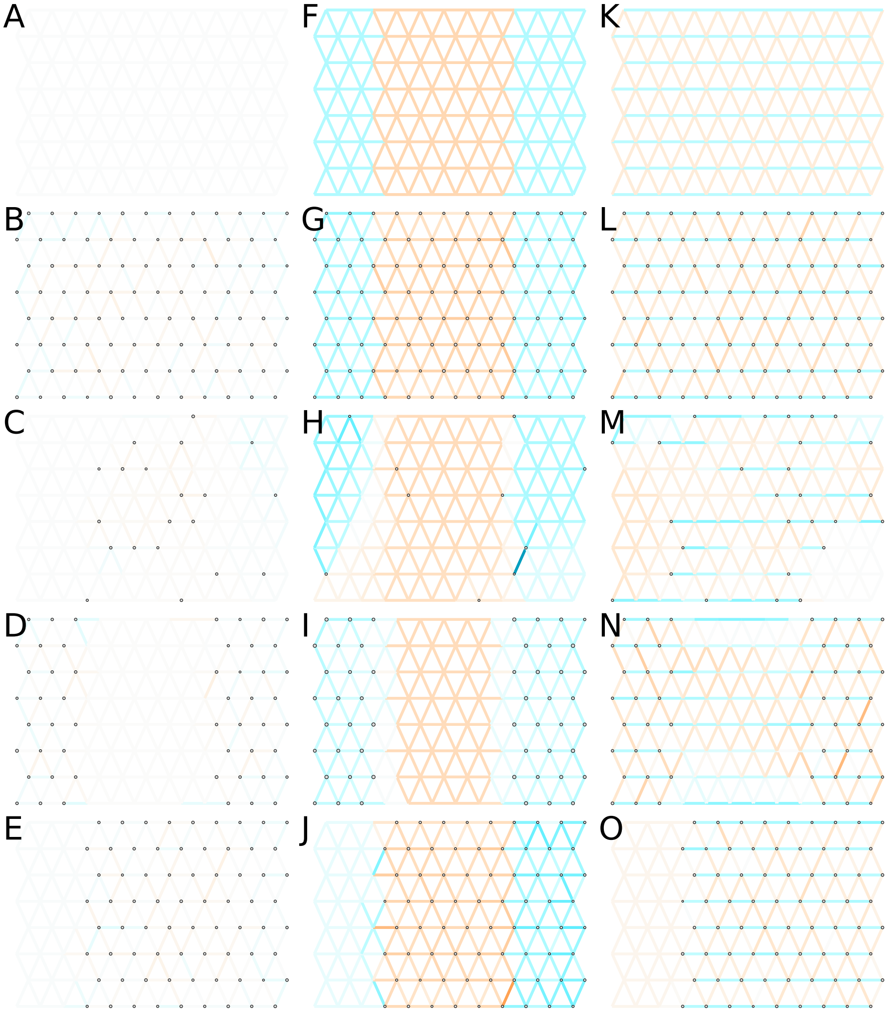


Figure 3.23: **Application of ℓ_1-norm-based FEEMS to an extended set of coalescent simulations:** We display an extended set of coalescent simulations with the same migration scenarios and sampling designs as Supp. Fig. 3.6. The sample sizes across the grid are represented by the size of the grey dots at each node. The migration rates are obtained by solving ℓ_1 norm based FEEMS objective (3.20) where the regularization parameters are specified at $\lambda = 10^{-1}, \alpha = 30$ (I), $\lambda = 10^{-3}, \alpha = 30$ (N), and $\lambda = 10^{-2}, \alpha = 30$ for the rest. (A, F, K) display the ground truth of the underlying migration rates. (B, G, L) Shows simulations where there is no missing data on the graph. (C, H, M) Shows simulations with sparse observations and nodes missing at random. (D, I, N) Shows simulations of biased sampling where there are no samples from the center of the simulated habitat. (E, J, O) Shows simulations of biased sampling where there are only samples on the right side of the habitat.

CHAPTER 4

EMPHASIZING SHARED EVOLUTIONARY HISTORIES WHEN INFERRING REPRESENTATIONS OF POPULATION STRUCTURE

Joseph H. Marcus, Jason Willwerscheid*, Peter Carbonetto, John Novembre, and Matthew Stephens*

* *denotes co-first authorship*

4.1 Abstract

Describing the genetic relatedness in a population genetic sample, often referred to as population structure, is a long-standing and important problem for many applications. One current challenge in population structure inference is that the input genotype matrix is high dimensional with datasets often containing thousands of samples and hundreds of thousands of genetic variants. Matrix factorization has been a unifying tool underlying many methods to reduce the dimensionality of the genotype matrix and find interpretable biological structure in the data. Existing matrix factorization methods when applied to population genetic data tend to emphasize clustered solutions due to constraints assumed in the models to fit the data. These constraints can be unnatural for population genetic data where the samples experience shared evolutionary change over time, often referred to as "shared genetic drift". Here we propose a new Bayesian matrix factorization method called *drift* that emphasizes shared evolutionary histories in the output summaries of population structure. Particularly, we assume an individual's expected genotype can be decomposed into a linear combination of unconstrained variables representing ancestral shared allele frequency change and where the coefficients of this linear combination are constrained to be between zero and one. These assumptions result in a new probabilistic semi-non-negative matrix

factorization model for population structure inference. We develop an efficient variational inference algorithm that learns approximate posterior distributions using an empirical Bayes inspired optimization algorithm. We apply *drift* to simulations and large-scale population genetic datasets from multiple species and find it has advantages over existing methods but is fraught with optimization challenges that need to be solved before widespread use and application. Overall *drift* is a promising new framework for population structure inference and lays the groundwork for exciting new methods and applications.

4.2 Introduction

Using genetic variation to understand the relationship among individuals is important for several applications. In genome-wide association studies estimates of genetic relatedness, often the top principal components (PCs) computed from a large genotype matrix, are used to control for confounding of sub-population membership with a quantitative trait or case-control status (Price et al., 2006; Yang et al., 2011). In studies of demographic history, genetic relatedness within a population genetic sample is used to learn about migration events or within-population processes that have shaped patterns of current and ancient genetic diversity (Schraiber and Akey, 2015; Pickrell and Reich, 2014). Systematic variation in genetic relatedness in a population genetic sample is often referred to as population structure, because within a meta-population, individuals are rarely randomly mating and hence the population is sub-structured (Hao and Storey, 2019; Lawson et al., 2012). When estimating population structure, researchers are commonly faced with a genotype matrix where the rows represent individuals and the columns represent different genetic variants that have been typed along the genome (Novembre and Peter, 2016). The elements of this matrix store the count of some predefined allele, for instance the minor allele, for each individual and genetic variant. Visualizing the genotype matrix is difficult because of its high dimensionality. In typical applications a researcher is faced with thousands of samples and hundreds of thousands of genetic variants. There are many ways to represent population structure using this large-scale genotype matrix as input, but matrix factorization has been an essential and unifying tool among many of the common approaches taken (Engelhardt and Stephens, 2010; Wang and Stephens, 2018; Cabreros and Storey, 2019).

Principal Component Analysis (PCA) is a widely used method to visualize population structure by reducing the dimensionality of the genotype matrix and focusing on axes of variation that are most important in the data (Novembre et al., 2008). One way to interpret PCA, among many, is that it is the solution to an optimization problem that finds the best rank-K approximation to a data matrix (Agrawal et al., 2020; Murphy, 2012). Because of the orthogonality constraints of the underlying factors that compose this low-rank approximation the optimization problem is solved efficiently and robustly using the singular value decomposition (SVD) (Strang, 2019). This computational tractability and simplicity have made PCA the "workhorse" of population structure inference, yet it has many limitations. While the orthogonality constraint of PCA is computationally attractive, it is difficult to interpret or provide a biological motivation (Engelhardt and Stephens, 2010). PCA has also been shown to be difficult to interpret in several population genetic models and datasets including homogeneous spatial processes with biased sampling (McVean, 2009; Novembre and Stephens, 2008). Importantly, it is difficult to visualize many PCs at once and typically each PC is plotted against each other in "bi-plots" leading to dozens of visualizations to interpret.

The Pritchard, Stephens and Donnelly (PSD) model has also been an essential tool in inferring population structure as it allows for visualization of many underlying latent factors at once in a stacked bar chart, commonly referred to as a "structure plot" (Pritchard et al., 2000). Since its initial conception many different algorithms, both Bayesian and maximum likelihood based, have been proposed to fit the PSD model (Alexander et al., 2009; Raj et al., 2014; Gopalan et al., 2016; Cabreros and Storey, 2019; Frichot et al., 2014). Fundamentally, the PSD model can be seen as a matrix factorization method where the individual's loadings (admixture proportions) are constrained to be non-negative and sum to one (defined on the K-dimensional simplex) and the SNP factors are constrained to be probabilities (Engelhardt and Stephens, 2010; Alexander et al., 2009). These constraints are motivated by an interpretation of each individual's genome-wide average "ancestry" as coming from a mixture of K different latent source ancestries (Alexander et al., 2009; Pritchard et al., 2000). An analogous framework called topic modeling was developed in the machine learning community to find interpretable low-dimensional representations of text in documents (Blei et al., 2003). While the biological motivation of admixture models has made them extremely popular and widely used they output results that are difficult to interpret under a number of important population genetic processes (Lawson et al., 2018). Particularly, due to the simplex constraint, admixture models tend to provide "clustered" results and poorly represent shared evolutionary histories in a single visualization (Lawson et al., 2018).

Many extensions of PCA and admixture models have been proposed to help alleviate some of the problems of these approaches. When side information about the spatial locations or temporal periods of the samples is available, multiple methods have been proposed to account for spatial or temporal confounding in the results of PCA or PSD like models (Bradburd et al., 2018; Caye et al., 2018; François et al., 2019; Frichot et al., 2012; Joseph and Pe'er, 2019b). Other approaches have emphasized that the matrix factors should be sparse to help boost interpretability (Frichot et al., 2014; Engelhardt and Stephens, 2010). Our work heavily depends on and extends the statistical model and algorithms developed in Wang and Stephens (2018), which takes an empirical Bayes approach to sparse matrix factorization. In their proposed model called "Factors and Loadings by Adaptive Shrinkage" (FLASH), flexible non-parametric mixture model priors are used to approximate any unimodal distribution with a specified mode, naturally assumed to be zero for a shrinkage effect (Stephens, 2017). This unimodal assumption encompasses a broad family of shrinkage priors and allows the data to decide what type of regularization (strength and shape of shrinkage) is best. The prior parameters are estimated from the data in an empirical Bayes inspired variational inference algorithm, allowing each latent factor to come from its own distribution with a specific regularization strength (Blei et al., 2017).

Here we develop a new model called *drift* which helps to emphasize shared evolutionary histories experienced by individuals, leading to a more natural and easily interpretable representation of population structure then previously proposed approaches like PCA or admixture models. In our work we extend the flexible and computationally efficient approach taken by Wang and Stephens (2018) but design a new model and algorithm tailored for population structure inference. Particularly, we develop a new Bayesian semi-non-negative matrix factorization algorithm that can be fit using a similar empirical Bayes variational inference algorithm but with a new bi-modal shrinkage prior family and variational approximation that allows for correlations in the approximate posterior for the factors (Ding et al., 2008). We demonstrate through simulations and empirical applications that the *drift* model is a promising approach for population structure inference but also has some optimization challenges.

4.3 Results

4.3.1 Overview of the drift model

The input to the *drift* model, like many other approaches, is a genotype matrix, $\boldsymbol{Y}$, with n rows representing individuals and p columns representing SNPs. The elements of this matrix, $y_{ij} \in \{0, 1, 2\}$, store the count of chromosomes that are labeled with some predefined allele. We typically orient to the derived allele when ancestral/derived alleles status information is available. To motivate our model, Figure 4.1 shows an idealized evolutionary scenario with individuals from three sub-populations sampled at the present, A, B, and C. The sub-populations are related to one another through a simple tree with two splits, where individuals B and C are sampled from sub-populations that share a more recent common ancestor. In this idealized tree, the genotypes of individuals A, B and C arise from independent allele frequency drift that occurs on branches 3, 4, and 5 as well as shared drift that occurs on internal branches 1 and 2. We use a Brownian motion approximation to allele frequency diffusion on the tree which has been commonly applied in the population genetics literature (Felsenstein, 1973). Here we refer to the branches of the tree as “drift events”, some of which are experienced by more than one sub-population, “shared events”, and some of which are “specific” to particular sub-populations. Note that we use the term “event” to describe some notion of time averaged allele frequency change and is thus not a discrete characterization of a evolutionary event that occurred instantaneously. We can then formulate the genotypes arising from this idealized model as a linear combination of SNP-specific changes in allele frequencies that occur on each of these K branches or drift events (McCullagh, 2009):

$$y_{ij} = \sum_{k=1}^{K} \ell_{ik} f_{jk} + e_{ij}. \tag{4.1}$$

Here, $\ell_{ik} \in [0, 1]$ is an individual-specific loading on the kth drift event, $f_{jk} \in \mathbb{R}$ is the change in allele frequency that occurs in the kth drift event for the jth SNP, $e_{ij} \sim \mathcal{N}(0, \sigma_r^2)$ is residual Gaussian noise and σ_r^2 is a residual variance parameter that is shared across all individuals and SNPs. Note that a more realistic error model for the genotype count data is a Binomial sampling model with $y_{ij} \sim \text{Binomial}(2, \pi_{ij})$, assuming Hardy-Weinberg equilibrium, but we make a Gaussian approximation for computational tractability (Hao et al., 2016). This Gaussian approximation should hold well for common variants whose frequencies are not too close to the boundaries of 0 or 1.

This model is similar to admixture models in that the individual loadings are, like the admixture proportions, non-negative, but one crucial difference is that in the *drift* model the loadings are bounded between 0 and 1, whereas in admixture models the admixture proportions for the ith individual are additionally constrained to sum to 1 (Pritchard et al., 2000; Alexander et al., 2009). In *drift*, each individual’s expected genotype is modeled as a linear combination of K unconstrained variables that represent genetic drift in an ancestral component of shared evolutionary change. We relax the simplex constraint on the coefficients in this linear combination to model how individuals may differ from the mean due to the effects of multiple ”shared drift” events with varying magnitudes.

We assume the loadings for each drift event are independent and identically distributed

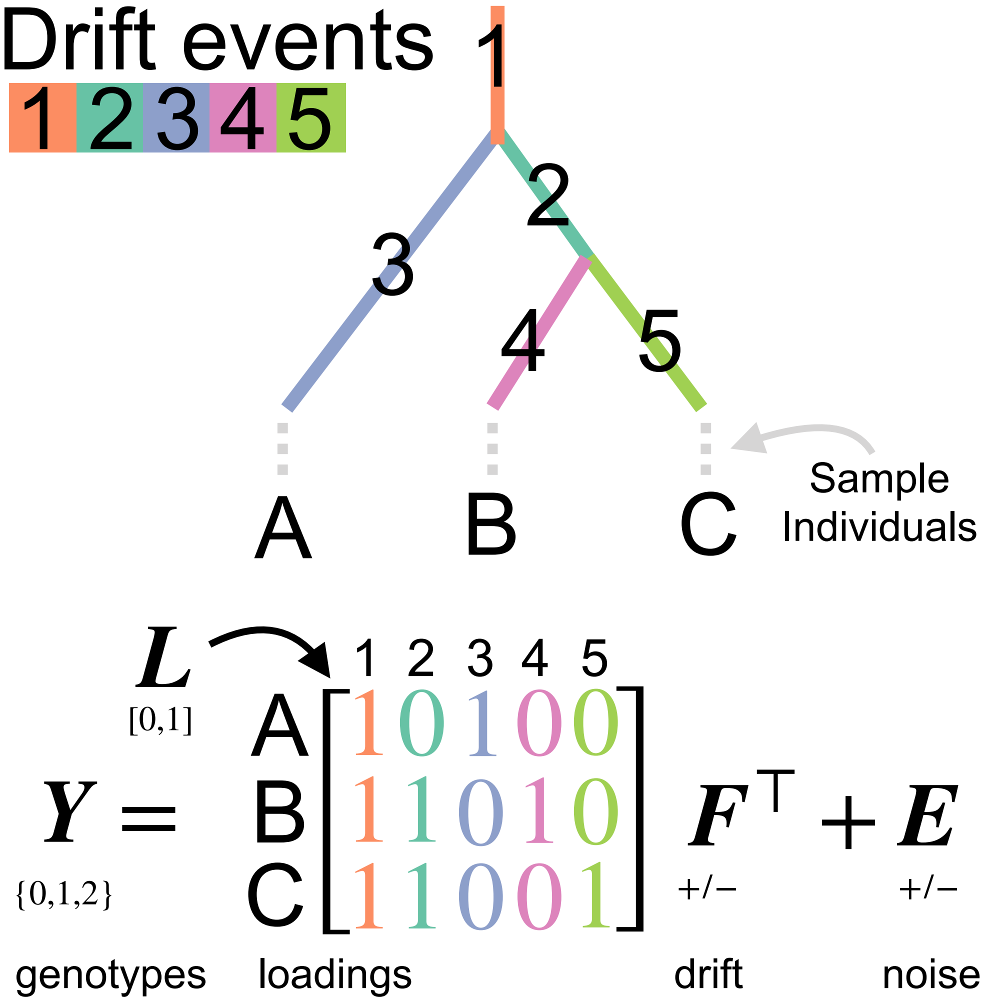


Figure 4.1: **Schematic of a motivating evolutionary scenario for the *drift* model**: This figure displays a schematic of a tree model for three individuals A, B, and C that are each sampled from sub-populations at the tips of the tree. While, in general, our method does not constrain a tree structure to the data we find it a useful motivating example on how "drift events", which correspond to shared allele frequency drift on each branch of the tree, can be represented in a matrix factorization frame-work. Here we show both the schematic tree and the corresponding loadings matrix and factorization model that is associated with the underlying evolutionary history.

from the following prior,

$$\ell_{1k}, \ell_{2k}, \ldots, \ell_{nk} \overset{iid}{\sim} g_k \in \mathcal{G}_{\text{BM}}, \tag{4.2}$$

where g_k is to be estimated from among the family $\mathcal{G}_{\text{BM}}$ of all bi-modal distributions on

$[0,1]$ with modes at 0 and 1. This bi-modal prior helps to regularize the loadings towards interpretable "tree-like" solutions where individuals either experience or do not experience particular drift events. We emphasize, however, that these priors are flexible enough to adapt to admixed individuals that have intermediate loadings on drift events. Furthermore, we assume that the SNP-specific drift terms come from a Gaussian distribution,

$$f_{1k}, f_{2k}, \ldots, f_{pk} \overset{iid}{\sim} \mathcal{N}(0, \sigma_k^2), \tag{4.3}$$

where σ_k^2 represents the prior variance in the change in allele frequencies for the kth drift event. In the machine-learning literature this type of model is a special case of semi-non-negative matrix factorization because the loadings are non-negative and the factors are unconstrained (Ding et al., 2008).

The inference challenge is to compute posterior distributions for the loadings and factors conditional on the observed genotype data. This problem is not analytically tractable. Here we extend the approach taken by Wang and Stephens (2018) and develop a new inference algorithm which uses a variational approximation to the posterior distributions, allowing the problem to be approximately solved without recourse to computationally expensive methods such as Markov Chain Monte Carlo (MCMC) (Blei et al., 2017). While the Bayesian model described above could be fit using the approach outlined in Wang and Stephens (2018), we developed a new algorithm that allows for correlations in the approximate posteriors for the factors (Wang and Stephens (2018) assumes a mean-field approximation that factorizes over each drift event). This new variational approximation is better suited for the types of evolutionary scenarios we are interested in inferring because the true posterior under such scenarios could have strong correlations which, if not properly accounted for, can lead to fitted representations of the data that are less interpretable.

We particularly emphasize that while we motivated this work using a tree-like evolutionary scenario, and while interpretation is often eased by assuming that a tree-like population genetic model holds, the model we propose is not constrained to fit a tree or admixture graph and thus is much more general. See methods for more details on the model and variational inference algorithm we developed.

4.3.2 Reconstruction of shared evolutionary histories in a simple simulation

To confirm that the *drift* model performs well in the idealized evolutionary scenario displayed in Figure 4.1 and to contrast the results with those obtained using other commonly applied tools such as PCA and the PSD model, we simulated data under a generative matrix factorization model. We simulated 50 individuals for each sub-population A, B, and C, for a total of 150 individuals, and $20,000$ independent genetic variants. We set the ancestral mean allele frequency to be $\mu = .5$ for every genetic variant and the prior standard deviations for the factors to $\sigma_2 = .1, \sigma_2 = .1, \sigma_3 = .05, \sigma_4 = .05, \sigma_5 = .05$. We then simulated genotypes by truncating the linear combination of drift terms to be between 0 and 1 and then subsequently performing binomial sampling to generate the diploid genotypes assuming Hardy-Weinberg equilibrium within each sub-population i.e. $y_{ij} \sim Binomial(2, \pi_{ij})$. The loadings for each individual were fixed at the values corresponding to the simple tree described above (see the schematic matrix for 3 individuals in Figure 4.1). We then applied PCA, the PSD model as implemented by the ALStructure algorithm (Cabreros and Storey, 2019), and *drift* to the simulated data. We display the results for each of these method in Figure 4.2.

In Figure 4.2C we see the results of applying the PSD model to this idealized simulation for values $K = 2, \ldots, K = 5$. While generally we see individuals from sub-populations B and C having more similar admixture fractions throughout the visualization it is very difficult to interpret the results in light of the tree model the data was simulated under.

In Figure 4.2B we plot the first two principal components against one another. These PCs capture the majority of the proportion variance explained. The other PCs (not shown)

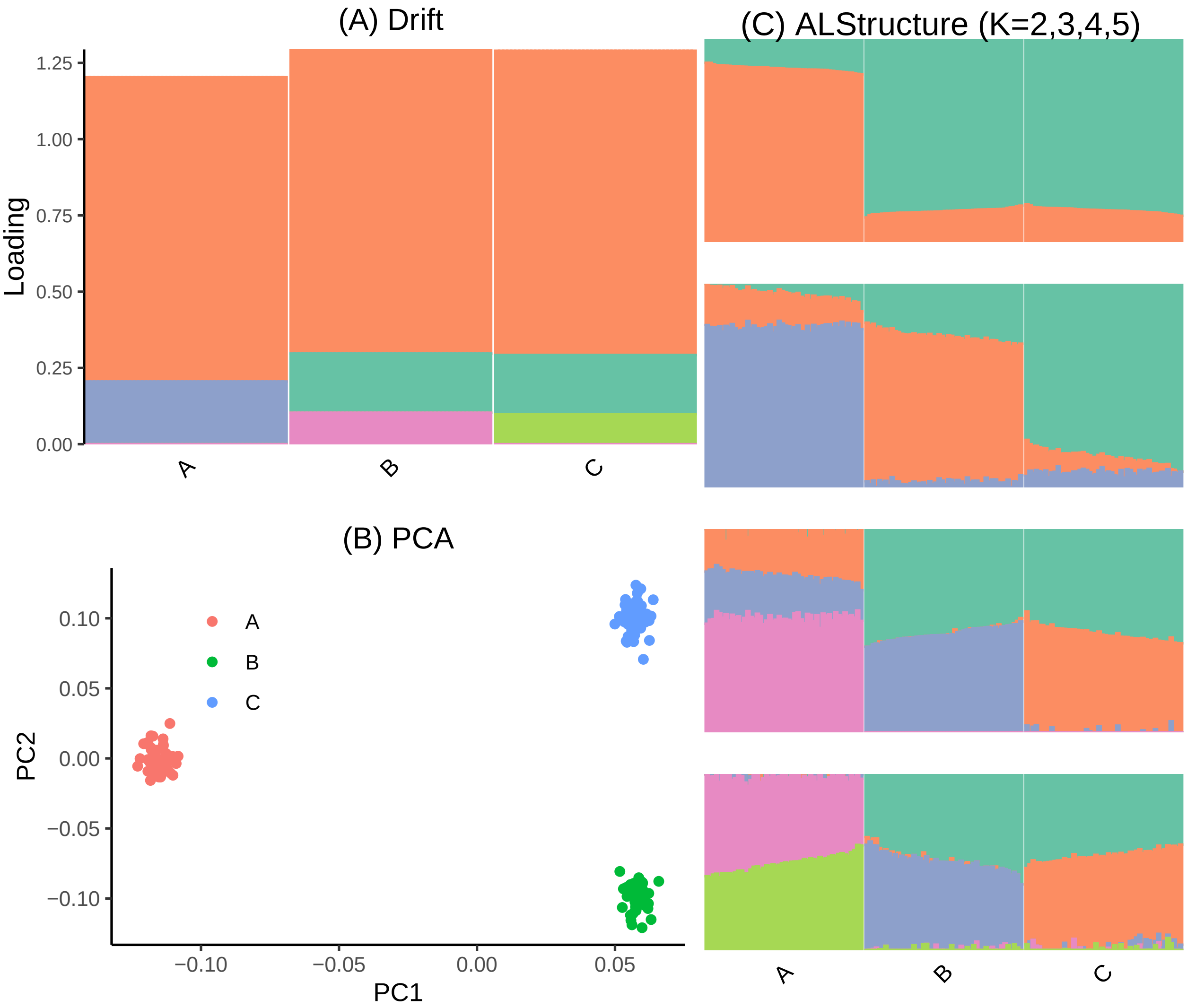
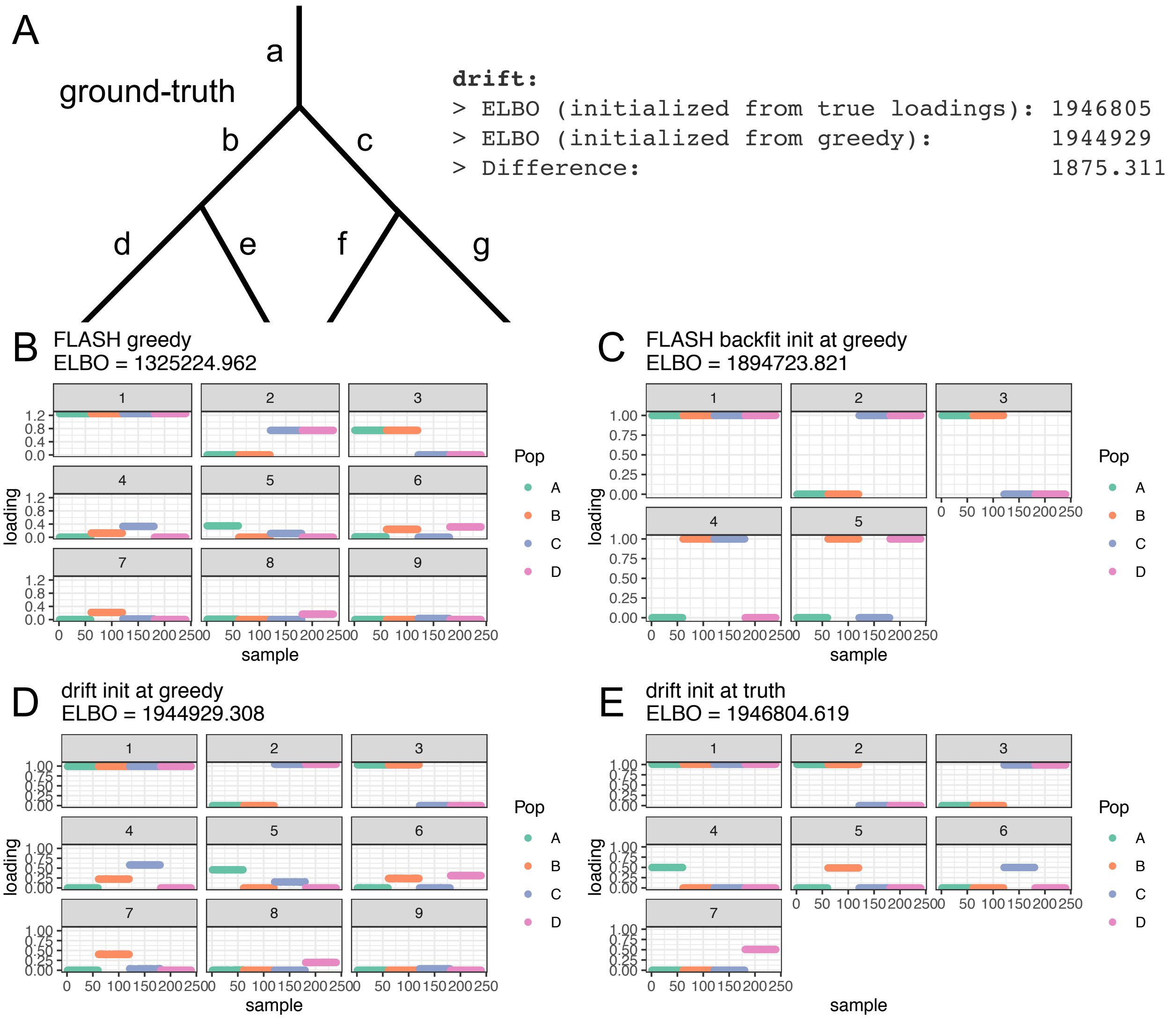


Figure 4.2: **Population structure inference methods applied to an idealized simulation**: We display the results of applying multiple population structure inference methods to an idealized 3-population tree simulation. 50 individuals are simulated from each sub-population at the tips of the tree. In the simulation the ancestral lineage that form individuals from sub-population A splits from the ancestral lineages of individuals from sub-populations B and C first and then the lineages that form B and C subsequently split. In panel A the results of the *drift* model are shown that naturally recapitulate the structure of the tree used to simulate the data. In B the top two PCs are shown. Finally, in C the results of fitting admixture from K=2 to K=5 are shown which is clearly difficult to interpret given the underlying structure of the tree. Note that we scale the drift loadings by the estimated prior variances of the corresponding drift events.

look like Gaussian noise. The sub-populations form three distinct clusters of points and we can read the tree-like nature of the data in the results by noting that PC1 separates individuals from sub-populations A from B and C and PC2 separates sub-populations B and C. That is, each PC corresponds to a split in the tree.

In Figure 4.2A we display the results of applying the *drift* model to the simulated data. The structure of the tree naturally pops out of the visualization. There is one shared drift event (orange) that is shared by all individuals. A second shared drift event (dark-green) is only shared by individuals from sub-populations B and C. Finally, each sub-population has a specific drift event (blue, purple and light-green, representing the external branches of the tree). The ability to capture shared drift events (orange and dark-green) is what primarily distinguishes our model from existing methods. While we acknowledge that real data is much more complex and, in particular, does not necessarily adhere to the strict assumptions of tree-like models, we believe that this pedagogical example usefully illustrates the potential advantages of our approach.

While we showed the advantage of the *drift* model in the three-population tree simulation, we also encountered optimization challenges in other simple tree simulation scenarios. We simulated data from a Gaussian matrix factorization representation of a four-population tree (Figure 4.3A). We compared the objective function values (Evidence Lower Bounds [ELBOs]) at convergence between two initialization strategies: (1) initializing with a greedy algorithm from the FLASH model with bi-modal priors for the loadings and Gaussian factors (2) initializing at the true loadings as specified by the tree. Both approaches showed evidence of converging, based on a small non-negative change in the ELBO between consecutive iterations, but the fit initialized at the truth shows a much higher value of the ELBO at convergence (a difference of 1875.311 log-likelihood units, Figure 4.3E). In Figure 4.3B and Figure 4.3D we display the fitted loadings for each factor for the greedy fit and the *drift* fit initialized at the greedy solution. We can see that there is no qualitative difference in the interpretation of the loadings between the two figures. While, both of the solutions do have some attractive features (e.g. they can capture the shared internal branches), the



```
drift:
> ELBO (initialized from true loadings): 1946805
> ELBO (initialized from greedy):        1944929
> Difference:                            1875.311
```

Figure 4.3: **Illustration of convergence to a local optimum in a simple simulation:** We display a set of visualizations to illustrate the convergence of the *drift* algorithm to a local optimum and sensitivity to initialization. In A we show the ground truth four-population tree we simulated data from and values of the ELBO initialized using the FLASH greedy algorithm, initialized from the ground truth, and the difference between these ELBOs. In B we show the fitted loadings from the FLASH greedy algorithm. In C we showed the fitted loadings from the FLASH backfitting algorithm initialized at the greedy solution. In D we show the fitted loadings from the *drift* algorithm initialized at the greedy solution. In E we show the fitted loadings from the *drift* algorithm initialized at the truth.

population specific factors, for the most part, are not recovered. Interestingly when we run the FLASH backfitting algorithm, initialized from the same greedy fit, it recovers a "sparser" representation of the tree, only using five factors to represent the data (Figure 4.3C). The recovery of this sparse representation is, presumably, because the FLASH algorithm uses a mean field variational approximation for the factors, assuming independence over each drift event, whereas in *drift* we allow for correlations in the approximate posterior (see methods section for details). Overall, this example points to some challenges in the non-convexity of the objective function because even though we simulate data from the underlying model that we fit, the algorithm can be sensitive to the initialization.

4.3.3 Empirical applications

Application to the 1000 Genomes Project

Next we applied PCA, ADMIXTURE and *drift* to the 1000 Genomes Project Phase 3 dataset. We downloaded and used the same dataset prepared for admixture analysis in the 1000 Genomes Phase 3 flagship paper (1000 Genomes Project Consortium et al., 2015; Alexander et al., 2009). This prepared dataset filtered out rare variants and thinned the genome to reduce the effect of linkage disequilibrium on the ADMIXTURE results. We then oriented each site so that the genotype data represents the count of chromosomes with the derived allele status for each individual and SNP using ancestral allele calls also inferred from the 1000 Genomes Project. We ran the *drift* model with a greedy initialization scheme for $K = 2, \dots, K = 12$ (Supp. Fig. 4.6, see Wang and Stephens (2018) for a description of the FLASH greedy algorithm). Specifically, in the greedy initialization algorithm we used the bimodal priors that are constrained to be between 0 and 1 for the loadings and Gaussian priors for the factors (drift events). For comparison to the results of ADMIXTURE emphasised in the 1000 Genomes Phase 3 flagship paper, we re-display admixture results for $K = 8$ and *drift* results for $K = 9$ in Figure 4.4. We add an additional factor because we fixed the first factor to be equally shared across all sub-populations and not updated in the algorithm, representing the ancestral mean shared across individuals (1000 Genomes Project Consortium et al., 2015).

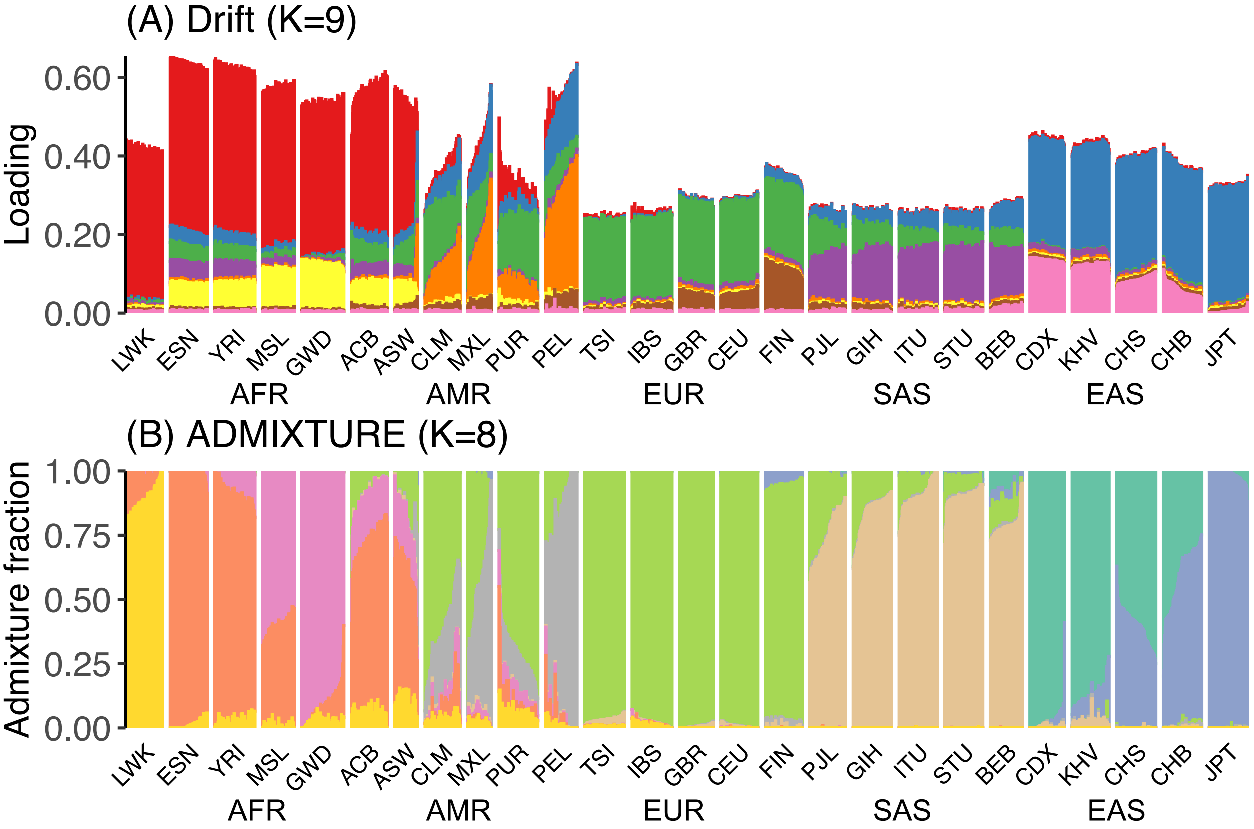


Figure 4.4: **Application of *drift* and ADMIXTURE to the 1000 Genomes Project**: We display the application of *drift* and ADMIXTURE to common variants from the 1000 Genomes Project dataset. In sub-panel A we represent the drift loadings in a stacked bar-char where each color represents a different drift event. Because the loadings are not constrained to sum to 1 the height of each bar can be different across individuals. Note that we scale each loading by the prior variances of their corresponding drift event. In sub-panel B we display a typical STRUCTURE plot of the application of ADMIXTURE to the same dataset. Overall we see that *drift* tends to emphasize more shared factors across individuals whereas ADMIXTURE leads to a more clustered solution with regional specific ancestries being emphasized.

We display the top PCs computed by running PCA on the normalized genotype matrix (Supp. Fig. 4.7). We note that the top PCs mainly capture global relationships among individuals, while subsequent PCs capture regional effects.

As expected from previous studies of global-scale ancestry, the results of ADMIXTURE displayed in a structure plot show a relatively "clustered" representation of population structure (Figure 4.4, Rosenberg et al. (2002)). For instance, individuals from Europe (TSI, IBS, GBR, CEU, FIN) all essentially have 100% of their ancestry proportions assigned to the light-green ancestry component, while individuals from East Asia (CDX, KHV, CHS, CHB, JPT) show two ancestry (dark-blue and dark-green) components that are mostly specific to the region. Individuals from the Americas show a signal of admixture with many ancestry components found in other continental regions as well as a grey ancestry component found in CLM, MXL, PUR and PEL. Finally, African individuals are represented by three ancestry components (red, yellow, and purple). This representation of population structure is not fully satisfying. This is especially apparent for the geographic regions that draw their ancestry almost exclusively from a single latent component (Europe) or from multiple latent components specific to that region (Africa). Only when looking at multiple fits of ADMIXTURE with different values of K can we begin to see nested patterns of shared ancestry (1000 Genomes Project Consortium et al., 2015).

We display the *drift* results in a stacked bar-chart, but because the individual loadings are constrained to be between 0 and 1 and not constrained to lie on a probability simplex like in ADMIXTURE, each bar can have a different height (Figure 4.4). For visual convenience we do not display the first shared factor which is fixed at one for all individuals and we scale the loadings for each individual by the estimated prior variances for their corresponding drift events. Like the results of the PSD model admixed individuals show loading on drift events that are found in many regions over the globe. For instance, African Ancestry in Southwest US (ASW) individuals and African Caribbean in Barbados Beyond (ACB) show loading on the red and yellow drift events primarily found in Africa as well as the blue, green and purple drift events founds outside of Africa. Also, individuals from the Americas (CLM, MXL, PUR and PEL) show signals of shared drift from 3-4 drift events which is consistent with their recent and complex admixture history. Particularly, the fact that the green component is at high levels in the Americans and primarily defines southern European individuals (Toscani in Italia [TSI], Iberian Population in Spain [IBS]) is consistent with this history.

In addition to this admixture signal we see many drift events that are shared across different regions of the globe, even in individuals who are not known to have a recent admixture signal in the results of the PSD model. This sharing of drift events is, indeed, the main point of developing this new representation of population structure. For instance, the green drift event is found in sets of individuals across the globe except in East Asia. The blue drift event is found in most regions of the globe except in Europe (barring Finland), though it is difficult to exactly understand the historical context of this factor and it might be an artifact of where our conception of the root is placed. The brown drift event is enhanced in individuals from Finland, relative to the other northern European individuals from Utah Residents (CEPH) with Northern and Western European Ancestry (CEU) and British in England and Scotland (GBR), where it's also found. This feature could make sense because Finland has gone through a recent bottleneck and is typically considered to be founder population (Martin et al., 2018). In addition to these shared factors there are also mostly regional specific factors such as the orange drift event found in Americas which possibly could represent the ancestral component of drift that occurred during the peopling of the Americas.

Application to inferring spatial population structure in gray wolves

While we use tree-like models as a simple motivator for the *drift* model and describe its ability to emphasize shared evolutionary histories, it has also has the potential to help provide interpretable visualizations of data generated under other types of stochastic processes that are relevant to population genetics. To help motivate this flexibility we applied *drift* to a previously published dataset of gray wolves from North America (Schweizer et al., 2016). This dataset is composed of 111 gray wolf samples, genotyped on an array at $17,729$ SNPs. It is well suited for visualization of spatial population structure because it includes the sample locations (latitude and longitude) for each wolf (note we do not use the spatial locations in the matrix factorization model itself). In Schweizer et al. (2016), the authors showed a

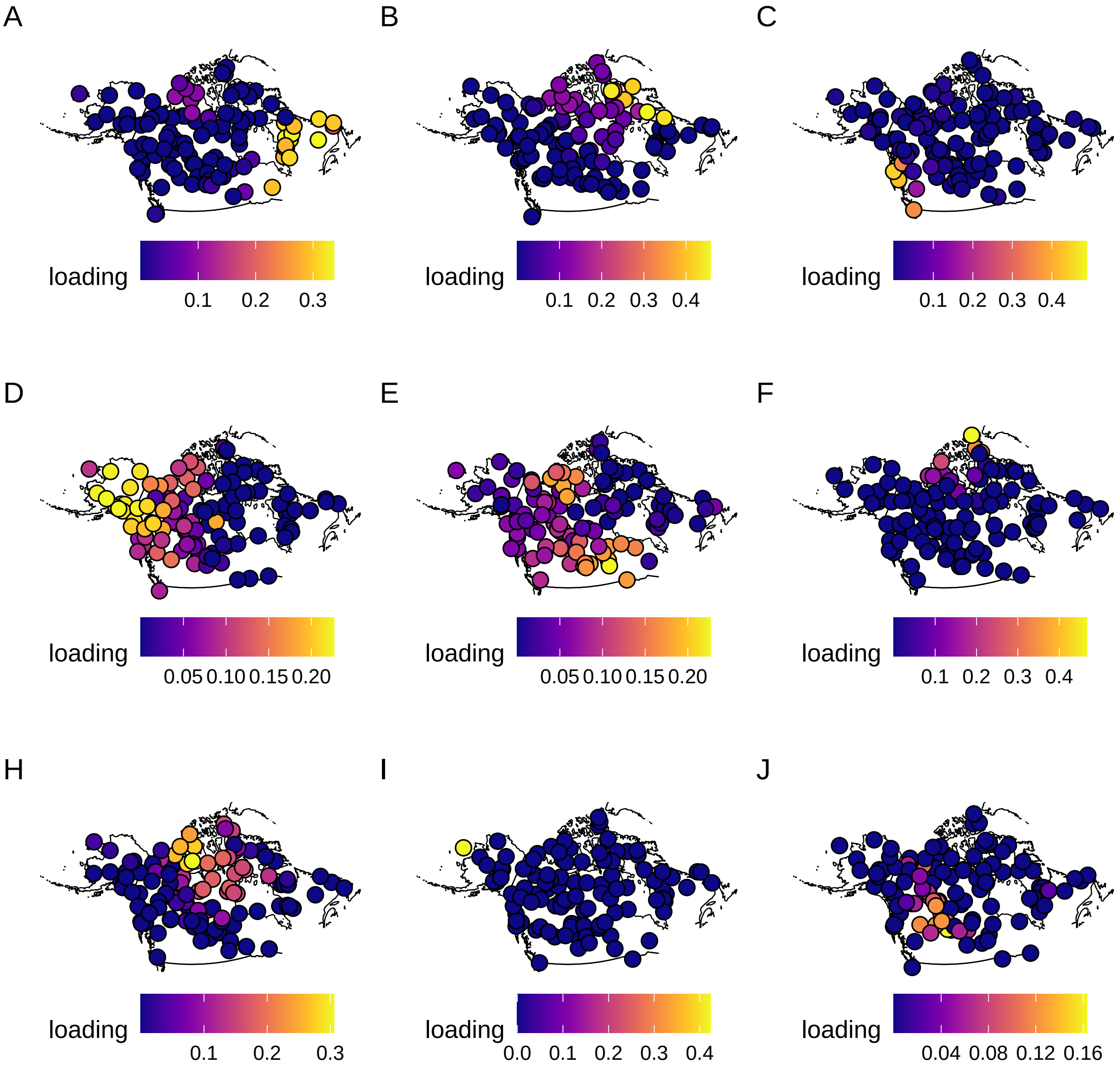


Figure 4.5: **Inferring spatial population structure in gray wolves**: We display the fit of *drift* to a spatial population genetic dataset of gray wolves from North America. Each subpanel shows points for each sample located at the longitude and latitude they were sampled at and colored by the loading on the focal factor represented by that sub-panel. The blue color corresponds to 0 loading on that factor and the more yellow colors correspond to higher loadings. Overall we see each latent factor represents spatial clusters of individuals that correspond well to known bio-geographic features that could plausible have affected wolf migration processes.

convincing signal of isolation-by-distance in this dataset, in which the genetic distance of wolf samples increases as they are more geographically separated. It has been previously shown that non-negative matrix factorization, including admixture models, is particularly useful for visualizing principal spatial patterns and clusters in the data (Wu et al., 2016; Miller et al., 2014).

In Figure 4.5 we display a representation of the *drift* fit. Each panel shows a map for the sampling range of wolves for a given factor, where each point represents a sample colored by its loading on that factor (brighter colors correspond to higher loadings). In general, we see that the factors show spatial clusters of samples throughout the habitat. For example, we observe high loadings for a set of samples east of Hudson Bay in Figure 4.5A, in samples from the Queen Elizabeth Islands in Figure 4.5F, and a quite sparse factor representing a sample from St. Lawrence Island in Figure 4.5I. Indeed, many of these spatial clusters correspond to the delineation of known geographical features in the dataset such as mountain ranges, island chains, other water boundaries, and transition zones between eco-regions. These results are consistent with many of the insights garnered from Schweizer et al. (2016), including those from running ADMIXTURE, but using a semi-non-negative matrix factorization approach that visualizes principal spatial patterns in the data.

4.4 Discussion

Overall, we found *drift* provides a promising new approach to visualizing population structure. The approach proposed here has the potential to emphasize shared evolutionary histories in a matrix factorization framework that has major advantages over existing methods. One could ask the question: why not represent population structure with a simple tree? Trees are a widely studied graphical structures in population genetics, and more flexible extensions like admixture graphs, which allow for admixture events over time, are well-suited for some aspects of this broad goal of representing shared evolutionary histories (Patterson et al., 2012). While tree-like models are of great use for many problems in evolutionary genetics they have several limitations: (1) in most tree-based methods individuals are pre-grouped into sub-populations, which are unknown and reflect human knowledge and bias in the choice of sub-population annotations, (2) trees and admixture graphs have an extremely large state space when the number of sub-populations is large and thus fitting them requires difficult combinatorial optimization problems and heuristics and (3) trees are unrealistic due to their discrete nature i.e. we don't believe real sub-populations evolve in a tree-like fashion and we don't believe individuals experience the evolutionary process on trees in a binary fashion (Yan et al., 2019; Pickrell and Pritchard, 2012). By formulating a model in a matrix factorization framework we hope to be able to capture some of the best features of tree-like models in a continuous representation, describing shared evolutionary histories, while overcoming some of these limitations described above.

ADMIXTURE models, which can be interpreted as another non-negative matrix factorization framework (Engelhardt and Stephens, 2010), are a natural comparison to the proposed method here (Pritchard et al., 2000; Alexander et al., 2009). Both ADMIXTURE and *drift* have the advantage over PCA that the admixture fractions or *drift* loadings can be displayed as a stacked bar chart, allowing the comparison of many factors at once to understand the underlying population structure of the sample. The challenge of ADMIXTURE is that it tends to lead to clustered representations of population structure in which understanding shared evolutionary histories is difficult. Lawson et al. (2018) elaborate a number of other scenarios in which data simulated under tree-like models are difficult to interpret. Particularly, they showed that three different evolutionary scenarios ("recent admixture", "ghost admixture", "recent bottle-neck"), lead to the same fitted admixture proportions. Understanding the behavior of the *drift* model in these types of evolutionary scenarios is an interesting future direction to explore.

While we have shown the *drift* model produces promising results, the proposed approach is not without a number of difficult problems that need to be solved before being ready to be applied to datasets by researchers. In many simulation experiments we have found the *drift* algorithm is highly sensitive to initialization (Marcus et al., 2020a). Indeed, this is expected, to some degree, because maximizing the empirical Bayes matrix factorization ELBO with respect to the variational parameters is a non-convex optimization problem (Wang and Stephens, 2018). In this work we initialized the optimization approach using the greedy algorithm described in (Wang and Stephens, 2018). While we found this initialization scheme to work well in some settings it also seemed to initialize close to a local optimum, allowing the *drift* algorithm to rapidly converge to a solution that gave the same qualitative interpretation as the greedy fit (Figure 4.3). Exploring different initialization strategies that allow the drift algorithm to escape this local optima would an important fruitful next step for this work. We note that initialization and non-convexity are indeed often overlooked problems for other non-negative matrix factorization algorithms such as ADMIXTURE. One unique advantage and fascinating feature of PCA is that the SVD is guaranteed to find the global optima of it's non-convex objective function, i.e. the best rank-K approximation to the data matrix (Murphy, 2012).

Related to the non-convexity of the objective function, a specific challenge in this work is interpreting the results in regards to evolutionary structures such as trees or admixture graphs (Patterson et al., 2012; Pickrell and Pritchard, 2012). In our applications to real data we pre-fixed a single drift event so that it was shared across all individuals i.e. fixed all individuals to have a one loading on this event while still estimating the factors. In tree-like histories this drift event has a natural interpretation of representing the value of the root of the tree (see the drift event one in Figure 4.1). Inferring the root value of stochastic processes on trees is known to be a difficult problem and sometimes this parameter is not identifiable (Pickrell and Pritchard, 2012; Felsenstein, 2004). In our application of the *drift* model to real data, the placement of the root can make interpretation difficult. For example in the 1000 Genomes application, a-priori, it would be natural to imagine that one of the drift events would be present in primarily non-African individuals, representing some notion of the “Out of Africa” bottleneck. Here, we observe that signal being represented as drift event being shared in only African individuals. This might suggest the root is being placed more internal in our conceptualized evolutionary tree and not at the node representing the most ancestral sub-population.

Overall, the *drift* probabilistic matrix factorization model and new variational inference algorithm is a promising approach for visualizing shared evolutionary histories and hence inferring population structure. While we have shown compelling results for multiple population genetic scenarios and species, more work is needed to understand the optimization and properties of the model before it can be used in practice. This work motivates an on-going more basic direction of research in fitting tree-like models to data using matrix factorization approaches, particular when not all samples adhere to the strict assumptions of a tree.

4.4.1 Model description

The *drift* model is a special case of the empirical Bayes matrix factorization (EBMF) model introduced in Wang and Stephens (2018). Like EBMF, *drift* models matrix data as

$$y_{ij} = \sum_{k=1}^{K} \ell_{ik} f_{jk} + e_{ij}, \tag{4.4}$$

with priors

$$\begin{aligned} \ell_{1k}, \ell_{2k}, \ldots, \ell_{nk} &\overset{iid}{\sim} g_k^{(\ell)} \in \mathcal{G}_k^{(\ell)}, \\ f_{1k}, f_{2k}, \ldots, f_{nk} &\overset{iid}{\sim} g_k^{(f)} \in \mathcal{G}_k^{(f)}, \\ \epsilon_{ij} &\overset{iid}{\sim} \mathcal{N}(0, s_{ij}^2) \end{aligned} \tag{4.5}$$

In EBMF, the priors $g_k^{(\ell)}$ and $g_k^{(f)}$ are estimated from among the families of priors $\mathcal{G}_k^{(\ell)}$ and $\mathcal{G}_k^{(f)}$ by maximizing a variational approximation of the log likelihood. Here we assume a more restricted error model where there is a single residual variance parameter $s_{ij}^2 = \sigma_r^2 \; \forall \, i, j$. The residual variance parameter σ_r^2 is likewise estimated via maximum likelihood.

Bi-modal prior family

As mentioned in the text we restrict the prior family on the loadings to be in set of all bi-modal distributions on $[0, 1]$ with modes at 0 and 1. This non-parametric prior constraint can be approximated in practice by using a discrete mixture of uniform distributions,

$$g_k\big(.; \boldsymbol{\pi}_k^{(\ell)}\big) = \sum_m \pi_{km}^{(\ell)} \text{Uniform}(.; a_m, b_m), \tag{4.6}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ are a large prefixed grid of end points of the uniform components that guarantee the mixture distribution will bi-modal with modes at 0 and 1. Specifically, in half of the grid the left end points of the uniform component are fixed to 0 and the right end points get increasingly larger towards 1, whereas in the other half of the grid the right endpoints, are fixed to 1 while the left end point get smaller towards 0. The mixture proportions are unknown parameters to be estimated by maximum likelihood in the variational inference algorithm. This family can be easily implemented using the `ashr` software by simply changing the prefixed grid of the uniform component end points. For more details see the `ebnm_bimodal` analysis in Marcus et al. (2020a).

4.4.2 Variational approximation

Our goal is to infer the posterior distribution of the loadings and factors conditional on the genotype data. Unfortunately, this is analytically intractable so we use a variational approximation to implement a fast optimization algorithm (Blei et al., 2017). We assume the following variational approximation on the posterior distribution of the loadings and factors

$$q_{\boldsymbol{L},\boldsymbol{F}}(\boldsymbol{L},\boldsymbol{F}) = \prod_k q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k) q_{\boldsymbol{F}}(\boldsymbol{F}), \tag{4.7}$$

where $q_{\boldsymbol{\ell}_k(\boldsymbol{\ell}_k)}$ denotes the approximate posterior distribution kth loadings and $q_{\boldsymbol{F}}(\boldsymbol{F})$ denotes the approximate posterior distribution for the factors.

Specifically, the optimal form for the approximate posterior for the factors is a multivariate Gaussian which factorizes over the SNPs,

$$q_{\boldsymbol{f}_j}(\boldsymbol{f}_j) = N\Big(\bar{\boldsymbol{f}}_j, \bar{\boldsymbol{\Sigma}}^{(j)}\Big), \tag{4.8}$$

where $\bar{\boldsymbol{f}}_j$ is the approximate posterior mean and $\bar{\boldsymbol{\Sigma}}^{(j)}$ is the approximate posterior covariance for the jth SNP. Our objective function, the Evidence Lower Bound (ELBO), can be written as:

$$\begin{aligned} F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = {} & \mathbb{E}_{q_{\boldsymbol{L}}, q_{\boldsymbol{F}}}\Big[log\ p(\boldsymbol{Y}|\boldsymbol{L},\boldsymbol{F})\Big] \\ & - \sum_k D_{KL}\left(q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k) \| g_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)\right) \\ & - \sum_j D_{KL}\Big(q_{\boldsymbol{f}_j}(\boldsymbol{f}_j) \| g_{\boldsymbol{f}_j}(\boldsymbol{f}_j)\Big)\,, \\ & + const \end{aligned} \tag{4.9}$$

where the first term is the expected log-likelihood which measures how well the variational approximation reconstructs the observed data, whereas the next two terms are KL-divergences between the approximate posteriors on the loadings and factors and their corresponding priors. These KL terms help to provide regularization for the approximate posteriors to be not too distant from the priors. More specifically,

$$
\begin{aligned}
F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = -\frac{1}{2}\sum_{i,j}\left[\log\left(2\pi s_{ij}^2\right) - \frac{1}{s_{ij}^2}\mathbb{E}_{q_{\boldsymbol{L}},q_{\boldsymbol{F}}}\left(y_{ij} - \sum_k \ell_{ik} f_{jk}\right)^2\right] \\
- \sum_k D_{KL}\left(q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k) \| g_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)\right) \\
- \sum_j D_{KL}\left(N\left(\bar{\boldsymbol{f}}_j, \bar{\boldsymbol{\Sigma}}^{(j)}\right) \| N(\boldsymbol{0}, \boldsymbol{D})\right), \\
+ \, const
\end{aligned}
\tag{4.10}
$$

where $\boldsymbol{D}$ is a $K \times K$ diagonal matrix with diagonal entries $\sigma_1^2, \ldots, \sigma_K^2$, storing the prior variances on the factors. The KL-divergence from $N(\boldsymbol{0}, \boldsymbol{D})$ to $N\left(\bar{\boldsymbol{f}}_j, \bar{\boldsymbol{\Sigma}}^{(j)}\right)$ is:

$$
D_{KL}\left(N\left(\bar{\boldsymbol{f}}_j, \bar{\boldsymbol{\Sigma}}^{(j)}\right) \| N(\boldsymbol{0}, \boldsymbol{D})\right) = \frac{1}{2}\left(\sum_k \frac{\bar{\Sigma}_{kk}^{(j)} + \bar{f}_{jk}^2}{\sigma_k^2} + \sum_k \log \sigma_k^2 - \log\det \bar{\boldsymbol{\Sigma}}^{(j)} - K\right)
\tag{4.11}
$$

The expected log-likelihood can be written as:

$$
\begin{aligned}
E_{q_{\boldsymbol{L}},q_{\boldsymbol{F}}}\left[log\ p(\boldsymbol{Y}|\boldsymbol{L}, \boldsymbol{F})\right] = -\frac{1}{2}\sum_{i,j}\left[\log\left(2\pi s_{ij}^2\right) - \frac{1}{s_{ij}^2}\mathbb{E}_{q_{\boldsymbol{L}},q_{\boldsymbol{F}}}\left(y_{ij} - \sum_k \ell_{ik} f_{jk}\right)^2\right] = \\
- \frac{np}{2}\log(2\pi) - \frac{1}{2}\sum_{i,j}\log s_{ij}^2 - \frac{1}{2}\sum_{i,j}\frac{y_{ij}^2}{s_{ij}^2} + \sum_{i,j,k}\frac{y_{ij}}{s_{ij}^2}\bar{\ell}_{ik}\bar{f}_{jk} \\
- \frac{1}{2}\sum_{i,j,k}\frac{1}{s_{ij}^2}\bar{\ell}^2{}_{ik}\overline{f^2}_{jk} - \frac{1}{2}\sum_{i,j,k,m:k\neq m}\frac{1}{s_{ij}^2}\bar{\ell}_{ik}\bar{\ell}_{im}\overline{f_{jk}f_{jm}}
\end{aligned}
\tag{4.12}
$$

where $\bar{\ell}_{ik}$ and $\bar{\ell}^2{}_{ik}$ denotes the approximate posterior first and second moments on the loaindgs respectively and $\overline{f_{jk}f_{jm}} = E_{q_{\boldsymbol{f}_j}}[f_{jk}f_{jm}]$. Overall our objective is thus,

$$\max_{q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}} F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}), \tag{4.13}$$

which can be solved using an empirical Bayes inspired variational inference algorithm which, in essence, uses coordinate ascent to maximize the ELBO.

4.4.3 Optimization

The EBFA algorithm requires four kinds of updates:

- Updates to the prior and posterior for the kth loading ($g_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)$ and the $q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)s$).
- Updates to the priors on the factors (the σ_k^2s).
- Updates to the posteriors on the factors (the $\bar{\boldsymbol{f}}_j$s and $\bar{\boldsymbol{\Sigma}}^{(j)}$s).
- Updates to the residual variance parameters s_{ij}^2 (which are assumed to be constant here, $s_{ij}^2 = \sigma_r^2$).

Each update is done via coordinate ascents steps on the ELBO.

Update to posteriors on the loadings

The terms in the ELBO that are relevant to the kth loading are:

$$\begin{aligned} F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = \sum_{i,j} \frac{y_{ij}}{s_{ij}^2} \bar{\ell}_{ik} \bar{f}_{jk} - \frac{1}{2} \sum_{i,j} \frac{1}{s_{ij}^2} \bar{\ell^2}_{ik} \bar{f^2}_{jk} - \sum_{i,j,m:m \neq k} \frac{1}{s_{ij}^2} \bar{\ell}_{ik} \bar{\ell}_{im} \overline{f_{jk} f_{jm}} \\ - D_{KL}\left(q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k) \| g_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)\right) + const \end{aligned} \tag{4.14}$$

Using the fact that $\overline{f_{jk} f_{jm}} = \bar{f}_{jk} \bar{f}_{jm} + \bar{\Sigma}_{km}^{(j)}$, we can rewrite the ELBO as:

$$
\begin{aligned}
&F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = \\
&-\frac{1}{2}\sum_i \left[\left(\sum_j \frac{\overline{f^2_{jk}}}{s^2_{ij}} \right) \overline{\ell^2_{ik}} - 2 \left(\sum_j \left[\frac{1}{s^2_{ij}} \bar{f}_{jk} \left(y_{ij} - \sum_{m:m \neq k} \bar{\ell}_{im} \bar{f}_{jm} \right) - \frac{1}{s^2_{ij}} \sum_{m:m \neq k} \bar{\ell}_{im} \bar{\Sigma}^{(j)}_{km} \right] \right) \bar{\ell}_{ik} \right] \\
&- D_{KL} \left(q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k) \| g_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k) \right) + const
\end{aligned}
\tag{4.15}
$$

Now we can use Lemma 2 from Wang and Stephens (2018) to conclude that $g_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)$ and $q_{\boldsymbol{\ell}_k}(\boldsymbol{\ell}_k)$ can simultaneously be optimized by solving an EBNM problem ($x_i \sim N(\theta_i, 1/\tau_i), \theta_i \sim g_k \in \mathcal{G}_k$) with

$$
\tau_i = \sum_j \frac{\overline{f^2_{jk}}}{s^2_{ij}}, \tag{4.16}
$$

and

$$
\tau_i x_i = \sum_j \left[\frac{1}{s^2_{ij}} \bar{f}_{jk} \left(y_{ij} - \sum_{m:m \neq k} \bar{\ell}_{im} \bar{f}_{jm} \right) - \frac{1}{s^2_{ij}} \sum_{m:m \neq k} \bar{\ell}_{im} \bar{\Sigma}^{(j)}_{km} \right]. \tag{4.17}
$$

Notice that these updates are similar to the EBMF updates given in Wang and Stephens (2018), but that there is an extra term in the expression for $\tau_i x_i$ that comes from no longer assuming that the posteriors on factors are independent.

Updates to priors on the factors

These updates are easy, since only KL-divergence terms include the σ_k^2s. Ignoring the common factor of $\frac{1}{2}$, the problem is to maximize:

$$
D_{KL} \left(N\left(\bar{\boldsymbol{f}}_j, \bar{\boldsymbol{\Sigma}}^{(j)} \right) \| N(\boldsymbol{0}, \boldsymbol{D}) \right) = \sum_{j,k} \frac{\bar{\Sigma}^{(j)}_{kk} + \overline{f^2_{jk}}}{\sigma_k^2} - p \sum_k \log \frac{1}{\sigma_k^2} + const \tag{4.18}
$$

Set the derivative with respect to $\frac{1}{\sigma_k^2}$ equal to zero and solve to get:

$$\sigma_k^2 = \frac{1}{p} \sum_j \left(\bar{\Sigma}^{(j)}_{kk} + \bar{f}^2_{jk} \right) \tag{4.19}$$

In other words, σ_k^2 is just the mean value of $\overline{f^2}_{jk}$ across all j.

Updates to the posteriors on the factors

For each j, all K factors must be updated at once. Multiplying the ELBO by two and retaining terms that include $\bar{\boldsymbol{f}}_j$ or $\bar{\boldsymbol{\Sigma}}^{(j)}$ yields:

$$\begin{aligned} F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = {} & 2 \sum_{i,k} \frac{y_{ij}}{s_{ij}^2} \bar{\ell}_{ik} \bar{f}_{jk} - \sum_{i,k} \frac{1}{s_{ij}^2} \overline{\ell^2}_{ik} \left(\bar{f}^2_{jk} + \bar{\Sigma}^{(j)}_{kk} \right) \\ & - \sum_{i,k,m:k \neq m} \frac{1}{s_{ij}^2} \bar{\ell}_{ik} \bar{\ell}_{im} \left(\bar{f}_{jk} \bar{f}_{jm} + \bar{\Sigma}^{(j)}_{km} \right) \\ & - \sum_k \frac{\bar{\Sigma}^{(j)}_{kk} + \bar{f}^2_{jk}}{\sigma_k^2} + \log \det \bar{\boldsymbol{\Sigma}}^{(j)} + \mathit{const} \end{aligned} \tag{4.20}$$

Further restricting to terms that include $\bar{\boldsymbol{\Sigma}}^{(j)}$ yields:

$$\begin{aligned} F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = {} & \sum_k \left(\bar{\Sigma}^{(j)}_{kk} \sum_i \frac{\overline{\ell^2}_{ik}}{s_{ij}^2} + \sum_{m:m \neq k} \bar{\Sigma}^{(j)}_{km} \sum_i \bar{\ell}_{ik} \bar{\ell}_{im}) s_{ij}^2 \right) \\ & - \operatorname{tr} \left(\boldsymbol{D}^{-1} \bar{\boldsymbol{\Sigma}}^{(j)} \right) + \log \det \bar{\boldsymbol{\Sigma}}^{(j)} + \mathit{const} \end{aligned} \tag{4.21}$$

Now let $\tilde{\boldsymbol{L}}^{(j)}$ be the matrix obtained by scaling each column of $\boldsymbol{L}$ by the n-vector s_{ij} (i.e., $\tilde{\ell}^{(j)}_{ik} = \frac{\ell_{ik}}{\sqrt{s_{ij}^2}}$). Then the above may be rewritten:

$$F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = -\operatorname{tr} \left(\left(\overline{\tilde{\boldsymbol{L}}^{(j)T} \tilde{\boldsymbol{L}}^{(j)}} + \boldsymbol{D}^{-1} \right) \bar{\boldsymbol{\Sigma}}^{(j)} \right) + \log \det \bar{\boldsymbol{\Sigma}}^{(j)} + \mathit{const}, \tag{4.22}$$

where $\overline{\tilde{\boldsymbol{L}}^{(j)T} \tilde{\boldsymbol{L}}^{(j)}} = \mathbb{E}_{q_{\boldsymbol{L}}}[\tilde{\boldsymbol{L}}^{(j)T} \tilde{\boldsymbol{L}}^{(j)}]$. Setting the matrix derivative equal to zero yields the solution:

$$\bar{\boldsymbol{\Sigma}}^{(j)} = \left(\overline{\tilde{\boldsymbol{L}}^{(j)T} \tilde{\boldsymbol{L}}^{(j)}} + \boldsymbol{D}^{-1} \right)^{-1} \tag{4.23}$$

Next, we use similar notation to set $\tilde{y}_{ij} = \frac{y_{ij}}{s_{ij}}$ and restrict to terms that include $\bar{\boldsymbol{f}}_j$. Using vector notation, we write:

$$F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = 2\tilde{\boldsymbol{y}}_j^T \bar{\tilde{\boldsymbol{L}}}^{(j)} \bar{\boldsymbol{f}}_j - \bar{\boldsymbol{f}}_j^T \overline{\tilde{\boldsymbol{L}}^{(j)T} \tilde{\boldsymbol{L}}^{(j)}} \bar{\boldsymbol{f}}_j - \bar{\boldsymbol{f}}_j^T \boldsymbol{D}^{-1} \bar{\boldsymbol{f}}_j + const \tag{4.24}$$

Again we set the derivative equal to zero and solve:

$$\bar{\boldsymbol{f}}_j = \left(\overline{\tilde{\boldsymbol{L}}^{(j)T} \tilde{\boldsymbol{L}}^{(j)}} + \boldsymbol{D}^{-1} \right)^{-1} \bar{\tilde{\boldsymbol{L}}}^{(j)T} \tilde{\boldsymbol{y}}_j = \left(\bar{\tilde{\boldsymbol{L}}}^{(j)} \bar{\boldsymbol{\Sigma}}^{(j)} \right)^T \tilde{\boldsymbol{y}}_j \tag{4.25}$$

Note the similarity of these equations to the solutions for the ridge regression problem: this is of course not coincidental, as the problem is basically the same except that we have to keep track of where we are taking expectations.

Updates to residual variance parameters

The residual variance parameters only enter into the data log likelihood. Reorganizing the ELBO thus yields:

$$F(q_{\boldsymbol{L}}, q_{\boldsymbol{F}}, g_{\boldsymbol{L}}, g_{\boldsymbol{F}}, \boldsymbol{S}; \boldsymbol{Y}) = \\ -\frac{1}{2} \sum_{i,j} \left[\log s_{ij}^2 + \frac{1}{s_{ij}^2} \left(y_{ij}^2 - 2y_{ij} \sum_k \bar{\ell}_{ik} \bar{f}_{jk} + \sum_k \bar{\ell^2}_{ik} \bar{f}_{jk}^2 + \sum_{k,m:k \neq m} \bar{\ell}_{ik} \bar{\ell}_{im} \overline{f_{jk} f_{jm}} \right) \right] + const \tag{4.26}$$

The solution will depend on the variance structure used. For example, if all residual variance parameters are assumed to be equal ($s_{ij}^2 \equiv s^2$), then:

$$s^2 = \frac{1}{np}\sum_{i,j}\left(y_{ij}^2 - 2y_{ij}\sum_k \bar{\ell}_{ik}\bar{f}_{jk} + \sum_k \bar{\ell^2_{ik}}\bar{f^2}_{jk} + \sum_{k,m:k\neq m} \bar{\ell}_{ik})\bar{\ell}_{im}\overline{f_{jk}f_{jm}}\right) \tag{4.27}$$

Doing some rearranging:

$$\begin{aligned} s^2 &= \frac{1}{np}\sum_{i,j}\left[\left(y_{ij} - \sum_k \bar{\ell}_{ik}\bar{f}_{jk}\right)^2 - \left(\sum_k \bar{\ell}_{ik}\bar{f}_{jk}\right)^2 + \sum_k \left(\bar{\ell^2}_{ik} - \bar{\ell}_{ik})^2\right)\bar{f^2}_{jk} + \sum_{k,m} \bar{\ell}_{ik}\bar{\ell}_{im}\overline{f_{jk}f_{jm}}\right] \\ &= \frac{1}{np}\sum_{i,j}\left[\left(y_{ij} - \sum_k \bar{\ell}_{ik}\bar{f}_{jk}\right)^2 + \sum_k \mathrm{Var}_{q_{\ell_{ik}}}(\ell_{ik})\bar{f^2}_{jk} + \sum_{k,m} \bar{\ell}_{ik}\bar{\ell}_{im}\mathrm{Cov}_{q_{\boldsymbol{f}_j}}(f_{jk}, f_{jm})\right] \end{aligned} \tag{4.28}$$

As with EBMF, the estimated residual variance is equal to the mean squared residual (where the mean is across all entries, or row-wise, or column-wise, depending on the variance structure), but there is an extra term that again comes from not assuming that posteriors on factors are independent.

4.4.4 Population structure inference

To compare the application of the *drift* model to existing population structure inference methods we ran multiple methods on the genotype data we analyzed here. We ran the ADMIXTURE software for multiple values of K visualizing among multiple replicates the one that achieves the high log-likelihood. We use the truncated SVD algorithm implemented in the LFA `R` package for all the principal components analysis visualizations displayed (Hao et al., 2016). Finally we used the ALStructure `R` package to run an implementation of admixture models on the simulated datasets (Cabreros and Storey, 2019).

4.5 Supplementary Information

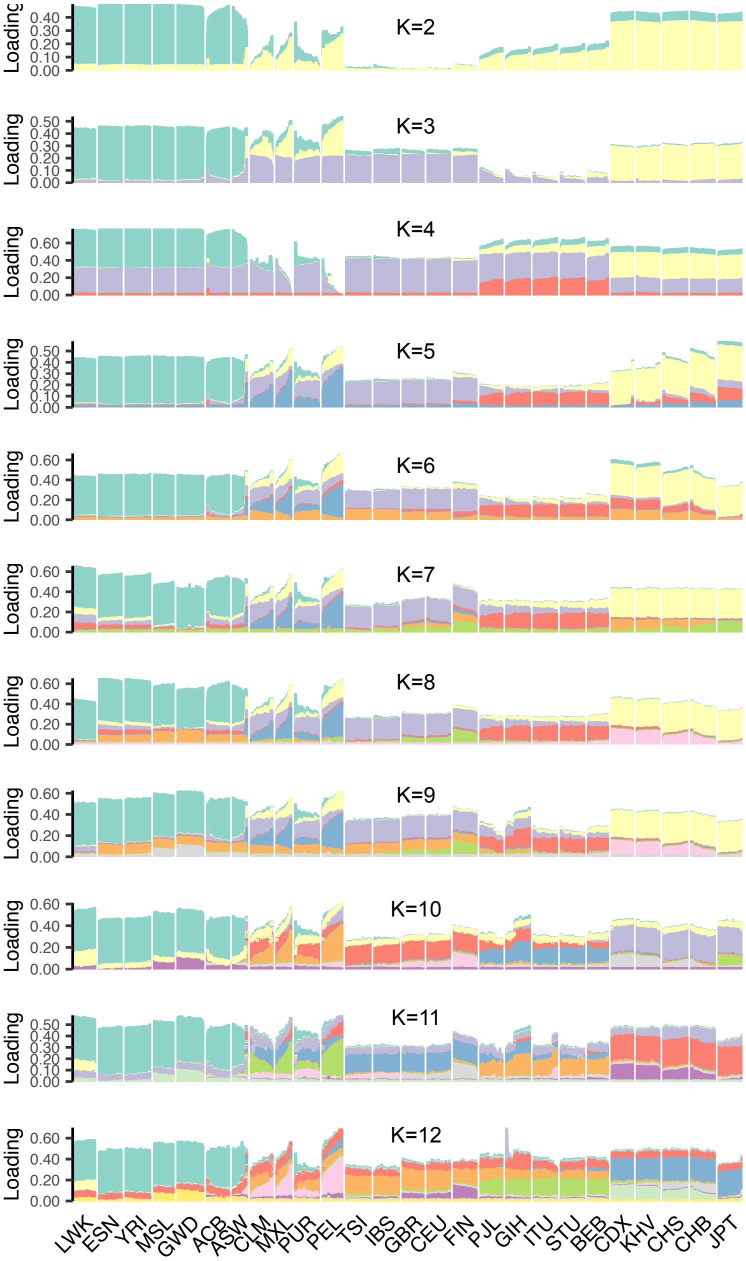


Figure 4.6: **Gallery plot of *drift* results from $K = 2$ to $K = 12$:** We display a gallery plot of *drift* results applied to the 1000 Genomes Project Phase 3 dataset.

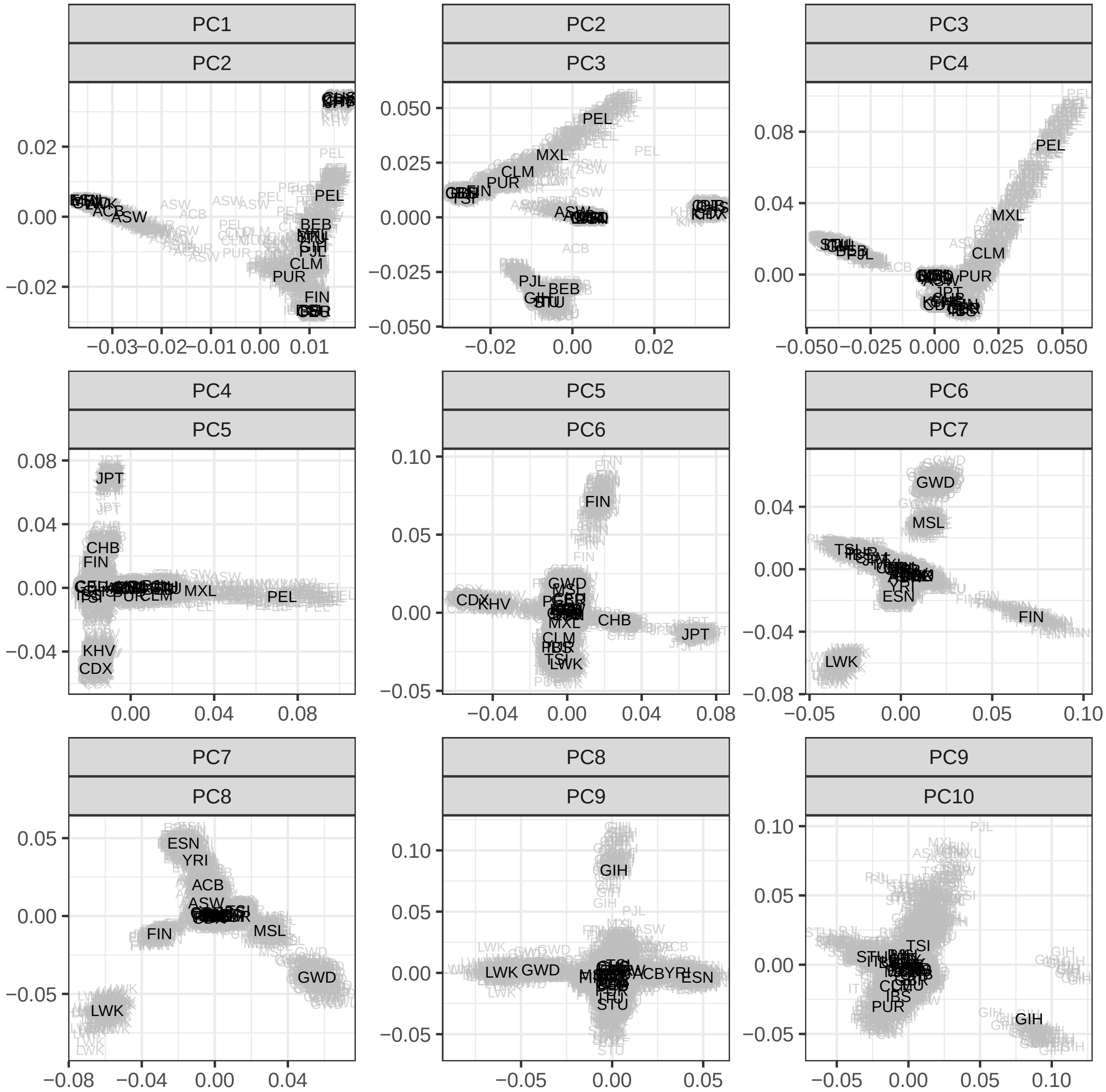


Figure 4.7: **PCA applied to the 1000 Genomes Project Dataset**: We display PCA biplots applied to the normalized 1000 Genomes Project Phase 3 genotype matrix of common variants.

CHAPTER 5

CONCLUSION

5.1 Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia

In chapter two titled "*Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia*", I, along with my collaborators, applied existing population structure inference methods to learn about the genetic history of the Mediterranean population of Sardinia (Marcus et al., 2020c). In this chapter we generated a new genome-wide ancient DNA dataset of 70 individuals from a broad temporal range and geographic locale across the island of Sardinia. The work builds upon the observations of Chiang et al. (2018) but with a greater resolution into the past because of the temporal component to the data. This unique dataset offered a new lens into the genetic history of Sardinia and understanding of migration processes in the broader Mediterranean region as well as Western Eurasia. The application of population structure inference tools allowed us to learn that ancestry in Sardinia was relatively stable for thousands of years, from the Neolithic to the Nuragic period, a period named after an unique Sardinian culture known for its distinctive stone-towers scattered throughout the island. In post-Nuragic times migration throughout the Mediterranean was much more dynamic leading to changes in ancestry from sources from the eastern and northern Mediterranean and northern Africa. Overall the utility of population structure inference methods are highlighted in an applied dataset, shedding light into the genetic history of Sardinia.

5.2 Fast and Flexible Estimation of Effective Migration Surfaces

In chapter three titled "*Fast and Flexible Estimation of Effective Migration Surfaces*", I, along with my collaborators, developed a new statistical model and optimization algorithm called FEEMS for inferring effective migration and visualizing non-homogeneous spatial population structure (Marcus et al., 2020b). As the name suggest, we extend and draw inspiration from the EEMS method, using a similar spatial graph based approach but propose two key differences for greater flexibility and speed; (1) We assign unique parameters to each edge-weight of the graph rather than each node and constraining each edge to be the average of node parameters. This parameterization is more flexible and allows the possibility of inferring anisotropic migration histories which we highlight in coalescent simulations. (2) Instead of MCMC, we develop a gradient based optimization algorithm. We develop a new penalty which encourages neighboring edges to be smooth over the graph. Specifically, we use a fast quasi-Newton algorithm to optimize a challenging non-convex program. We apply FEEMS to a dataset of gray wolves from North America and find FEEMS provides similar results to EEMS but orders of magnitude faster. We hope for FEEMS to be a helpful addition to the toolkit of spatial population structure inference methods.

5.3 Emphasizing shared evolutionary histories when inferring representations of population structure

Finally, in chapter four titled "*Emphasizing shared evolutionary histories when inferring representations of population structure*", I, along with my collaborators, developed a new Bayesian matrix factorization method for visualizing population structure. PCA and admixture models are the most widely used tools for inferring low-rank representations of population structure but both of these methods have difficulties when applied to population genetic data. Particularly, the output of admixture models tends to lead to clustered representations of population structure because of the model constraints on the admixture proportions. In this chapter we developed a new matrix factorization model called *drift* which emphasises shared evolutionary histories that individuals have experienced over time. We extend the work (Wang and Stephens, 2018) and developed a fast empirical Bayes inspired variational inference algorithm which returns approximate posterior distributions on the matrix factors using optimization. We applied *drift* to multiple species and find it has advantageous over existing approaches but also unique challenges in the optimization.

5.4 Future Directions

A fundamental focus of this dissertation can be described as "unsupervised learning", with a particular emphasis on inferring interpretable representations that are tailored for population genetic data. The output of unsupervised learning methods are often used for prediction or imputation tasks (Mazumder et al., 2010) but directly interpreting the underlying learned representation is an important and challenging problem in many fields (Friedman et al., 2001).

In chapter 2, we used existing approaches to population structure inference applied to ancient and modern genetic variation data (Marcus et al., 2020c). One major challenge was that the ancient data was extremely noisy and had high levels of missingness due to the inherit sample input and capture technologies used to generate the data (Orlando et al., 2015). In our work, as is the standard in the field, we fixed the genotypes to pseudo-haploid gentoype calls to make the application of standard population structure inference tools practically easier (Haak et al., 2015). We also leveraged the high quality genotypes of modern samples to help, in some-sense, regularize the estimates of ancestries for the ancient samples. We approached this by projecting ancient samples on to a sub-space defined by modern samples, while correcting for an out-of-sample bias induced when performing PC score regression (Lee et al., 2010). While this approach does not identify the population structure specific to the ancient samples it can still help to provide an interpretable representation of ancestry with respect to the high quality modern genotype data.

Another reasonable approach would be to model the raw data for ancient samples, which are read-counts, not pre-fixed pseudo-haploid genotype calls. The challenge then becomes integrating over the uncertainty of the latent genotypes for each ancient sample and SNP while estimating population structure along with high quality modern genotype data. There are some promising initial approaches in different population genetic applications, such as in admixture models (Skotte et al., 2013) or inferring runs of homozygosity (Ringbauer et al., 2020), but developing scalable and simple population structure inference methods that can account for heterogeneous data inputs of reads and genotypes is in general an open and important problem. A fast and simple approach to this problem would provide exploratory tools that would allow for inference of population structure within ancient samples which would be of great use for many applications.

In chapter 3, we developed a new model and efficient optimization algorithm for inferring a graph embedded in geographic space. The edge-weights of the graph represent effective migration and the nodes represent allele frequencies. In the chapter we discussed many future directions for the FEEMS model specifically, but an exciting future direction for the field would be to combine the smoothness assumptions of graph learning as well as low-rank assumptions in matrix factorization. Caye et al. (2018) took a related approach to this idea in admixture models by constructing a graph based on a similarity function taking spatial coordinates as input. They then used a graph laplacian smoothing penalty on the admixture fractions to encourage them to be smooth over geographic space. This approach allowed for better interpretation and predictive accuracy of held-out genotypes for admixture models in a spatial context. One potential limitation of the approach is that the spatial regularization is assumed to be globally smooth over geographic space and did not allow for sharp-change points which allows for local-adaptivity. It would be interesting to explore spatially smooth matrix factorization with trend filtering penalties that allow for sharp changes in ancestries over space, representing geographic regions that reduce or enhance gene-flow. It could be also useful to impute these ancestries over the entire geographic region (in the method itself), not just inferring ancestries at sampled locations. This type of approach could lead to interpretable spatial maps of ancestries and result in a simple alternate matrix factorization based approach to the visualisations output by FEEMS. Smooth matrix factorizaton could have some other interesting applications to complex biological datasets, like those generated in spatial transcriptomics (Rodriques et al., 2019).

Finally in chapter 4, we developed a new Bayesian matrix factorization model and variational inference algorithm inspired by Wang and Stephens (2018) for visualizing population structure. Working on this model and algorithm opened open a number of fundamental questions on fitting trees to data and their relationship to matrix factorization or, relatedly, estimating tree-structured covariance matrices (McCullagh, 2009). While fitting trees to data is a common task, often performed with hierarchical clustering (Friedman et al., 2001), an interesting question to follow up on is how to fit trees to data that are generated from "tree-like" models. For example a dataset with admixed individuals is "tree-like" yet it is challenging to infer an underlying tree including admixed individuals in the model (Patterson et al., 2012). More importantly, in practice, which individuals are admixed is not known a-priori. Extracting trees from "tree-like" data is a ubiquitous problem in many fields and yet to our knowledge it has not been thoroughly explored. These fundamental questions is fodder for interesting research directions that could have wide applications in genomic data, including population genetics and inference of tree structures from gene expression data.

Looking forward, many challenges and open questions remain in all these chapters but in general they contribute to a body of knowledge that can be applied and built upon. I am excited to seeing how this work is extended and contributes to scientific knowledge on these challenging inference problems.

References

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

A. Agrawal, A. M. Chiu, M. Le, E. Halperin, and S. Sankararaman. Scalable probabilistic pca for large-scale genetic variation data. *PLoS Genetics*, 16(5):e1008773, 2020.

H. Al-Asadi, K. K. Dey, J. Novembre, and M. Stephens. Inference and visualization of DNA damage patterns using a grade of membership model. *Bioinformatics*, 35(8):1292–1298, 2018.

H. Al-Asadi, D. Petkova, M. Stephens, and J. Novembre. Estimating recent migration and population-size surfaces. *PLoS Genetics*, 15(1):e1007908, 2019.

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

M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, et al. Population genomics of Bronze Age Eurasia. *Nature*, 522(7555):167–172, 2015.

R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, and N. Howell. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics*, 23(2):147, 1999.

M. L. Antonio, Z. Gao, H. M. Moots, M. Lucci, F. Candilio, S. Sawyer, V. Oberreiter, D. Calderon, K. Devitofranceschi, R. C. Aikens, S. Aneli, F. Bartoli, A. Bedini, O. Cheronet, D. J. Cotter, D. M. Fernandes, G. Gasperetti, R. Grifoni, A. Guidi, F. La Pastina, E. Loreti, D. Manacorda, G. Matullo, S. Morretta, A. Nava, V. Fiocchi Nicolai, F. Nomi, C. Pavolini, M. Pentiricci, P. Pergola, M. Piranomonte, R. Schmidt, G. Spinola, A. Sperduti, M. Rubini, L. Bondioli, A. Coppa, R. Pinhasi, and J. K. Pritchard. Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science*, 366(6466): 708–714, 2019. ISSN 0036-8075. doi: 10.1126/science.aay6826.

W. K. Barnett. Cardial pottery and the agricultural transition in Mediterranean Europe. *Europe's First Farmers*, pages 93–116, 2000.

C. Battey, P. L. Ralph, and A. D. Kern. Space is the place: Effects of continuous spatial structure on analysis of population genetic data. *Genetics*, 215(1):193–214, 2020.

P. Beerli and J. Felsenstein. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8):4563–4568, 2001.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

W. F. Bodmer and L. L. Cavalli-Sforza. A migration matrix model for the study of random genetic drift. *Genetics*, 59(4):565, 1968.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011.

G. S. Bradburd and P. L. Ralph. Spatial population genetics: It's about time. *Annual Review of Ecology, Evolution, and Systematics*, 50:427–449, 2019.

G. S. Bradburd, P. L. Ralph, and G. M. Coop. A spatial framework for understanding population structure and admixture. *PLoS Genetics*, 12(1), 2016.

G. S. Bradburd, G. M. Coop, and P. L. Ralph. Inferring continuous and discrete population genetic structure across space. *Genetics*, 210(1):33–52, 2018.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

I. Cabreros and J. D. Storey. A likelihood-free estimator of population structure bridging admixture models and principal components analysis. *Genetics*, 212(4):1009–1029, 2019.

C. Calò, A. Melis, G. Vona, and I. Piras. Review Synthetic Article: Sardinian population (italy): A Genetic Review. *International Journal of Modern Anthropology*, 1(1):39–64, 2008.

L. L. Cavalli-Sforza. The human genome diversity project: past, present and future. *Nature Reviews Genetics*, 6(4):333, 2005.

K. Caye, F. Jay, O. Michel, O. François, et al. Fast inference of individual admixture coefficients using geographic data. *The Annals of Applied Statistics*, 12(1):586–608, 2018.

A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari. The electrical resistance of a graph captures its commute and cover times. *Computational Complexity*, 6(4):312–340, 1996.

C. W. Chiang, J. H. Marcus, C. Sidore, A. Biddanda, H. Al-Asadi, M. Zoledziewska, M. Pitzalis, F. Busonero, A. Maschio, G. Pistis, et al. Genomic history of the Sardinian population. *Nature Genetics*, page 1, 2018.

L. Contu, M. Arras, C. Carcassi, G. L. Nasa, and M. Mulargia. HLA structure of the Sardinian population: a haplotype study of 551 families. *Tissue Antigens*, 40(4):165–174, 1992.

N. Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, et al. Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*, page 201314445, 2013.

C. H. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.

T. Dobzhansky and S. Wright. Genetics of natural populations. x. dispersion rates in drosophila pseudoobscura. *Genetics*, 28(4):304, 1943.

X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst. Learning laplacian matrix in smooth graph signal representations. *IEEE Transactions on Signal Processing*, 64(23): 6160–6173, 2016.

X. Dong, D. Thanou, M. Rabbat, and P. Frossard. Learning graphs from data: A signal representation perspective. *IEEE Signal Processing Magazine*, 36(3):44–63, 2019.

N. Duforet-Frebourg and M. G. Blum. Nonstationary patterns of isolation-by-distance: inferring measures of local genetic differentiation with bayesian kriging. *Evolution*, 68(4): 1110–1123, 2014.

S. Dyson and R. Rowland. *Archaeology and History in Sardinia from the Stone Age to the Middle Ages. Shepherds, Sailors, and Conquerors.* University of Pennsylvania Museum Press, 2007.

H. E. Egilmez, E. Pavez, and A. Ortega. Graph learning from data under structural and laplacian constraints. *arXiv preprint arXiv:1611.05181*, 2016.

B. E. Engelhardt and M. Stephens. Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics*, 6(9):e1001117, 2010.

M. Feldman, D. M. Master, R. A. Bianco, M. Burri, P. W. Stockhammer, A. Mittnik, A. J. Aja, C. Jeong, and J. Krause. Ancient DNA sheds light on the genetic origins of early Iron Age Philistines. *Science Advances*, 5(7), 2019.

J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American journal of human genetics*, 25(5):471, 1973.

J. Felsenstein. How can we infer geography and history from gene frequencies? *Journal of Theoretical Biology*, 96(1):9–20, 1982.

J. Felsenstein. *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA, 2004.

D. M. Fernandes, A. Mittnik, I. Olalde, I. Lazaridis, O. Cheronet, N. Rohland, S. Mallick, R. Bernardos, N. Broomandkhoshbacht, J. Carlsson, et al. The arrival of Steppe and Iranian related ancestry in the islands of the Western Mediterranean. *bioRxiv*, page 584714, 2019.

P. Francalacci, L. Morelli, A. Angius, R. Berutti, F. Reinier, R. Atzeni, R. Pilu, F. Busonero, A. Maschio, I. Zara, et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science*, 341(6145):565–569, 2013.

O. François, S. Ancelet, and G. Guillot. Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2):805–816, 2006.

O. François, S. Liégeois, B. Demaille, and F. Jay. Inference of population genetic structure from temporal samples of dna. 2019.

E. Frichot, S. D. Schoville, G. Bouchard, and O. François. Correcting principal component maps for effects of spatial autocorrelation in population genetic data. *Frontiers in Genetics*, 3:254, 2012.

E. Frichot, F. Mathieu, T. Trouillon, G. Bouchard, and O. François. Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4):973–983, 2014.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, 524(7564):216, 2015.

C. Gamba, E. R. Jones, M. D. Teasdale, R. L. McLaughlin, G. Gonzalez-Fortes, V. Mattiangeli, L. Domboróczki, I. Kővári, I. Pap, A. Anders, et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5:5257, 2014.

F. Germanà. Crani della seconda età del Bronzo da S'isterridolzu (ossi, sassari) nel contesto umano paleosardo recente (Antropologia e Paleontologia). In *Atti del XX Congresso Internazionale d'Antropologia e d'Archeologia Preistorica, Cagliari, Poligraf, Aprilia*, pages 377–394. Università di Cagliari, 1980.

S. Ghirotto, S. Mona, A. Benazzo, F. Paparazzo, D. Caramelli, and G. Barbujani. Inferring genealogical processes from patterns of Bronze-Age and modern DNA variation in Sardinia. *Molecular Biology and Evolution*, 27(4):875–886, 2009.

A. Ginolhac, M. Rasmussen, M. T. P. Gilbert, E. Willerslev, and L. Orlando. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, 27(15):2153–2155, 2011.

P. Gopalan, W. Hao, D. M. Blei, and J. D. Storey. Scaling probabilistic models of genetic variation to millions of humans. *Nature Genetics*, 48(12):1587, 2016.

M. Guirguis, C. Murgia, and R. Pla Orquín. Archeoantropologia e bioarcheologia nella necropoli di Monte Sirai (Carbonia-Italia). risultati delle analisi su alcuni contesti della prima età Punica (fine VI-inizi IV sec. a.c.). In F. Serra, editor, *From the Mediterranean to the Atlantic: People, Goods and Ideas between East and West. Proceedings of the 8th International Congress of Phoenician and Punic Studies (Italy, Sardinia-Carbonia, Sant'Antioco, 21-26 October 2013). (Folia Phoenicia, 1).*, pages 282–299, Pisa-Roma, 2017. Fabrizio Serra Editore.

T. Günther, C. Valdiosera, H. Malmström, I. Ureña, R. Rodriguez-Varela, Ó. O. Sverrisdóttir, E. A. Daskalaki, P. Skoglund, T. Naidoo, E. M. Svensson, et al. Ancient genomes link early farmers from atapuerca in spain to modern-day basques. *Proceedings of the National Academy of Sciences*, 112(38):11917–11922, 2015.

W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015.

E. M. Hanks. A constructive spatio-temporal approach to modeling spatial covariance. *arXiv preprint arXiv:1506.03824*, 2015.

E. M. Hanks and M. B. Hooten. Circuit theory and model-based inference for landscape connectivity. *Journal of the American Statistical Association*, 108(501):22–33, 2013.

W. Hao and J. D. Storey. Extending tests of hardy–weinberg equilibrium to structured populations. *Genetics*, 213(3):759–770, 2019.

W. Hao, M. Song, and J. D. Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, 2016.

G. Hellenthal, G. B. Busby, G. Band, J. F. Wilson, C. Capelli, D. Falush, and S. Myers. A genetic atlas of human admixture history. *Science*, 343(6172):747–751, 2014.

Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Díez-del Molino, L. van Dorp, S. López, A. Kousathanas, V. Link, et al. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, 113 (25):6886–6891, 2016.

R. R. Hudson et al. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

H. Jónsson, A. Ginolhac, M. Schubert, P. L. Johnson, and L. Orlando. mapdamage2. 0: fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics*, 29 (13):1682–1684, 2013.

T. A. Joseph and I. Pe'er. Inference of population structure from time-series genotype data. *The American Journal of Human Genetics*, 2019a.

T. A. Joseph and I. Pe'er. Inference of population structure from time-series genotype data. *The American Journal of Human Genetics*, 105(2):317–333, 2019b.

V. Kalofolias. How to learn a graph from smooth signals. In *Artificial Intelligence and Statistics*, pages 920–929, 2016.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5), 2016.

A. Keller, A. Graefen, M. Ball, M. Matzas, V. Boisguerin, F. Maixner, P. Leidinger, C. Backes, R. Khairat, M. Forster, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nature Communications*, 3:698, 2012.

M. Kimura. Stepping stone model of population. *Annual Report of the National Institute of Genetics Japan*, 3:62–63, 1953.

M. Kimura and G. H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561, 1964.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT Press, 2009.

T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, 15(1):356, 2014.

M. Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of drosophila melanogaster. *Nature*, 304(5925):412–417, 1983.

A. La Fragola and D. Rovina. Il cimitero romano di monte carru (alghero) e la statio di carbia. *Sardinia, Corsica et Baleares antiquae*, 16(16):59–79, 2018.

L. Lai, R. H. Tykot, E. Usai, J. F. Beckett, R. Floris, O. Fonzo, E. Goddard, D. Hollander, M. R. Manunza, and A. Usai. Diet in the Sardinian Bronze Age: models, collagen isotopic data, issues and perspectives. *Préhistoires Méditerranéennes*, 2013.

R. Lampis, L. Morelli, S. De Virgiliis, M. Congia, and F. Cucca. The distribution of HLA class II haplotypes reveals that the Sardinian population is genetically differentiated from the other Caucasian populations. *Tissue Antigens*, 56(6):515–521, 2000.

S. L. Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

D. J. Lawson, G. Hellenthal, S. Myers, and D. Falush. Inference of population structure using dense haplotype data. *PLoS Genet*, 8(1):e1002453, 2012.

D. J. Lawson, L. Van Dorp, and D. Falush. A tutorial on how not to over-interpret structure and admixture bar plots. *Nature Communications*, 9(1):1–11, 2018.

I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, N. Rohland, S. Mallick, D. Fernandes, M. Novak, B. Gamarra, K. Sirak, et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.

I. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, N. Rohland, S. Pfrengle, A. Furtwängler, A. Peltzer, C. Posth, A. Vasilakis, et al. Genetic origins of the Minoans and Mycenaeans. *Nature*, 548(7666):214–218, 2017.

M. Le Lannou. Pâtres et Paysans de la Sardaigne. *Tours*, 8:364, 1941.

S. Lee, F. Zou, and F. A. Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, 38(6):3605, 2010.

G. Lettre and J. N. Hirschhorn. Small island, big genetic discoveries. *Nature Genetics*, 47 (11):1224–1225, 2015.

H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

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

M. Lipson, A. Szécsényi-Nagy, S. Mallick, A. Pósa, B. Stégmár, V. Keerl, N. Rohland, K. Stewardson, M. Ferry, M. Michel, et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, 2017.

L. T. Liu, E. Dobriban, A. Singer, et al. *e* pca: High dimensional exponential family pca. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018.

P.-R. Loh, M. Lipson, N. Patterson, P. Moorjani, J. K. Pickrell, D. Reich, and B. Berger. Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, 193(4):1233–1254, 2013.

C. Lugliè. Your path led trough the sea... the emergence of Neolithic in Sardinia and Corsica. *Quaternary International*, 470:285–300, 2018.

E. Lundgren and P. L. Ralph. Are populations like a circuit? comparing isolation by resistance to a new coalescent-based method. *Molecular ecology resources*, 19(6):1388–1406, 2019.

G. Malécot. Les mathématiques de l'hérédité. masson et cie. *Paris, France*, 1948.

J. Marcus, J. Willwerscheid, and P. Carbonetto. jhmarcus/drift-workflow: dissertation, Sept. 2020a. URL `https://doi.org/10.5281/zenodo.4012448`.

J. H. Marcus, W. Ha, R. F. Barber, and J. Novembre. Fast and flexible estimation of effective migration surfaces. *bioRxiv*, 2020b.

J. H. Marcus, C. Posth, H. Ringbauer, L. Lai, R. Skeates, C. Sidore, J. Beckett, A. Furtwängler, A. Olivieri, C. W. Chiang, et al. Genetic history from the middle neolithic to present on the mediterranean island of sardinia. *Nature communications*, 11(1): 1–14, 2020c.

A. R. Martin, K. J. Karczewski, S. Kerminen, M. I. Kurki, A.-P. Sarin, M. Artomov, J. G. Eriksson, T. Esko, G. Genovese, A. S. Havulinna, et al. Haplotype sharing provides insights into fine-scale population history and disease in finland. *The American Journal of Human Genetics*, 102(5):760–775, 2018.

A. Mastino. *Storia della Sardegna antica*, volume 2. Il Maestrale, 2005.

G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro. Connecting the dots: Identifying network structure via graph signal processing. *IEEE Signal Processing Magazine*, 36(3): 16–43, 2019.

I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, et al. The genomic history of southeastern Europe. *Nature*, 555(7695):197, 2018.

E. Matisoo-Smith, A. Gosling, D. Platt, O. Kardailsky, S. Prost, S. Cameron-Christie, C. Collins, J. Boocock, Y. Kurumilian, M. Guirguis, et al. Ancient mitogenomes of Phoenicians from Sardinia and Lebanon: A story of settlement, integration, and female mobility. *PloS ONE*, 13(1):e0190169, 2018.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

P. McCullagh. Marginal likelihood for distance matrices. *Statistica Sinica*, pages 631–649, 2009.

B. H. McRae. Isolation by resistance. *Evolution*, 60(8):1551–1561, 2006.

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

P. G. Meirmans. The trouble with isolation by distance. *Molecular ecology*, 21(12):2839–2846, 2012.

P. Melis. *Un Approdo della costa di Castelsardo, fra età nuragica e romana.* Carocci, 2002.

G. M. Meloni. Ricerche archeologiche nelle località di Corona Moltana e Zarau. In Conca C., editor, *Bonnanaro e il suo patrimonio culturale*, pages 90–99. Segnavia, Sassari, 2004.

M. Meyer and M. Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6):pdb–prot5448, 2010.

A. Miller, L. Bornn, R. Adams, and K. Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *International conference on machine learning*, pages 235–243, 2014.

A. Modi, F. Tassi, R. R. Susca, S. Vai, E. Rizzi, G. De Bellis, C. Lugliè, G. G. Fortes, M. Lari, G. Barbujani, et al. Complete mitochondrial sequences from Mesolithic Sardinia. *Scientific Reports*, 7:42869, 2017.

S. Moscati. La penetrazione fenicia e punica in Sardegna. *Memorie della Accademia Nazionale dei Lincei, classe di scienze morali, storiche e filologiche*, 8.7.3:215–250, 1966.

K. P. Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

M. R. Nelson, K. Bryc, K. S. King, A. Indap, A. R. Boyko, J. Novembre, L. P. Briley, Y. Maruyama, D. M. Waterworth, G. Waeber, et al. The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358, 2008.

J. Nocedal and S. Wright. *Numerical optimization.* Springer Science & Business Media, 2006.

J. Novembre and B. M. Peter. Recent advances in the study of fine-scale population structure in humans. *Current opinion in genetics & development*, 41:98–105, 2016.

J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, 40(5):646–649, 2008.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

I. Olalde, H. Schroeder, M. Sandoval-Velasco, L. Vinner, I. Lobón, O. Ramirez, S. Civit, P. García Borja, D. C. Salazar-García, S. Talamo, et al. A common genetic origin for early farmers from Mediterranean Cardial and Central European LBK cultures. *Molecular Biology and Evolution*, 32(12):3132–3142, 2015.

I. Olalde, S. Brace, M. E. Allentoft, I. Armit, K. Kristiansen, T. Booth, N. Rohland, S. Mallick, A. Szécsényi-Nagy, A. Mittnik, et al. The Beaker phenomenon and the genomic transformation of Northwest Europe. *Nature*, 555(7695):190, 2018.

I. Olalde, S. Mallick, N. Patterson, N. Rohland, V. Villalba-Mouco, M. Silva, K. Dulias, C. J. Edwards, F. Gandini, M. Pala, P. Soares, M. Ferrando-Bernal, N. Adamski, N. Broomandkhoshbacht, O. Cheronet, B. J. Culleton, D. Fernandes, A. M. Lawson, M. Mah, J. Oppenheimer, K. Stewardson, Z. Zhang, J. M. Jiménez Arenas, I. J. Toro Moyano, D. C. Salazar-García, P. Castanyer, M. Santos, J. Tremoleda, M. Lozano, P. García Borja, J. Fernández-Eraso, J. A. Mujika-Alustiza, C. Barroso, F. J. Bermúdez, E. Viguera Mínguez, J. Burch, N. Coromina, D. Vivó, A. Cebrià, J. M. Fullola, O. García-Puchol, J. I. Morales, F. X. Oms, T. Majó, J. M. Vergès, A. Díaz-Carvajal, I. Ollich-Castanyer, F. J. López-Cachero, A. M. Silva, C. Alonso-Fernández, G. Delibes de Castro, J. Jiménez Echevarría, A. Moreno-Márquez, G. Pascual Berlanga, P. Ramos-García, J. Ramos-Muñoz, E. Vijande Vila, G. Aguilella Arzo, Á. Esparza Arroyo, K. T. Lillios, J. Mack, J. Velasco-Vázquez, A. Waterman, L. Benítez de Lugo Enrich, M. Benito Sánchez, B. Agustí, F. Codina, G. de Prado, A. Estalrrich, Á. Fernández Flores, C. Finlayson, G. Finlayson, S. Finlayson, F. Giles-Guzmán, A. Rosas, V. Barciela González, G. García Atiénzar, M. S. Hernández Pérez, A. Llanos, Y. Carrión Marco, I. Collado Beneyto, D. López-Serrano, M. Sanz Tormo, A. C. Valera, C. Blasco, C. Liesau, P. Ríos, J. Daura, M. J. de Pedro Michó, A. A. Diez-Castillo, R. Flores Fernández, J. Francès Farré, R. Garrido-Pena, V. S. Gonçalves, E. Guerra-Doce, A. M. Herrero-Corral, J. Juan-Cabanilles, D. López-Reyes, S. B. McClure, M. Merino Pérez, A. Oliver Foix, M. Sanz Borràs, A. C. Sousa, J. M. Vidal Encinas, D. J. Kennett, M. B. Richards, K. Werner Alt, W. Haak, R. Pinhasi, C. Lalueza-Fox, and D. Reich. The genomic history of the Iberian Peninsula over the past 8000 years. *Science*, 363 (6432):1230–1234, 2019. ISSN 0036-8075. doi: 10.1126/science.aav4040.

A. Olivieri, C. Sidore, A. Achilli, A. Angius, C. Posth, A. Furtwängler, S. Brandini, M. R. Capodiferro, F. Gandini, M. Zoledziewska, et al. Mitogenome diversity in Sardinians: a genetic window onto an island's past. *Molecular Biology and Evolution*, 34(5):1230–1239, 2017.

L. Orlando, M. T. P. Gilbert, and E. Willerslev. Reconstructing ancient genomes and epigenomes. *Nature Reviews Genetics*, 16(7):395–408, 2015.

L. Ortu. *Storia della Sardegna dal Medioevo all'età contemporanea*. Cuec, 2011.

J. A. Palacios, J. Wakeley, and S. Ramachandran. Bayesian nonparametric inference of population size changes from sequential genealogies. *Genetics*, 201(1):281–304, 2015.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2(12):e190, 2006.

N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

A. Peltzer, G. Jäger, A. Herbig, A. Seitz, C. Kniep, J. Krause, and K. Nieselt. Eager: efficient ancient genome reconstruction. *Genome Biology*, 17(1):60, 2016.

B. M. Peter, D. Petkova, and J. Novembre. Genetic landscapes reveal how human genetic diversity aligns with geography. *BioRxiv*, page 233486, 2018.

D. Petkova, J. Novembre, and M. Stephens. Visualizing spatial population structure with estimated effective migration surfaces. *Nature Genetics*, 48(1):94, 2016.

D. I. Petkova. *Inferring effective migration from geographically indexed genetic data*. The University of Chicago, 2013.

J. Pickrell and J. Pritchard. Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, pages 1–1, 2012.

J. K. Pickrell and D. Reich. Toward a new history and geography of human genes informed by ancient DNA. *Trends in Genetics*, 30(9):377–389, 2014.

E. Pompianu and C. Murgia. Nuovi scavi nella necropoli punica di Villamar. un primo bilancio delle ricerche 2013-2015. In G. Serreli, R. Melis, C. French, and F. Sulas, editors, *Sa Massarìa: ecologia storica dei sistemi di lavoro contadino in Sardegna. (Europa e Mediterraneo. Storia e immagini di una comunità internazionale 37)*, pages 455–504. CNR, Cagliari, 2017.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

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

A. Raj, M. Stephens, and J. K. Pritchard. faststructure: variational inference of population structure in large snp data sets. *Genetics*, 197(2):573–589, 2014.

C. E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5), 2014.

G. Renaud, V. Slon, A. T. Duggan, and J. Kelso. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology*, 16 (1):224, 2015.

H. Ringbauer, A. Kolesnikov, D. L. Field, and N. H. Barton. Estimating barriers to gene flow from distorted isolation-by-distance patterns. *Genetics*, 208(3):1231–1245, 2018.

H. Ringbauer, J. Novembre, and M. Steinruecken. Detecting runs of homozygosity from low-coverage ancient dna. *bioRxiv*, 2020.

S. G. Rodriques, R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, 2019.

N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, and D. Reich. Partial uracil–dna–glycosylase treatment for screening of ancient dna. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1660):20130624, 2015.

N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

D. Rovina and P. Fiori M. Olia. Il Duomo e il cimitero di San Nicola. In F. M. Rovina D., editor, *Sassari : Archeologia Urbana*, pages 120–129. Felici Editore, 2013.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

T. Safner, M. P. Miller, B. H. McRae, M.-J. Fortin, and S. Manel. Comparison of bayesian clustering and edge detection methods for inferring boundaries in landscape genetics. *International Journal of Molecular Sciences*, 12(2):865–889, 2011.

K. Sahr, D. White, and A. J. Kimerling. Geodesic discrete global grid systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003.

S. Sarno, A. Boattini, L. Pagani, M. Sazzini, S. De Fanti, A. Quagliariello, G. A. G. Ruscone, E. Guichard, G. Ciani, E. Bortolini, et al. Ancient and recent admixture layers in Sicily and Southern Italy trace multiple migration routes along the Mediterranean. *Scientific Reports*, 7(1):1984, 2017.

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

J. G. Schraiber and J. M. Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16(12):727–740, 2015.

M. Schubert, S. Lindgreen, and L. Orlando. Adapterremoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, 9(1):88, 2016.

R. M. Schweizer, B. M. Vonholdt, R. Harrigan, J. C. Knowles, M. Musiani, D. Coltman, J. Novembre, and R. K. Wayne. Genetic subdivision and candidate genes under selection in north american grey wolves. *Molecular Ecology*, 25(1):380–402, 2016.

C. Sidore, F. Busonero, A. Maschio, E. Porcu, S. Naitza, M. Zoledziewska, A. Mulas, G. Pistis, M. Steri, F. Danjou, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature Genetics*, 47(11):1272–1281, 2015.

M. Sikora, M. L. Carpenter, A. Moreno-Estrada, B. M. Henn, P. A. Underhill, F. Sánchez-Quinto, I. Zara, M. Pitzalis, C. Sidore, F. Busonero, et al. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. *PLoS Genetics*, 10(5):e1004353, 2014.

M. Siniscalco, L. Bernini, G. Filippi, B. Latte, P. M. Khan, S. Piomelli, and M. Rattazzi. Population genetics of haemoglobin variants, thalassaemia and glucose-6-phosphate dehydrogenase deficiency, with particular reference to the malaria hypothesis. *Bulletin of the World Health Organization*, 34(3):379, 1966.

R. Skeates, M. G. Gradoli, and J. Beckett. The cultural life of caves in Seulo, central Sardinia. *Journal of Mediterranean Archaeology*, 26(1):97–126, May 2013.

P. Skoglund, H. Malmström, M. Raghavan, J. Storå, P. Hall, E. Willerslev, M. T. P. Gilbert, A. Götherström, and M. Jakobsson. Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science*, 336(6080):466–469, 2012.

P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, C. Valdiosera, T. Günther, P. Hall, K. Tambets, J. Parik, K.-G. Sjögren, et al. Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. *Science*, 344(6185):747–750, 2014a.

P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, S. Pääbo, J. Krause, and M. Jakobsson. Separating endogenous ancient dna from modern day contamination in a siberian neandertal. *Proceedings of the National Academy of Sciences*, 111(6):2229–2234, 2014b.

L. Skotte, T. S. Korneliussen, and A. Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, 2013.

M. Slatkin. Gene flow in natural populations. *Annual review of ecology and systematics*, 16 (1):393–430, 1985.

M. Slatkin. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, 47(1):264–279, 1993.

D. Speed and D. J. Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 2017.

G. Strang. *Linear algebra and learning from data.* Wellesley-Cambridge Press, 2019.

H. Tang, J. Peng, P. Wang, and N. J. Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 28(4):289–301, 2005.

R. H. Tykot. Obsidian procurement and distribution in the central and western Mediterranean. *Journal of Mediterranean Archaeology*, 9:39–82, 1996.

P. Van Dommelen. Punic farms and Carthaginian colonists: surveying Punic rural settlement in the central Mediterranean. *Journal of Roman Archaeology*, 19:7–28, 2006.

M. Van Oven and M. Kayser. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, 30(2):E386–E394, 2009.

D. Vianello, F. Sevini, G. Castellani, L. Lomartire, M. Capri, and C. Franceschi. Haplofind: A new method for high-throughput mt DNA haplogroup assignment. *Human Mutation*, 34(9):1189–1194, 2013.

C. Viganó, C. Haas, F. J. Rühli, and A. Bouwman. 2,000 year old β-thalassemia case in Sardinia suggests malaria was endemic by the Roman period. *American journal of physical anthropology*, 164(2):362–370, 2017.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: https://doi.org/10.1038/s41592-019-0686-2.

N. K. Vishnoi et al. Lx= b. *Foundations and Trends in Theoretical Computer Science*, 8 (1–2):1–141, 2013.

J. Wakeley. *Coalescent theory: an introduction*. Number 575: 519.2 WAK. 2009.

C. Wang, X. Zhan, L. Liang, G. R. Abecasis, and X. Lin. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *The American Journal of Human Genetics*, 96(6):926–937, 2015.

W. Wang and M. Stephens. Empirical bayes matrix factorization. *arXiv preprint arXiv:1802.06931*, 2018.

Y.-X. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

G. Webster. *The Archaeology of Nuragic Sardinia*, volume 14 of *Monographs in Mediterranean Archaeology*. Equinox Publishing, 2016.

A. Wilson, M. Kasy, and L. Mackey. Approximate cross-validation: Guarantees for model assessment and selection. *arXiv preprint arXiv:2003.00617*, 2020.

S. Wright. Breeding structure of populations in relation to speciation. *The American Naturalist*, 74(752):232–248, 1940.

S. Wright. Isolation by Distance. *Genetics*, 28(2):114, 1943.

S. Wright. Isolation by distance under diverse systems of mating. *Genetics*, 31(1):39, 1946.

S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise. Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proceedings of the National Academy of Sciences*, 113(16):4290–4295, 2016.

J. Yan, N. Patterson, and V. M. Narasimhan. miqograph: Fitting admixture graphs using mixed-integer quadratic optimization. *bioRxiv*, page 801548, 2019.

J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

P. Zalloua, C. J. Collins, A. Gosling, S. A. Biagini, B. Costa, O. Kardailsky, L. Nigro, W. Khalil, F. Calafell, and E. Matisoo-Smith. Ancient DNA of Phoenician remains indicates discontinuity in the settlement history of Ibiza. *Scientific Reports*, 8(1):17567, 2018.

J. Zilhão. Radiocarbon evidence for maritime pioneer colonization at the origins of farming in west Mediterranean Europe. *Proceedings of the National Academy of Sciences*, 98(24): 14180–14185, 2001.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.