The Distribution of Neanderthal Ancestry Across Populations And Within Genomes

Aaron B. Wolf

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jay Shendure, Chair

Philip Green

Joshua Akey

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

The Distribution of Neanderthal Ancestry Across Populations And Within Genomes

Aaron B. Wolf

Chair of the Supervisory Committee:
Professor Jay Shendure
Dept. of Genome Sciences, University of Washington

For many millennia, modern humans overlapped in time and space with archaic humans such as Neanderthals and Denisovans. We now know that modern and archaic humans interbred, and that modern human populations carry some amount of archaic ancestry. The complexities of this admixture history, however, have persisted as outstanding questions in the field of archaic genomics. This dissertation focuses on some of these questions.

In the first chapters of this dissertation, I focus on methods for detecting archaic sequence in modern humans and their application to geographically diverse populations. These studies provide insight regarding the distribution of Neanderthal ancestry across populations, and illustrate the complexity of the admixture dynamics between archaic and modern humans. Specifically, they show how pervasive archaic ancestry is across modern populations, being found in every population examined to date. Furthermore, they highlight a role for both Neanderthal-to-human gene flow and human-to-Neanderthal gene flow.

These initial chapters also describe the heterogeneous distribution of archaic sequence across the genome, and raise questions about the dynamics and mechanisms by which archaic sequence was retained or lost. The later chapter of this dissertation extensively examines the mechanisms responsible for forming regions significantly depleted of archaic ancestry. By modeling demographic histories of admixture, I show that these regions are possible under neutral processes only through extreme conditions.

Altogether, this work elucidates some of the complexities of archaic and modern human admixture and answers several of the outstanding questions in this field. Leveraging the discovery of archaic sequence provides an opportunity to better understand the evolutionary processes that have shaped the modern human genome and led to exceptional human phenotypes.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This research has truly been a peripatetic journey, requiring two cross-country moves and taking me to two different universities and three different departments. As such, the number of faculty, staff, students, colleagues, and friends who have made this work possible and made my life enjoyable during it, are too many to list comprehensively. Pretty much, if you think you were involved in any way, you certainly were, so thank you.

There are some individuals that deserve explicit acknowledgement. I am exceptionally grateful to my advisor, Joshua Akey, who took me into the lab, despite my inexperience, and allowed me to work on a project that had been my passion for the many years prior and remains so today.

I would not have gotten this far in my research without the incredible mentoring from many in the Akey lab, especially Josh Schraiber, Rajiv McCoy, and Selina Vattathil. Other past and present members of the lab, Katherine Xue, Anne Clark, Rachel Gittelman, Sunjin Moon, Lu Chen, Liming Li, Ben Vernot, Serena Tucci, and Troy Comi contributed massively to my work and my experiences in Seattle and Princeton.

I am also grateful to the members of my committee, Jay Shendure, Phil Green, Christine Queitsch, and Dan Eisenberg, who have always provided excellent feedback and direction, and were supportive and patient during our lab's move to Princeton. Christine, especially, has been a great source of inspiration and mentorship. Her excitement for my research and future ambitions has helped me regain my confidence when it has faltered.

Lastly, I have to acknowledge the incredible support of my friends and family. I was only able to complete my work because I pursued passions outside of this research. My parents, teammates, housemates, and close friends were instrumental in this. Thank you.

Chapter 1. Outstanding Questions in the Study of Archaic Hominin Admixture

Parts of this chapter are adapted from:

> **Wolf, AB**, Akey JM (2018) Outstanding questions in the study of archaic hominin admixture. PLoS Genet 14(5):e1007349

1.1 Introduction

Today, humans are the only hominin species walking the planet. This exclusivity is a recent feature of our species' history. Specifically, though humans with anatomically modern features first appear in the archaeological record 200kya-300kya [1–4], other hominins persisted until as recently as 30kya-40kya [5,6]. In some cases, modern humans overlapped temporally and spatially with archaic humans, for example Neanderthals and Denisovans and perhaps others [7,8]. Neanderthals left a rich archaeological and paleontological record and resided in the Middle East, Europe, and parts of Asia [9–11]. Denisovans, whom we only discovered from ancient DNA taken from a single finger bone and three teeth [12–14], are believed to have resided in parts of East and Southeast Asia.

There has been long standing interest in whether modern humans and archaic human ancestors hybridized. Historically, attempts at answering this question focused on archaeological remains and compared dental, cranial, and post-cranial features from modern human and archaic human sites for evidence of hybrid morphologies [15]. By the early 2000s, technological innovations enabled the extraction and sequencing of mitochondrial DNA from archaic human remains [16–19], and eventually facilitated the capture and sequencing of the full nuclear genome [20–22].

The complete sequencing of archaic and modern human nuclear genomes led to the discovery that modern non-African human populations shared more genetic ancestry with archaic humans than did African populations [22]. Initial inferences demonstrated a strong likelihood of hybridization between archaic humans and the ancestors of all modern non-African populations, and these results proved robust to alternative explanations, such as archaic population structure. The continued development of ancient DNA technology facilitated

extraction and sequencing of high-quality Neanderthal [23] and Denisovan [13] reference genomes. These foundational resources, coupled with advances in statistical and computational tools for analyzing ancient genomes, enabled the identification of sequences inherited from archaic ancestors (i.e. introgressed sequences) in the genomes of modern human individuals.

Considerable progress has been made in the study of archaic hominin admixture, which has been reviewed elsewhere [24–28]. However, many outstanding questions remain, whose resolution is critical to more completely understand the history and consequence of admixture between archaic and modern humans. Below, I outline several of these questions, which include cataloguing Neanderthal sequence across global populations, refining models of admixture history, and determining the mechanisms responsible for the loss and retention of archaic sequence. This dissertation focuses on further exploration of these questions.

1.2 How high was the initial level of archaic-modern human admixture?

All modern non-African genomes are estimated to carry ~2-7% archaic human sequence; approximately 2% ancestry from Neanderthals and an additional 2-5% ancestry from Denisovans in Melanesian populations [29–31]. However, present day levels of archaic ancestry need not reflect initial admixture levels, which is of special interest in understanding human history. Specifically, an accurate estimate of initial admixture levels would provide significant insights into models of hybridization and admixture dynamics.

Following the discovery of ~2% Neanderthal ancestry in modern non-Africans [22], it was estimated the initial level of admixture between Neanderthals and modern humans was also ~2% [32–34]. Further analyses revealed, however, large depletions of Neanderthal ancestry across the human genome, indicating widespread purging of deleterious Neanderthal sequence. For example, in the 20% of the genome with the lowest density of functionally important elements, Neanderthal ancestry is 1.54x the genome-wide average [34]. Assuming this subset of the genome to be unaffected by selection, the implication is that the initial proportion of Neanderthal ancestry after admixture was >3%.

Several recent analyses have estimated the initial Neanderthal admixture proportion was dramatically higher than 3%. These studies propose the prolonged small effective population size of Neanderthals led to a high frequency of weakly deleterious alleles in the Neanderthal

population [35,36]. When these Neanderthal alleles entered the human population, with a comparatively larger effective population size, they were more readily removed by selection. Using simulations and models reflecting this expectation, these studies estimate the initial admixture proportion to have been 2x-5x the level present in modern human genomes.

Analyses of additional ancient DNA samples support the projection that initial admixture levels were much higher than those found today. Genome-wide data from Eurasian samples ranging in age from 7kya-45kya suggest an initial Neanderthal admixture proportion close to 6%, which decreased gradually over time to a contemporary level of 2% [37]. Because all the individuals analyzed are descended from a single founding population, the authors argue the steady decline in Neanderthal ancestry is driven by natural selection against introgressed sequence, and not dilution from a non-admixed population.

Recent re-analysis of the these same ancient Eurasian samples, but with an updated admixture statistic, have contradicted the loss of Neanderthal ancestry as a gradual decline [38]. Instead, the authors show that the loss of Neanderthal ancestry occurred rapidly, within only a few generations after admixture. Simulations of selection against introgressed archaic sequence indicate high frequency and weakly deleterious alleles best replicate the initial rapid loss of Neanderthal sequence followed by a more gradual loss observed in empirical data [35,38].

Answering questions regarding the initial admixture level, the duration of admixture, and the rate of archaic sequence loss will depend on the continued collection of ancient human and Neanderthal samples closer to the time of admixture. Furthermore, updated measures have highlighted the sensitivity of certain admixture statistics to assumptions about demographic histories [38]. Re-analysis of current data with new methods that are less sensitive to these features or more accurately capture past demographic events will also provide new insights.

1.3 How many distinct pulses of admixture occurred with Neanderthals?

Closely related to the question of how high initial levels of admixture were, the number of independent admixture events (sometimes referred to as “pulses” in the literature) is also uncertain. Initially, this question seemed to have a simple answer—admixture with Neanderthals occurred once in human history. Early studies found that all non-Africans carried approximately equal levels of Neanderthal ancestry [22]. Therefore, the most parsimonious model involved a

single pulse of admixture between Neanderthals and an Out-Of-Africa wave of human migrants, before the ancestral Eurasian population split into European and Asian lineages (**Figure 1.1**).

However, as researchers analyzed more globally diverse populations with refined methods, they found that levels of Neanderthal ancestry varied among populations. Analyses of introgressed Neanderthal sequence using the high-quality Altai reference genome [23] noted more regions of Neanderthal origin in Asian and American populations than European ones [32], as well as higher levels of Neanderthal ancestry in East Asian populations compared to European populations, and lower levels of Neanderthal ancestry in Melanesians compared to either East Asians or Europeans [30,31] (**Figure 1.2A**). The differences in these Neanderthal ancestry proportions are on the order of 0.1-0.5%. This small difference, however, manifests as an ~20% enrichment of Neanderthal ancestry in East Asians compared to Europeans. How we interpret the global variation in archaic human ancestry has a profound impact on our understanding of human history, informing our estimate for the frequency of archaic-modern human admixture—either as isolated in time and space, or recurrent and pervasive.

Considering the variation in levels of Neanderthal ancestry among populations, a single pulse of admixture may still be the most parsimonious explanation. For example, admixture between ancient Europeans and populations lacking Neanderthal ancestry could have diluted the amount of Neanderthal ancestry remaining in modern European populations [13,39] (**Figure 1.1**). It has also been proposed that East Asian demographic history led to retention of more Neanderthal sequence [34].

Alternatively, several analyses using statistical and simulation approaches demonstrate that models incorporating multiple pulses of admixture better explain the data [33,30,40–42]. These models include an initial admixture event into an ancestral non-African population followed by additional admixture events into an ancestral Eurasian population and ancestral East Asian population [42], or just an ancestral Asian population [33] (**Figure 1.1**). Studies simulating admixture over a range of selection and demographic models can only account for the higher proportion of Neanderthal ancestry in East Asians compared to Europeans by including multiple pulses of admixture [41,43]. Even a “two-pulse” model may be too simplistic a representation for the history of human and Neanderthal admixture. Simulations that included additional admixture events, like an intermediate admixture pulse into the ancestral population of Europeans and East Asians but not Southeast Asians, are also compatible with the empirical data [30].

Lastly, it is important to note that the estimated number of "pulses" of admixture is an oversimplification assumed for modeling, representing the minimum number of distinct admixture events that provided sequence still present in modern humans. The true frequency of admixture events may therefore have been much higher, with gene flow between modern human and archaic hominin populations occurring at low rates.

1.4 Did archaic hominin admixture happen in Africa?

While the genetic evidence, collected from archaic and modern human DNA samples, persuasively demonstrates archaic hominin admixture in non-African populations, similar studies of archaic admixture in African populations have been limited. This is despite the fact that numerous archaic hominin lineages are known to have existed in Africa [44], and may have overlapped in time and space with modern humans [45]. Studies of archaic admixture in Africans have been hindered by the historical underrepresentation of African populations in large genomic datasets and the absence of reference genomes for archaic African hominins—the combined effects of the greater age of archaic samples and challenging climate impeding the recovery of ancient DNA.

Several studies, however, have made a concerted effort to investigate the likelihood of archaic admixture in African populations, leveraging LD-based [46,47] and demographic model-based [48] methods for detecting signals of archaic admixture without an archaic reference genome. Evidence from these early studies does indicate admixture occurred between an unidentified archaic hominin ancestor and several African populations, and contributed functionally relevant genetic variation at specific loci, such as the salivary *MUC7* locus [49]. In the absence of any recovered ancient DNA samples, "excavating" archaic sequences from modern African genomes may be the best strategy to identify archaic hominin lineages. Although studies of archaic admixture in Africa are limited and have been necessarily cautious in their conclusions, significant new discoveries should be anticipated as more genomic data from diverse African populations become available and new methods are developed to analyze them.

1.5 Was there gene-flow from modern humans into Neanderthals?

Most studies to date have focused on the Neanderthal contribution to the modern human genome through hybridization. However, investigating the contribution of modern human admixture to these archaic hominin genomes is also of great interest. New research is uncovering instances of potential gene flow from early humans into Neanderthal populations. Analyses of nuclear DNA from multiple Neanderthal samples and modern humans [50] support models in which an early human population—diverged from the population ancestral to contemporary Africans and non-Africans—contributed low levels (0.1-2%) of sequence to a Neanderthal lineage ~100kya (**Figure 1.1**). More complete data from these archaic DNA samples [51] improved these estimates, demonstrating the gene flow event occurred at least 130-145kya into a lineage ancestral to both Vindija and Altai Neanderthal populations. Analyses of mitochondrial DNA from multiple Neanderthal and human samples support an even earlier gene-flow event from humans into Neanderthals, potentially as early as ~300kya [52]. These studies highlight the importance of collecting and analyzing additional archaic samples, as well as illustrate the complexity and likely pervasiveness of admixture between different hominin groups.

Characterizing human gene flow into archaic hominins has important implications for identifying archaic sequence in modern human populations. If gene flow occurred from humans into Neanderthals at a time before the split of African and non-African populations, as detailed in numerous analyses [50,51,53], this would increase the level of sequence matching between Neanderthal and African genomes. For methods that rely on matching a Neanderthal reference genome to a modern human genome to identify introgressed sequence in human populations, this could increase the rate of false-discoveries, especially in African populations that have not experienced direct Neanderthal hybridization. In methods that rely on an African reference panel to control for shared ancestry between hominin populations (e.g. S* and Reich's CRF), human to Neanderthal gene flow could have the opposite effect, limiting the power to detect tracts of introgressed Neanderthal sequence in modern humans.

1.6 What caused deserts of archaic sequence to form?

Compiling the surviving introgressed archaic human haplotypes in hundreds of individuals from geographically diverse populations led to a "map" of introgressed sequence across the human genome. While introgressed sequence tends to be widespread across the genome, covering all 22 autosomes and the 2 sex chromosomes, it was a striking discovery to find that there also exist large depletions—"deserts"—of archaic ancestry (**Figure 1.2A**).

On the autosomes, the largest deserts span multiple megabases, with a handful extending up to 10Mb in length [33,34]. The mechanisms responsible for the heterogeneous distribution of Neanderthal sequence across the autosomes are not yet fully understood, and several may act in combination. Understanding the processes responsible for this heterogeneous distribution could be informative about what distinguished modern and archaic humans.

One proposed explanation for autosomal deserts is that they resulted from intense bottlenecks in the human population [34]. Theoretically, a bottleneck soon after admixture with Neanderthals could cause the rapid loss of large introgressed haplotypes, before they could be broken apart by generations of recombination. Simulations exploring these extreme demographic scenarios have found genetic drift able to explain some, but not all, of the observed data [30].

Alternatively, selection against Neanderthal haplotypes at desert loci might also generate large depletions of archaic sequence. Selection against specific deleterious Neanderthal alleles in the admixed population could remove large swaths of linked archaic sequence. Deserts of introgressed sequence do exhibit higher levels of background selection and human-Neanderthal sequence divergence [30,33]. Furthermore, deserts of Neanderthal sequence overlap with deserts of Denisovan sequence significantly more often than expected by chance [30]. These data indicate the repeated loss of archaic DNA at specific loci across multiple independent admixture events.

If selection played a part in the removal of large Neanderthal haplotypes and the formation of deserts, an obvious question is whether selection acted strongly on a very few sites, or weakly across multiple sites (**Figure 1.2B**). Studies modeling the effective population sizes of Neanderthals and humans before and during admixture, demonstrate that the small size of the Neanderthal population would have allowed weakly deleterious alleles to drift as if neutral and accumulate at a high frequency [35]. When these alleles entered the human population through

admixture, the effective size of the human population need only have been marginally larger than the Neanderthal population to increase the strength of selection against these alleles and effect their removal. At the same time, deserts of introgression tend to exhibit higher levels of background selection and are also significantly enriched for genes expressed in the brain, such as *FOXP2*, which is essential to speech and language development [30,33,54]. These patterns suggest strong selection at a single locus could drive the loss of Neanderthal sequence across a wide region. Furthermore, environmental differences between modern and archaic humans may have meant that, rather than just the force of selection changing, the selection pressures themselves might have changed for archaic alleles when they entered the human population. What was potentially advantageous or neutral in a Neanderthal population may have been deleterious in a human one.

Finally, it is important to note explanations beyond drift and selection in forming deserts. For instance, large inversions, on either the human or Neanderthal and Denisovan lineages, could theoretically prevent introgression in these regions by suppressing recombination. Considering the overlap of Neanderthal and Denisovan deserts [30], large inversions seem unlikely to explain all of the archaic depletions found to date, but remain a formal possibility. Unfortunately, identifying potential lineage-specific inversions is incredibly difficult given the deterioration of ancient DNA samples and short sequencing read lengths.

1.7 What are the functional and phenotypic consequences of hybridization?

A critical question in studying archaic-modern human hybridization is the functional impact of the remaining introgressed archaic sequence in the modern human genome. How has introgressed sequence shaped human evolution? How is it currently affecting modern human phenotypes and health and disease? Is the effect of Neanderthal sequence on human phenotypes proportional to the low amount of Neanderthal ancestry present in the human genome, or are there instances where Neanderthal ancestry has a disproportionately large effect?

Several studies examining certain modern human populations have identified introgressed Neanderthal haplotypes that have risen to higher frequency than expected by drift (**Figure 1.2D**). The functional significance of these genes has been hypothesized based on prior biological studies and association with normal and disease phenotypes [55,56]. For example, a

Neanderthal version of the gene *BNC2* was identified at high frequency in several non-African populations—a sign of putative adaptive introgression—and is associated with skin pigmentation levels in Europeans [55,57]. Additionally, putatively adaptive introgressed sequences have been identified at several genes that play key roles in immunological function, such as *STAT2* [58], *OAS1* [59], and *TLR1/6/10* [57,60]. There are also examples of certain populations carrying putatively adaptive introgressed sequences from Denisovans, such as in Greenlandic Inuit the genes *TBX15* and *WARS2* [61]—associated with adipose tissue differentiation and distribution—and in Tibetans the high-altitude adaptation gene *EPAS1* [62]. These and other instances of possible adaptive introgression [63] support the hypothesis that archaic hominins, who inhabited Eurasia for 400ky before humans, would have been a source of advantageous genetic variants pre-adapted to local environmental features, like colder climates, lower UV-exposure, and endemic pathogens.

Alternatively, archaic hominins may simply have provided a reservoir of additional genetic variation to modern humans, some of which happened to be advantageous following introgression into modern humans, but which was not necessarily pre-adapted to the Eurasian environment. It should also be noted that selection against introgressed archaic alleles seems to have been the predominant pattern across the genome, as indicated by the low levels of retained archaic ancestry, especially in functionally important regions. This raises questions of whether admixture was beneficial overall for ancient humans migrating out of Africa, and to what extent the benefits of some alleles were able to outweigh the costs of others.

While there is increasing power to detect these archaic introgressed segments in modern human populations, our understanding of the evolutionary and fitness consequences remains murky. In the case of introgressed Neanderthal sequence, researchers have leveraged large association studies to infer the effects of archaic sequences [55,56]. Applying a similar approach to introgressed Denisovan sequences has proven more difficult, since those populations with Denisovan ancestry are underrepresented in large association studies and genomic data sets. Furthermore, if our intent is to understand the features under selection at the time of introgression, it is important to remember that the phenotypic effects of archaic sequence we see manifested today appear in very different environments than the ones for which they would have been selected. This will confound attempts to draw connections between the phenotypic effects of introgressed archaic sequences today and the original selected phenotypes.

In addition to leveraging association studies or prior biological findings to infer the effects of introgressed sequence, several recent studies have examined the direct effects of introgressed variants on gene regulation (**Figure 1.2C**). At sites of putatively adaptive introgressed archaic sequences, researchers have observed Neanderthal alleles affecting expression levels of immunologically relevant genes *OAS1/2/3* and *TLR1/6/10*, and observed that these expression differences can be cell-type specific and influenced by environmental stimuli [57]. Others have correlated the genotypes of putatively introgressed Neanderthal alleles with the expression of nearby genes and found introgressed archaic alleles contribute proportionally more to expression variation than non-archaic alleles [64]. In individuals that are heterozygous for the Neanderthal and human alleles, researchers found frequent instances of allele-specific expression, and a significant down regulation of Neanderthal allele expression in specific brain sub-regions and the testes, relative to other tissues [65]. These findings suggest the phenotypic effects of introgressed archaic sequences are more likely mediated through gene regulation than protein changes (**Figure 1.2C**). Recent developments in genomic editing technologies should allow future studies to more thoroughly explore these regulatory effects through *in vitro* experiments.

1.8 Summary

The complete sequencing of archaic and modern human genomes and the discovery that all non-African populations carry ~2% Neanderthal ancestry was a significant breakthrough in anthropology and paleogenomics. Subsequent studies have expanded on this research, cataloguing a richly complex history of human admixture, migration, and evolution. Despite this, many questions still remain about the extent and direction of admixture between archaic and modern humans and the roles selection and demography have played in the retention and loss of introgressed sequences.

Future progress will certainly depend on the recovery and analysis of additional archaic samples—Neanderthals, Denisovans, and others—and additional ancient human samples dated closer to the time of admixture. Data from more archaic samples [66,67] improve our understanding of Neanderthal and Denisovan genetic diversity, population structure, and the frequency of admixture events between these human lineages. For instance, the analysis of a

new, high-coverage Neanderthal genome from Vindija Cave [51] has improved estimates of the Neanderthal effective population size, determined the admixing Neanderthal population to be closer to the Vindija Neanderthal populations than the Altai one, and marginally increased estimates of Neanderthal ancestry in non-African populations outside Oceania to 1.8-2.6%. Analyzing older human samples, closer to the time of admixture, will be informative about the true initial level of admixture, as well as the rate at which archaic sequence was lost, and thereby provide insight into the mechanisms responsible for the loss and retention of archaic sequence in the modern human genome.

As well, analyses using improved methods to detect introgressed archaic sequence in existing data remain critical. We have seen how updated admixture statistics can reshape our view on historical archaic ancestry levels [38]. Importantly, methods that are less sensitive to complex demographic histories, like human-to-Neanderthal gene flow or population structure, will facilitate the cataloguing of archaic sequence in more geographically diverse modern populations. The unique histories of these populations may mean they carry distinct archaic introgressed haplotypes, and will aid in answering many of the outstanding questions in the study of modern and archaic hominin admixture.

1.9 Figures and Tables

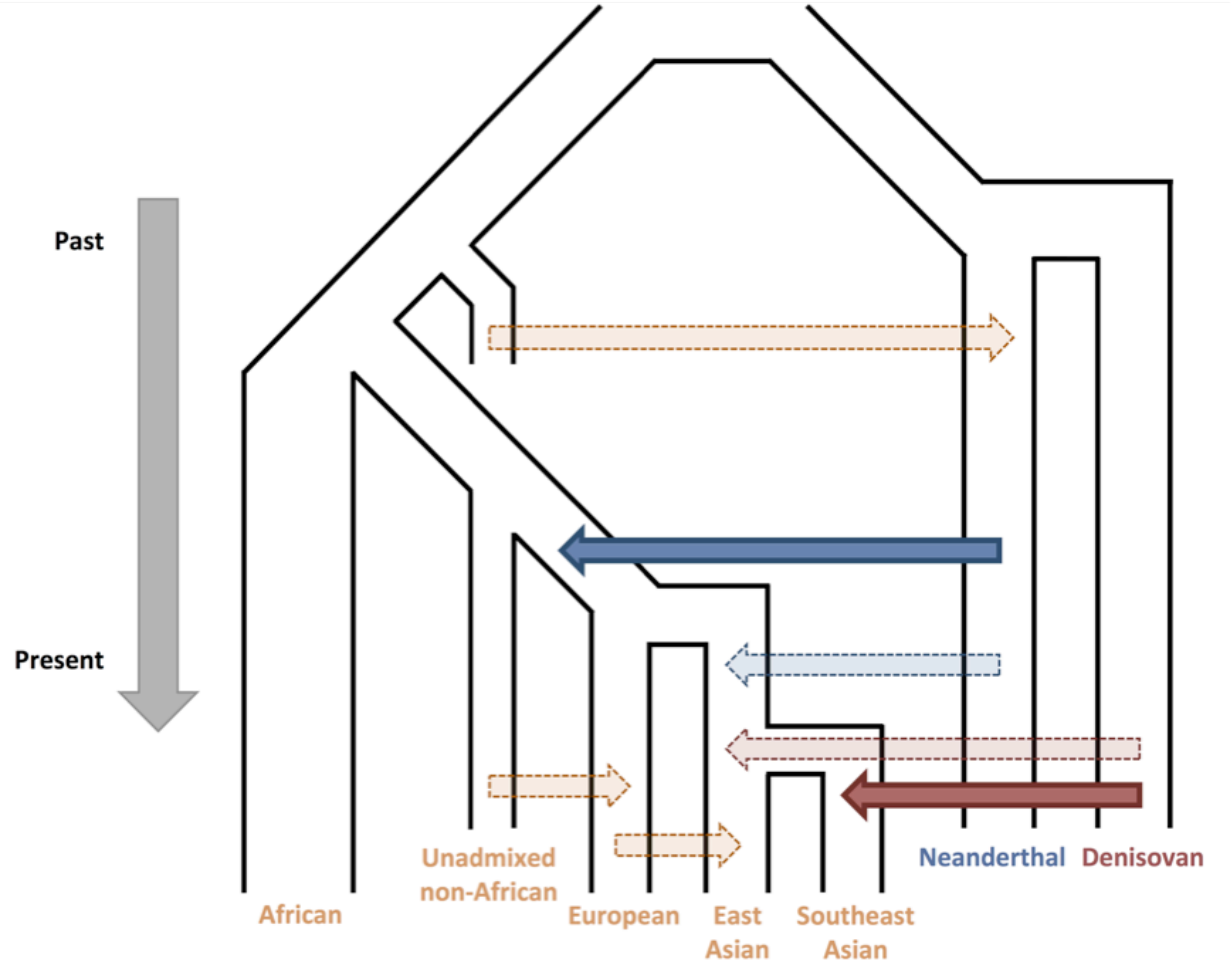


Figure 1.1. Simplified model of admixture history between archaic and anatomically modern human populations. There is consensus that at least two independent gene flow events occurred (solid arrows)—admixture from Neanderthals into an ancestral Eurasian population (solid-blue) and from Denisovans into an ancestral Southeast Asia population (solid-red). It is likely that additional instances of admixture occurred, explaining the variation in the percentage of archaic sequence across different global populations. These additional instances include a pulse of admixture from Neanderthals (dashed-blue) and from Denisovans (dashed-red) into an ancestral East Asian population. Alternatively, or in addition, global variation in archaic ancestry could be the result of admixture within human populations (dashed-orange) diluting archaic sequence. Admixture from human populations may also have introduced sequence into archaic populations.

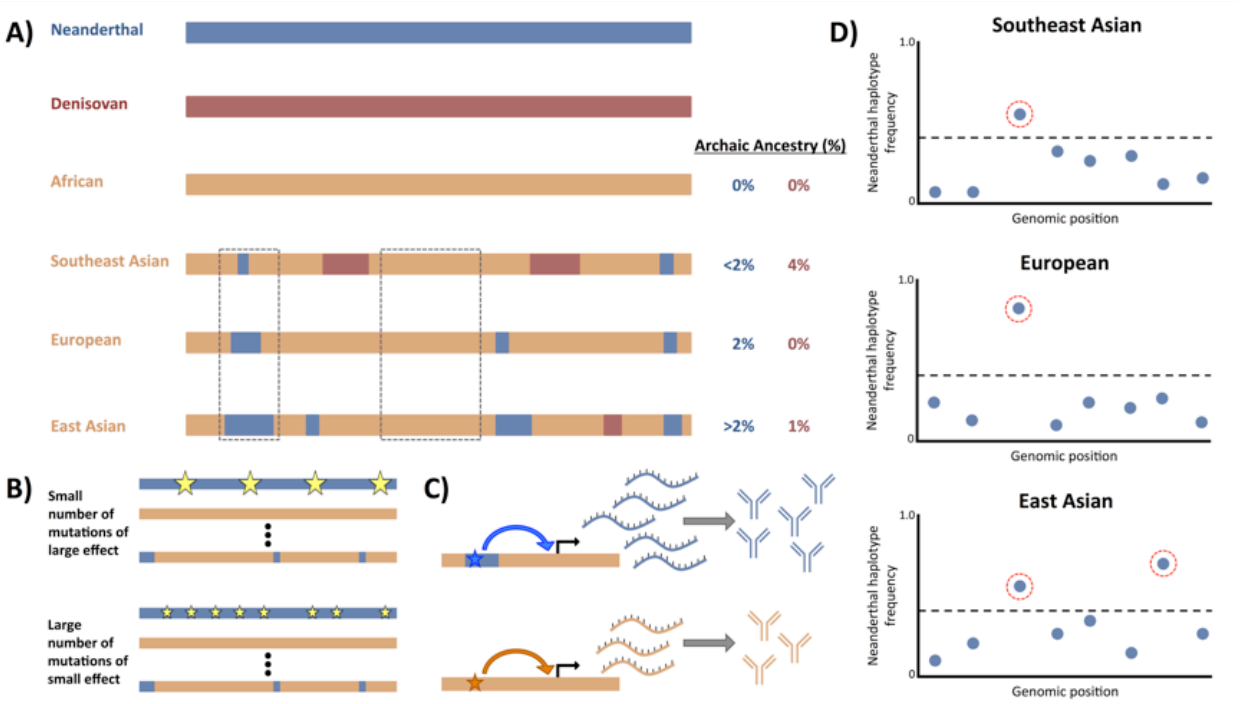


Figure 1.2. Patterns and characteristics of archaic sequence across the genome.
(A) A representation of individual genomes from archaic and modern human populations. The modern human genomes (orange) are ordered by increasing levels of Neanderthal (blue) admixture percentage (approximate). Only Asian populations carry Denisovan (red) sequence. Some introgressed archaic segments are shared across populations, and some large regions of the genome are depleted of introgressed archaic sequence in all populations examined.
(B) Large deserts may be a product of selection against deleterious archaic variants (gold stars) at those loci. Whether selection acted against a few strongly deleterious variants (top), or many weakly deleterious variants (bottom), remains uncertain.
(C) Many segments of introgressed archaic sequence are found to carry variants (stars) that affect gene regulation and expression. Altering gene expression may affect downstream protein levels (e.g. immunological proteins) and could have provided a mechanism of rapid adaptation for admixed modern humans.
(D) Putatively adaptive introgressed segments can be identified by examining the frequency of introgressed segments (blue dots) within a population and filtering for those that exceed a percentile cut-off (dashed black line).

Chapter 2. A Catalogue of Neanderthal Sequence in Global Populations Reveals A Heterogeneous Distribution Across the Genome

Parts of this chapter are adapted from:

> Vernot B, Tucci S, Kelso J, Schraiber JG, **Wolf AB**, Gittelman RM, et al. (2016) Excavating Neanderthal and Denisovan DNA from the genomes of Melanesian individuals. Science 352: 235–239

Recent studies have developed and applied methods to detect sequences in modern human genomes that were inherited from hybridization with Neanderthals and Denisovans [33,34]. Strikingly, the distribution of archaic sequence in the modern human genome is heterogeneous, with some large regions depleted of it. Regions depleted of archaic sequence, sometimes referred to as "deserts" of introgression, may represent loci where archaic sequence was strongly deleterious and rapidly purged from modern human populations. Alternatively, the stochastic loss of archaic sequences due to drift could also contribute to these archaic deserts. Understanding the formation and characteristics of archaic deserts in the modern human genome will help interpret how archaic admixture influenced human evolution and, possibly, what genes play a role in unique human behaviors.

We identified introgressed archaic hominin sequence in a geographically diverse data set of 503 European, 504 East Asian, 489 South Asian, [68] and 27 Melanesian individuals using the S* pipeline [33] and the Altai Neanderthal and Denisovan reference genomes [13,23]. This portion of the project was performed primarily by Ben Vernot.

As my contribution to the project, I leveraged this new dataset of identified archaic sequence to re-examine the distribution of archaic sequence across the genome. I characterized the heterogeneous distribution of archaic sequence as an empirical frequency of depletion. While introgressed archaic sequence appears throughout the modern human genome, several large regions are significantly depleted of it. The overlap of regions depleted of Neanderthal and Denisovan sequence is significantly greater than expected due to chance. To assess the significance of this distribution, and its likelihood under neutral demographic processes, I compared the empirical data to data from coalescent simulations of a wide variety of neutral

demographic models. I found that modern humans are significantly more enriched for large depletions than expected under neutral models.

The largest regions depleted of archaic sequence differ from the rest of the genome in several key characteristics, such as being significantly enriched for genes expressed in regions of the brain and differing in their levels of sequence diversity. The largest region depleted of archaic sequence contains the *FOXP2* gene, which is associated with speech and language and carries a regulatory change unique to modern humans.

2.1 Results

*2.1.1 Detected Archaic Sequence Using S**

Briefly, we first identified putative introgressed sequences using the statistic S* (which does not use information from an archaic reference genome) [33,69] and then refined this set by comparing significant S* haplotypes to the Neanderthal and Denisovan genomes and testing to determine whether they match more than expected by chance. Variation in neutral divergence between archaic groups across loci and incomplete lineage sorting complicate classification of archaic haplotypes as Neanderthal or Denisovan [30]. To address this issue, we developed a likelihood method that operates on the bivariate distribution of Neanderthal and Denisovan match p-values [30]. This framework estimates the proportion of Neanderthal, Denisovan, and null sequences in the set of S* significant haplotypes, identifies archaic haplotypes at a desired false discovery rate (FDR), and probabilistically categorizes them as Neanderthal, Denisovan, or ambiguous (i.e., Neanderthal or Denisovan status cannot be confidently distinguished).

In aggregate we analyzed 1523 non-African individuals, composed of 27 Melanesian individuals and 1496 European, East Asian, and South Asian individuals from the 1000 Genomes Project Phase 3 [68]. We recovered 1340 Mb and 304 Mb of the Neanderthal and Denisovan genomes, respectively, at a FDR = 5%. Melanesian individuals have on average 104 Mb of archaic sequences per individual (48.9, 42.9, and 12.2 Mb of Neanderthal, Denisovan, and ambiguous sequence, respectively) (**Figure 2.1**). In contrast, we only call between 0.026 Mb (in Esan) to 0.5 Mb (in Luhya) of sequences per individual as archaic in Africans, highlighting that our method and error rates are well calibrated. We identify on average 65.0, 55.2, and 51.2 Mb of archaic sequences in East Asians, South Asians, and Europeans, respectively (**Figure 2.1**).

Virtually all of the archaic sequences in these populations are Neanderthal in origin, although a small fraction (<1%) of introgressed sequences in East and South Asians are predicted to be Denisovan [30].

2.1.2 Quantifying Significant Depletions of Archaic Ancestry

The density of surviving Neanderthal sequences across the genome is heterogeneous **(Figure 2.2)** [33,34]**,** and regions that are strongly depleted of Neanderthal ancestry may represent loci where archaic sequences were deleterious in hybrid individuals and were purged from the population. In order to quantify levels of archaic depletion across the genome, I scanned our Neanderthal and Denisovan introgression callsets from 503 Europeans, 504 East Asians, 489 South Asians, and 27 Island Melanesians, in windows from 5Mb to 15Mb, with a 100kb step, and calculated the average amount of introgressed sequence in each window. I required that at least 90% of the sequence in a window be callable given the filters used in the primary analysis for identifying introgressed sequence [30]. I then counted, for each window size, the number of windows with an average percent introgression lower than $10^{-3.5}$ for Neanderthal and 10^{-4} for Denisovan (representing the lower 99th-percentiles) and present this as a proportion of the total windows tested (**Figure 2.3**).

2.1.3 Large Depletions Significantly More Frequent Than Expected Under Neutral Models

To measure how unusual Neanderthal depleted regions are under neutral models, I performed coalescent simulations **(Table 2.1),** focusing on individuals of European and East Asian ancestry whose demographic histories are known in most detail. Depletions of Neanderthal sequences that extend ≥8 Mb are significantly enriched in the observed compared with simulated data (permutation $P < 0.01$) **(Figure 2.3).** Neanderthal depleted regions that span at least 8 Mb are also significantly depleted of Neanderthal sequences in South Asians and Melanesians (Kolmogorov-Smirnov test, $p < 10^{-15}$). Based on these findings, I merged windows of 8Mb to 15Mb that were depleted of Neanderthal ancestry in EUR, EAS, and SAS populations to generate a final list of significantly depleted regions (**Table 2.2**).

If these desert regions do represent loci where strong selection against archaic sequence occurred, we hypothesized they would also be depleted of Denisovan ancestry. To test this, we measured the overlap of the Neanderthal set of desert windows with the comparable set of

Denisovan deserts to identify regions significantly depleted of both Neanderthal and Denisovan sequence (**Table 2.3**). Using a "sliding-genome" approach, we find significantly more overlap in regions depleted of Neanderthal and Denisovan lineages than expected by chance (permutation P = 0.0008) (**Figure 2.4**) [30], consistent with recurrent selection against archaic sequences.

2.1.4 Regions Depleted of Archaic Sequence Are Enriched For Neurodevelopmental Loci

Regions depleted of archaic lineages are also significantly enriched for genes expressed in specific brain regions, particularly in the developing cortex and adult striatum (permutation p < 0.05) (**Table 2.4**). A large region depleted of archaic sequence spans 11 Mb on chromosome 7 and contains the *FOXP2* gene **(Figure 2.5),** which has been associated with speech and language [54,70]. This region is also significantly enriched for genes associated with autism spectrum disorders (Fisher's exact test, p = 0.008) [30]. Another desert region on chromosome 3 also contains the pair of genes *ROBO1* and *ROBO2* (**Figure 2.5**), which are critical in the axon guidance toward and across the central nervous system midline [71–73], and are associated with sever dyslexia and language acquisition [74,75].

2.2 Discussion

We have demonstrated the wide-spread distribution of inherited Neanderthal and Denisovan sequence among geographically diverse populations. As genome-scale data from diverse populations continues to accumulate, a more complete catalogue of surviving archaic lineages can lead to a more robust "meta-genome", providing insight regarding the population level genetic diversity of these archaic lineages.

As well, detailed maps of archaic sequence in diverse human populations, and measures of the frequency of this archaic sequence, can provide novel insights into the evolutionary history of these human populations and the selection pressures they faced. Specifically, I have detailed our analysis of regions depleted of archaic ancestry in populations examined. Based on demographic simulations, I find the size and frequency of these desert regions to be inconsistent with neutral evolutionary processes. These regions may identify loci under strong selection for variants specific to the human lineage. Many of the variants at these desert loci are involved in

neurodevelopmental processes, and could be responsible for human behaviors that distinguish us from other archaic hominins.

Along with providing a more complete catalogue of global archaic ancestry, analyzing more geographically diverse populations will also aid in a better picture of the heterogeneous distribution of archaic ancestry across the genome. We have demonstrated how adding Melanesian samples and their inherited Denisovan ancestry has better defined regions significantly depleted of archaic ancestry. Including additional populations, and especially populations that may contain independent admixture events, will be instructive about the significance of these depleted regions. Specifically, African populations have been necessarily left out of this analysis because of limitations in the S* method for detecting loci of archaic ancestry. While African populations are expected to carry much less Neanderthal sequence, there is evidence of earlier admixture between African lineages and other African-specific archaic hominins [48,49]. In evaluating their evolutionary significance, it will be important to determine whether the desert loci are robust to multiple independent instances of admixture inside and outside of Africa.

Lastly, although our data show that large regions depleted of archaic ancestry are inconsistent with neutral evolution, mechanisms other than selection, such as structural variation, could also contribute to the appearance of archaic deserts by suppressing recombination at these loci. Based on the short read-length of sequenced archaic DNA, however, it will be extremely difficult to determine the frequency and location of large lineage-specific structural variants. Mapping features that contribute to structural variation, like segmental duplications, relative to large deserts could be helpful in exploring this mechanism. However, overlaps between these features would only provide circumstantial support for the role of structural variation. Additional work is therefore necessary to fully understand the origins of these desert regions.

2.3 Materials and Methods

*2.3.1 Identifying Introgressed Archaic Sequence Using S**

We extended our previously described framework to identify archaic sequences in the genomes of modern humans [33]. Specifically, we used a two-stage approach to first identify candidate introgressed sequences using the statistic S* [32,33] and then refined this set of

haplotypes by calculating a p-value to quantify whether a putatively introgressed haplotype matched an archaic sequence more than expected by chance.

S* is designed to detect highly divergent haplotypes whose variants are in strong linkage disequilibrium and are not found in a "reference" population, and to then maximize the S*-score by summing across variants in LD using an efficient dynamic programming algorithm. On average, introgressed haplotypes are expected to have an older TMRCA compared to non-introgressed lineages and therefore exhibit high levels of divergence. Because admixture occurred relatively recently, the introgressed haplotype will also tend to persist over large genomic regions (~50 kb in the case of Neanderthal introgression)[32]. Finally, because Neanderthal admixture is expected to have occurred only in non-African populations, variants on the introgressed haplotype should not be found in African individuals. We chose 107 Yoruba genomes as a reference population, as levels of Neanderthal variation in Yoruba are not statistically enriched (as measured by the D statistic), as opposed to Sandawe, Maasai, and African Americans [76].

The significance of S* scores is assigned based on a null distribution generated by simulating sequence data under a standard demographic model of non-African populations [33]. We retain putative introgressed haplotypes with an S* score in the 99th percentile of null simulations.

We now take the S* callset for each population, which is statistically enriched for archaic sequences but has not been compared to any archaic genome, and calculate archaic match p-values against both Neanderthal and Denisovan. To do this we generate an empirical distribution of the expected archaic match percentage in a population without archaic admixture using comparable Yoruba haplotypes. We used a likelihood method, which operates on the bivariate distribution of Neanderthal and Denisovan match p-values, to probabilistically assign them the labels of "Neanderthal", "Denisovan", or "Ambiguous" (archaic haplotypes that cannot be robustly distinguished as Neanderthal or Denisovan).

2.3.2 Coalescent Simulations of Admixture

To better understand the heterogeneous distribution of Neanderthal introgression, and the prevalence of significantly depleted regions of introgressed sequence, we performed extensive coalescent simulations. Specifically, the goal of these simulations was to test whether the large

windows depleted of introgressed sequence that we observe in the empirical data can be explained from demographic models of human history without invoking selection.

We used the coalescent simulator MaCS, which simulates genealogies spatially across chromosomes as a Markovian process [77]. Five demographic models, modified from previously published and accepted models, were used for simulating the demographic history of European and East Asian populations [78–81] **(Table 2.1).** In each model, an introgression event was added at 50 kya from a separate Neanderthal population into a larger Eurasia population, which subsequently split into Europeans and East Asians. The level of introgression was modified for each of the models so that modern sampled European and East Asian populations would retain 2% introgressed sequence. The Neanderthal population coalesced with the modern human population at 700 kya. For each model we simulated a sample of 503 European and 504 East Asian individuals (2,014 total haplotypes) over 1000 independent replicates of 15Mb of simulated sequence.

We then identified true introgressed haplotypes from reported coalescent trees. In empirical analyses, we are underpowered to identify short introgressed haplotypes and recover only a percentage of all introgressed bases. Therefore, we down-sampled simulated introgressed haplotypes to match the haplotype length distribution and mean percent introgression in our European and East Asian call sets. The mean percentage of introgressed bases in a 15Mb windows for the empirical data was 1.25%, while in simulated data it was ~2% for all models. The simulated call sets were down sampled by 63% to match the percentage of introgressed bases recovered in the empirical data. In addition, simulated introgressed haplotypes <15kb were down sampled to represent 0.57% of the total distribution, haplotypes that fell between 15kb and 30kb in size were down sampled to represent 5.6% of the total distribution, and haplotypes that fell between 30kb and 45kb were sampled to represent 20% of the total distribution.

After down sampling, we calculated the percentage of introgressed bases in windows of varying size (1- 15Mb) for the simulated data to compare with the empirical data for European and East Asian samples. When varying the window size, only one window per simulated chromosome was used so that each window represented an independent simulation of a given model. For example, having simulated 1000 iterations of 15Mb window, if we wanted to look at percent introgression in only a 1Mb window, we only took the first 1Mb of the original 15Mb simulation.

2.3.3 Comparison of Simulated Data to European and East Asian Neanderthal S Callsets*

We then scanned our Neanderthal introgression callset from 503 Europeans and 504 East Asians in windows from 1Mb to 15Mb, with a 100kb step, and similarly calculated the average amount of introgressed sequence in each window. We additionally require that at least 90% of each window is callable given the filters described in the primary analysis [82]. For this dataset and for the above simulations, we counted the number of windows with average percent introgression lower than $10^{-3.5}$ for each window size as a proportion of the total measured windows **(Figure 2.3).** Generally, we observed more depletions of Neanderthal sequence in real data than in simulations (**Figure 2.3**). To estimate significance, we resampled 5000 times from the simulated data, first correcting for the fact that we use sliding windows in real data by sampling a number of windows equal to the total size of the genomic regions dived by the window size. For example, for 10Mb windows we consider 2418.8Mb of genomic sequence, and we resampled 241 windows at a time from the simulated data: $2418.8\text{Mb} / 10\text{Mb}$. From these resamples, we calculated an empirical p-value for the significance of the observed number of depleted windows**.** At window sizes 8Mb and larger, we see significantly more windows depleted of Neanderthal sequence as compared to simulations.

We next identified regions 10Mb or larger and significantly depleted of Neanderthal sequence in Europeans, East Asians, South Asians and Melanesians, at a threshold of $10^{-3.5}$; these regions total 85.3Mb of the genome. We then compared these regions to patterns of Denisovan introgression in Island Melanesians, by identifying similar large depletions of Denisovan sequence. A complicating factor in this comparison is the relatively small number of Island Melanesian individuals. This results in many large regions with no identified Denisovan introgression, the largest of which is 21.8Mb, totaling 253Mb of the genome. However, by strictly considering windows with no Denisovan introgression, we would be unnecessarily splitting up large regions of depletion but with a small number of false positive calls. We therefore considered a threshold of 0.0001 average Denisovan introgression (2.1% of all windows 10Mb or larger are significant at this threshold) in regions of 10Mb or larger—totaling 356.6Mb of the genome. The 85.3Mb of Neanderthal depletion and 356.6Mb of Denisovan depletion overlap by 47.8Mb—over four regions of 10Mb or larger (**Table 2.3**). It is interesting to note that these overlaps are all larger than 10Mb, and not partial overlaps.

We next employed a "sliding genome" permutation algorithm to estimate the significance of the overlap between Neanderthal and Denisovan depletions. We first collapsed the genome by merging all chromosomes and removing uncallable regions, shifting the genomic positions of each Neanderthal and Denisovan depleted region appropriately. We advanced the positions of the Denisovan deserts by 1Mb steps, and for each step calculated the overlap between Neanderthal and Denisovan depletions (**Figure 2.4**). Using this distribution, we find that the 47.8Mb of overlapping Neanderthal and Denisovan depletions is significantly larger than random overlaps (empirical p-value = 0.0008).

2.4 Figures and Tables

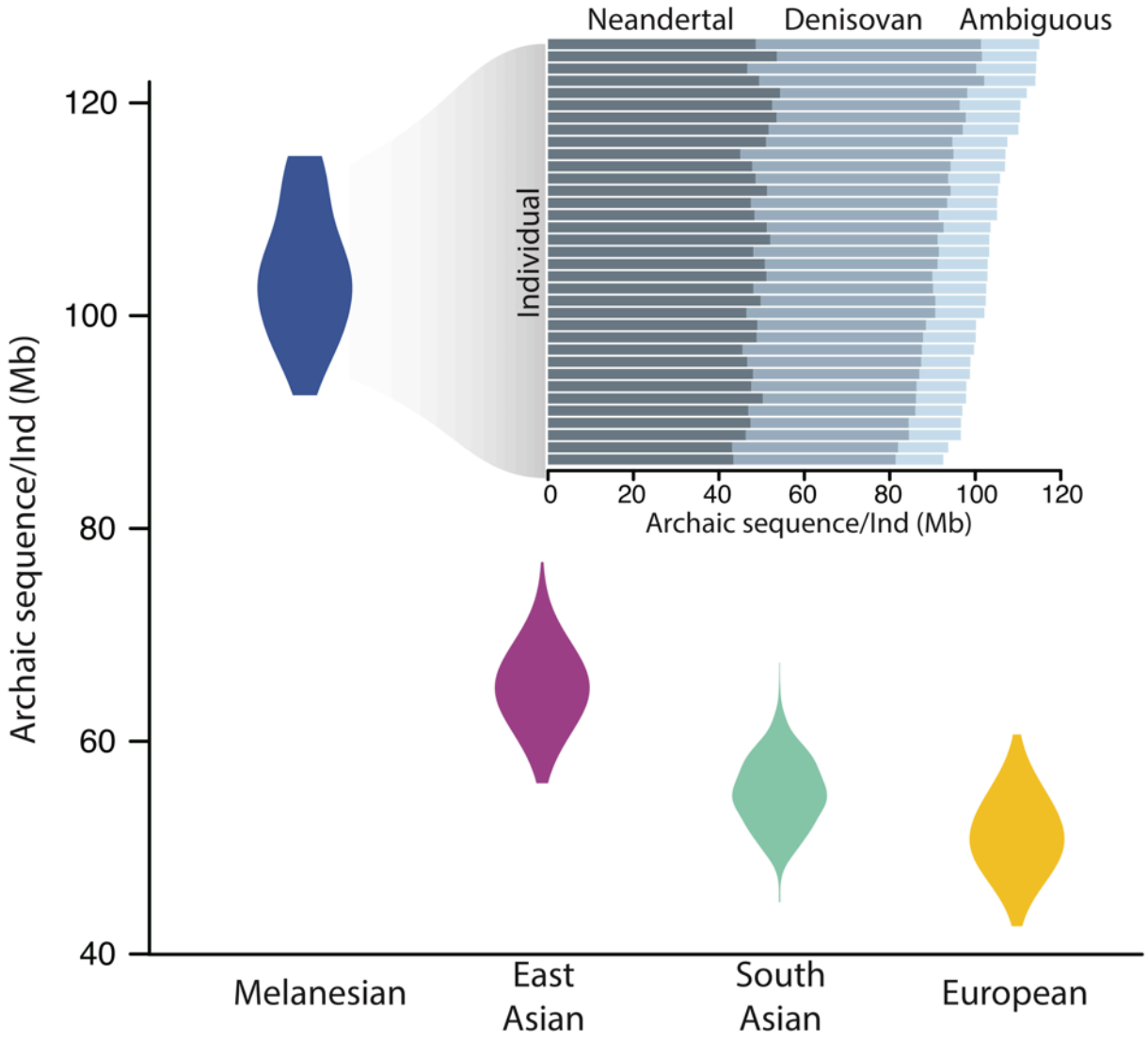


Figure 2.1. Amount of archaic introgressed sequence identified in each analyzed population. Inset: Amount of Neanderthal, Denisovan, and ambiguous (Neanderthal or Denisovan) introgressed sequence for each Melanesian individual.

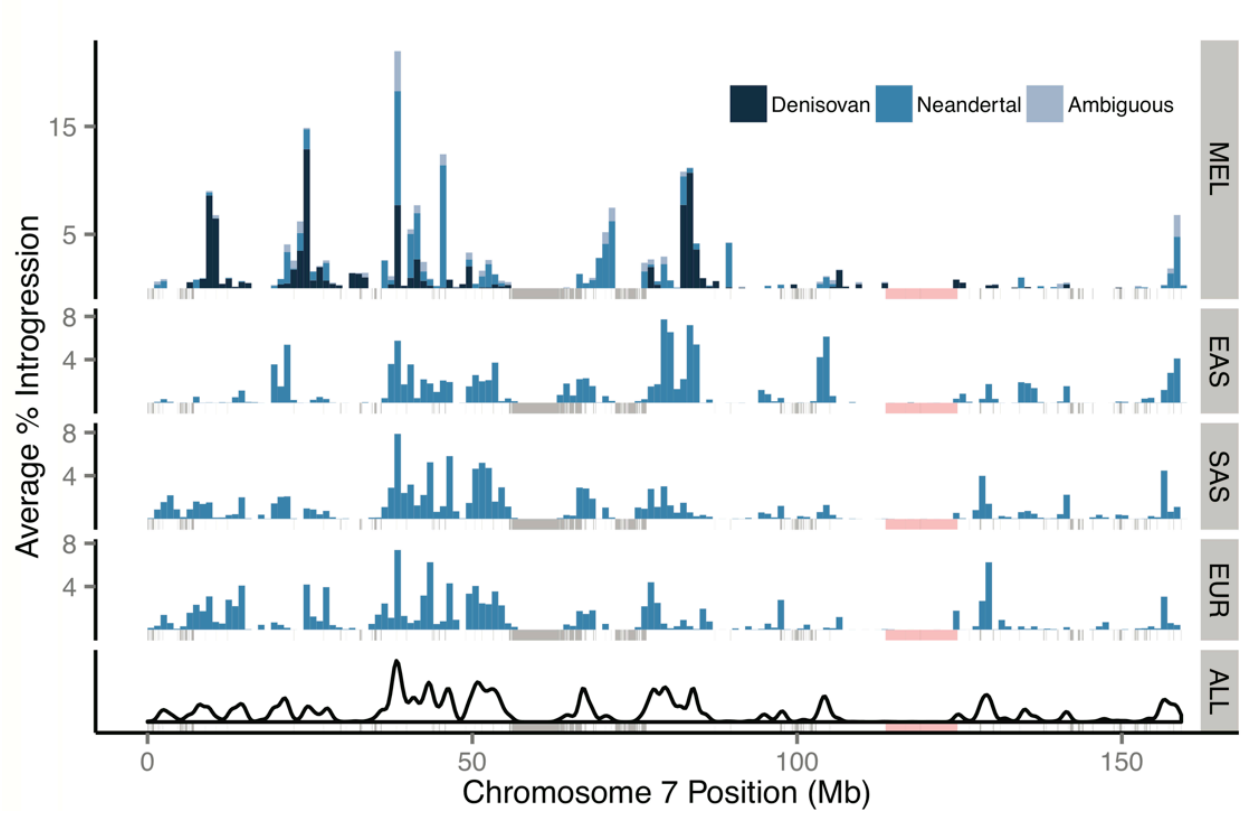


Figure 2.2. **Heterogeneous distribution of Neanderthal and Denisovan sequence across chromosome 7.** Visualization of the amount of archaic ancestry identified in Melanesians (MEL), East Asians (EAS), South Asians (SAS), and Europeans (EUR), and then summarized across all populations (ALL). Masked regions are shown as grey bars. An 11.1 Mb region significantly depleted of Denisovan and Neanderthal ancestry in all populations is show in light pink.

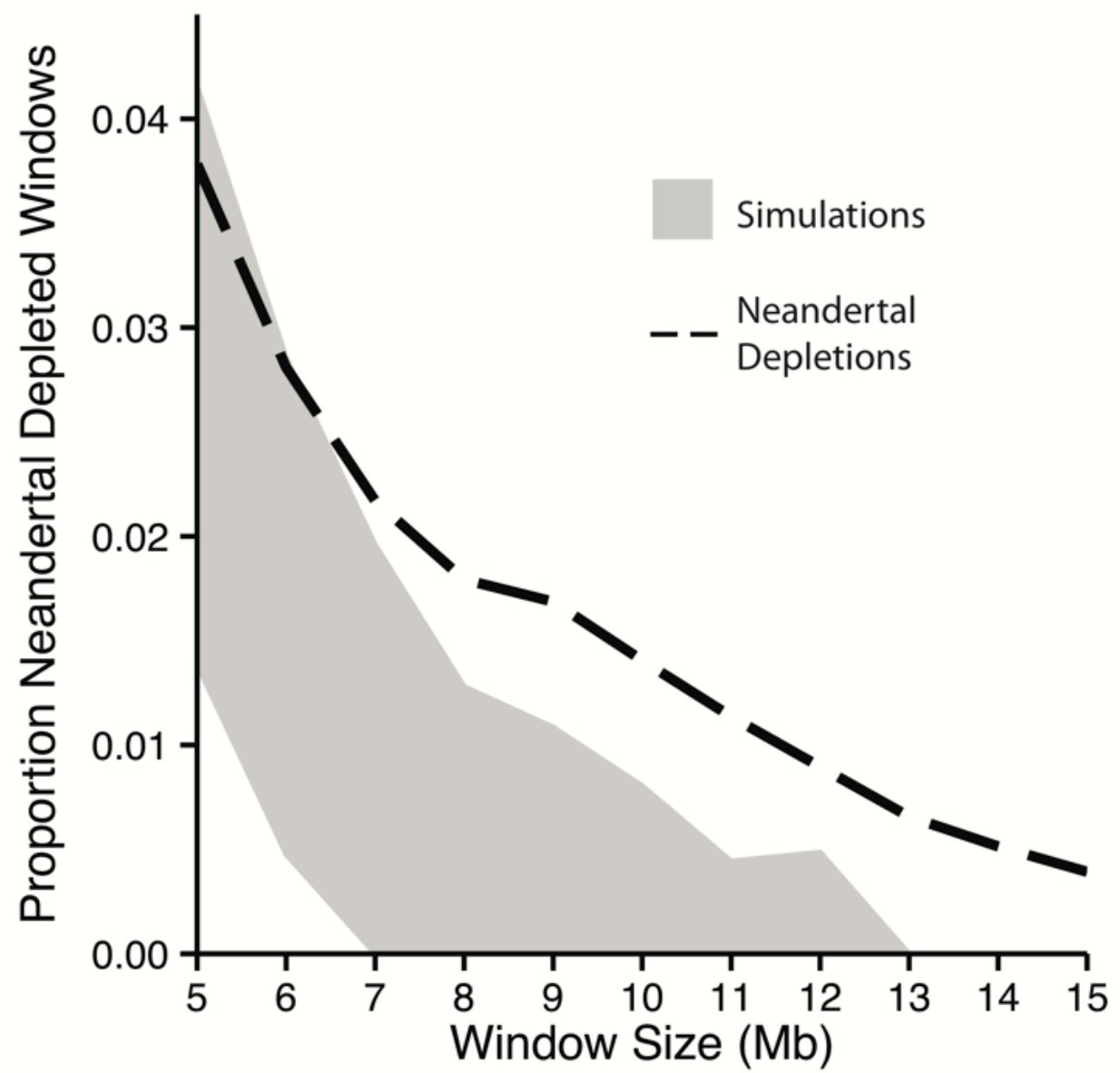


Figure 2.3. **Proportion of windows significantly depleted of Neanderthal introgression in Europeans and East Asians.** The dashed line represents the empirical values versus what is expected in neutral demographic models (95% confidence interval in grey).

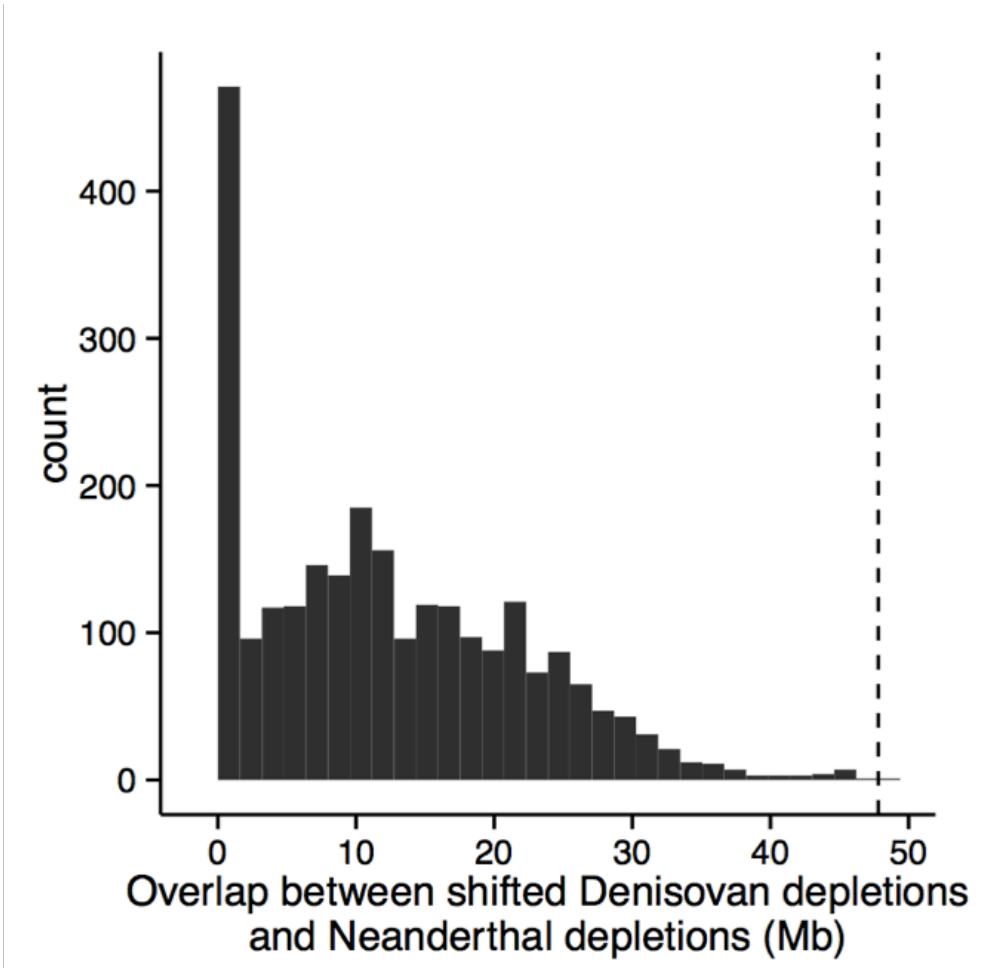


Figure 2.4. Neanderthal and Denisovan overlap in randomized archaic deserts. Distribution of the amount of overlap between Neanderthal and Denisovan deserts when randomized by shifting the location of Denisovan deserts by 1Mb along the length of the genome. This distribution is significantly below the observed overlap between these deserts, shown as a dotted line.

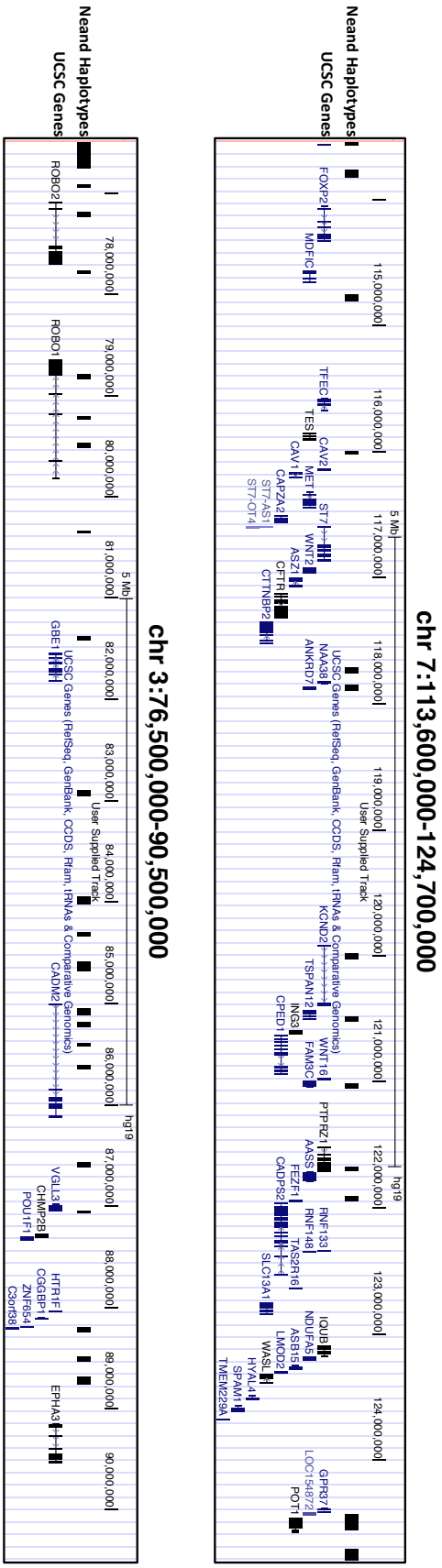


Figure 2.5. Desert regions on chr7 and chr3 contain genes *FOXP2* and *ROBO1/2*. Loci where Neanderthal haplotypes were called are demarcated on top and UCSC genes on bottom.

Table 2.1. Coalescent simulations for five standard demographic models: MaCS commands

Tennessen model [78]:
macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 58.00274 -n 2 70.041 -n 3 187.55 -eg 0.9e-10 1 482.46 -eg 1.0e-10 2 570.18 -eg 1.1e-10 3 720.23 -em 1.2e-10 1 2 0.731 -em 1.3e-10 2 1 0.731 -em 1.4e-10 3 1 0.2281 -em 1.5e-10 1 3 0.2281 -em 1.6e-10 2 3 0.9094 -em 1.7e-10 3 2 0.9094 -eg 0.007 1 0 -en 0.007001 1 1.98 -eg 0.007002 2 89.7668 -eg 0.007003 3 113.3896 -eG 0.031456 0 -en 0.031457 2 0.1412 -en 0.031458 3 0.07579 -eM 0.031459 0 -ej 0.03146 3 2 -en 0.0314601 2 0.2546 -em 0.0314602 2 1 4.386 -em 0.0314603 1 2 4.386 -eM 0.0697669 0 -ej 0.069767 2 1 -en 0.0697671 1 1.98 -en 0.2025 1 1 -ej 0.9575923 4 1 -em 0.06765 2 4 32 -em 0.06840 2 4 0

Gravel low coverage + exon model [79]:
macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 2.12 -n 2 4.911 -n 3 6.703 -eg 1.0e-10 2 111.11 -eg 1.1e-10 3 140.35 -em 1.2e-10 1 2 0.731 -em 1.3e-10 2 1 0.731 -em 1.4e-10 3 1 0.228 -em 1.5e-10 1 3 0.228 -em 1.6e-10 2 3 0.9094 -em 1.7e-10 3 2 0.9094 -eG 0.031456 0 -en 0.031457 2 0.1412 -en 0.031458 3 0.07579 -eM 0.031459 0 -ej 0.03146 3 2 -en 0.0314601 2 0.2546 -em 0.0314602 2 1 4.386 -em 0.314603 1 2 4.386 -eM 0.0697669 0 -ej 0.069767 2 1 -en 0.0697671 1 1.98 -en 0.2025 1 1 -ej 0.9575923 4 1 -em 0.06765 2 4 32 -em 0.06840 2 4 0

Gravel low coverage model [79]:
macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 2.12 -n 2 4.911 -n 3 6.703 -eg 1.0e-10 2 78.95 -eg 1.1e-10 3 90.64 -em 1.2e-10 1 2 0.491 -em 1.3e-10 2 1 0.491 -em 1.4e-10 3 1 0.1696 -em 1.5e-10 1 3 0.1696 -em 1.6e-10 2 3 1.725 -em 1.7e-10 3 2 1.725 -eG 0.03826 0 -en 0.03827 2 0.2216 -en 0.03828 3 0.1123 -eM 0.03829 0 -ej 0.03830 3 2 -en 0.03831 2 0.3773 -em 0.03832 2 1 5.848 -em 0.03833 1 2 5.848 -eM 0.1340 0 -ej 0.1341 2 1 -en 0.1342 1 2.105 -en 0.4322 1 1 -ej 0.9575923 4 1 -em 0.06765 2 4 32 -em 0.06840 2 4 0

Gutenkunst model [80]:
macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00069 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 1.685 -n 2 4.4 -n 3 8.6 -eg 1.0e-10 2 116.8 -eg 1.1e-10 3 160.6 -em 1.2e-10 1 2 0.876 -em 1.3e-10 2 1 0.876 -em 1.4e-10 3 1 0.5548 -em 1.5e-10 1 3 0.5548 -em 1.6e-10 2 3 2.8032 -em 1.7e-10 3 2 .8032 -eG 0.0290 0 -en 0.02901 2 0.1370 -en 0.02902 3 0.06986 -eM 0.02903 0 -ej 0.02904 3 2 -en 0.0290401 2 0.2877 -em 0.0290402 2 1 7.3 -em 0.0290403 1 2 7.3 -eM 0.19149 0 -ej 0.1915 2 1 -en 0.191501 1 1.685 -en 0.3014 1 1 -ej 0.9575923 4 1 -em 0.06774 2 4 34 -em 0.06849 2 4 0

Schaffner model [81]:
macs 2025 15000000 -s ${RANDOM}${SGE_TASK_ID} -i 10 -r 3.0e-04 -t 0.00075 -T -I 4 10 1006 1008 1 0 -n 4 0.205 -n 1 8 -n 2 8 -n 3 8 -em 1.2e-10 1 2 1.6 -em 1.3e-10 2 1 1.6 -em 1.4e- 10 3 1 0.4 -em 1.5e-10 1 3 0.4 -en 0.004 1 1.92 -en 0.007 2 0.616 -en 0.008 3 0.616 -en 0.03942 2 0.0574 -en 0.03998 2 0.616 -en 0.038 3 0.058 -en 0.03997 3 0.616 -eM 0.03999 0 -ej 0.040 3 2 -en 0.04001 2 0.616 -en 0.0686 2 0.032 -en 0.0696 1 0.0996 -ej 0.07 2 1 -en 0.07001 1 1.92 -en 0.34 1 1 -ej 0.56 4 1 -em 0.04002 2 4 29 -em 0.04077 2 4 0

Table 2.2. Regions significantly depleted of Neanderthal introgressed sequence in European, East Asian, and South Asian populations.

Regions significantly depleted of Neanderthal sequence in all populations			
chromosome	start	end	len_Mb
1	102200000	114900000	12.7
2	201100000	211500000	10.4
3	76500000	90500000	14.0
7	106300000	124700000	18.4
8	53900000	66000000	12.1
18	25000000	41800000	16.8

Table 2.3. Regions significantly depleted of Neanderthal introgressed sequence in European, East Asian, and South Asian populations and Denisovan sequence in Melanesian individuals.

Regions significantly depleted of both Neanderthal and Denisovan sequence			
chromosome	start	end	len_Mb
1	104000000	114900000	10.9
3	76500000	90500000	14.0
7	113600000	124700000	11.1
8	54500000	65400000	10.9

Table 2.4. GO enrichments of genes in regions depleted of archaic sequence.

GO ID	GO Term	*p-value*	FWER	Genes
GO:0004556	alpha-amylase activity	1.60E-09	<0.001	*AMY1C,AMY1A,AMY2BA MY2A,AMY1B*
GO:0016160	amylase activity	9.20E-09	<0.001	*AMY1C,AMY1A,AMY2BA MY2A,AMY1B*
GO:0004553	hydrolase activity, hydrolyzing O-glycosyl compounds	9.30E-09	<0.001	*MGEA5,AMY1C,AMY1A,A MY2B,HYAL4, AMY2A,OVGP1,SPAM1CH IA,GBE1, CHI3L2,AMY1B*
GO:0004364	glutathione transferase activity	1.90E-08	<0.001	*GSTO2,GSTM4,GSTM2,GS TO1,GSTM1, GSTM3,GSTM5*
GO:0016798	hydrolase activity, acting on glycosyl bonds	1.80E-07	<0.001	*MGEA5,AMY1C,AMY1AA MY2B,HYAL4, AMY2A,OVGP1,SPAM1CH IA,GBE1, CHI3L2,AMY1B*
GO:0015271	outward rectifier potassium channel activity	6.60E-07	<0.001	*KCNIP2,KCNA3,KCND3,KC NA2,KCND2*
GO:0005250	A-type (transient outward) potassium channel activity	5.30E-06	<0.001	*KCNIP2,KCND3,KCND2*
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	7.20E-06	0.001	*GSTO2,GSTM4,GSTM2,GS TO1,GSTM1, GSTM3,GSTM5*
GO:0048593	camera-type eye morphogenesis	2.00E-05	0.035	*RP1,PAX2,TSPAN12, WNT2B,HIPK1,PITX3, WNT16,GNAT2,WNT2*
GO:0005109	frizzled binding	5.00E-05	0.012	*MAGI3,WNT2B,SDCBP,W NT16,WNT2*
GO:0008076	voltage-gated potassium channel complex	7.10E-05	0.022	*KCNC4,KCNIP2,KCNA3,KC ND3,KCNA10, KCNA2,KCND2*
GO:0034705	potassium channel complex	7.10E-05	0.022	*KCNC4,KCNIP2,KCNA3,KC ND3,KCNA10, KCNA2,KCND2*

Chapter 3. A Reference Free Method Identifies Neanderthal Ancestry in Africans

Parts of this chapter are adapted from:

> Chen L[*], **Wolf AB**[*], Wenqing F, Akey JM. (2019) A reference free method to identify introgressed hominin sequences reveals insights into Neanderthal ancestry in African and non-African populations. Cell (*in press)*

Studies of ancient DNA are transforming our understanding of human evolutionary history, and in particular how admixture has shaped past and present patterns of human genomic variation [25,26,83,84]. Of particular interest has been the discovery that admixture with archaic hominins occurred multiple times throughout human history [12,13,22,23]. In particular, approximately 2% of all non-African ancestry is derived from Neanderthals [13,22,23,30–32], with Oceanic populations having an additional 2-4% of ancestry attributable to gene flow with Denisovans [30,31,85,86].

A consistent observation in all studies of archaic hominin admixture is that East Asian populations have approximately 20% more Neanderthal ancestry compared to Europeans [30–34,83]. Numerous models have been invoked to explain this difference, including the interaction of demography and selection [34,39,87], dilution by non-admixed populations [13,39], or additional population-specific admixture events [41,87,88]. Understanding variation in Neanderthal ancestry among non-African populations has important implications for refining our understanding of admixture between modern human ancestors and Neanderthals.

Moreover, the ability to identify introgressed hominin sequence in the genomes of modern humans enables inferences about the functional, evolutionary, and phenotypic significance of archaic admixture. There are examples both of strong selection against Neanderthal and Denisovan sequence in the human genome, and examples of beneficial sequences driven to high frequency in modern human populations.

We describe a novel method for detecting Neanderthal ancestry in modern humans, called IBDmix, which does not require an unadmixed reference human panel. We apply IBDmix to genotype data from a large set of modern human individuals from Eurasia, America, and Africa, and make novel discoveries regarding Neanderthal sequence in Africans and re-examine

the relative levels of Neanderthal ancestry in Eurasian populations. We also replicate, extend, and discover new instances of adaptive introgression that may offer insight into human evolution and phenotypic variation in modern humans.

W. Fu derived the analytical theory and wrote the software for IBDmix. W. Fu and L. Chen developed the method. A. Wolf and L. Chen conducted the analyses. A. Wolf, L. Chen, and J. Akey wrote the manuscript.

3.1 Results

3.1.1 Evaluating the Power and Robustness of IBDmix

Methods that identify introgressed Neanderthal lineages in modern humans must differentiate between sequences shared with Neanderthals because of ancient hybridization or because of a shared common ancestor. Previous approaches, such as S* [33,69], CRF [34], diCal-admix [89], and HMM [90], use an "unadmixed" modern reference panel, commonly an African population such as Yoruba (YRI), to control for false positives due to shared ancestry by "masking" putative archaic sequence present in the reference panel and the target sample. If the reference panel carries introgressed Neanderthal sequence, this will result in missing Neanderthal sequence in the target sample (**Figure 3.1A**).

Our new method, IBDmix, does not use a modern reference panel (**Figure 3.1A**). Instead, IBDmix calculates the probabilities that a variant site in an individual is and is not shared identical by descent (IBD) with a reference archaic genome, while accounting for genotyping errors in the reference archaic and modern human sequences. The ratio of these probabilities is used to construct a single site LOD score, where higher values indicate a greater likelihood an individual's genotype is shared IBD with the reference archaic genome. IBDmix then uses a dynamic programming algorithm to sum together single-site LOD scores and maximize this score in order to identify introgressed segments.

We evaluated IBDmix's performance and operating characteristics using simulated data generated from an accepted demographic model (**Figure 3.2**) [91], and compared it to results using *S**. The false positive rate for IBDmix is controlled by the LOD score threshold and introgressed segment size cutoff. Specifically, for introgressed segment sizes >30kb, the power of IBDmix is >60% with a FDR ≤ 10% (**Figure 3.1B**). Note, the power and FDR of IBDmix in

non-African populations is not influenced gene-flow from non-Africans into Africans, whereas it has a large effect on S* (**Figure 3.1B**). The power to detect introgressed sequence in non-African populations is particularly low for S* when this sequence is also found in the reference population (Africans), whereas IBDmix maintains power (**Figure 3.1C**). This observation implies that biases may arise in methods that use a modern human reference panel, as the power to detect introgressed sequence will be a function of whether it is or is not found in the reference panel.

We also tested the impact of genetic variation and recombination rate on IBDmix using simulated data. The performance of IBDmix improved overall with increasing mutation rates (**Figure 3.3**), We observed a noticeable improvement for shorter segments (both FPR, FDR, and power). This is not surprising since higher genetic variation provides more evidence of identity by descent, and would enhance the IBDmix LOD score. In testing the effect of recombination rate on IBDmix performance, we used data generated from a model with no Neanderthal introgression. We evaluated FPR of IBDmix under models with a recombination rate equal to the genome-wide average (1cM/Mb), and models $1/10^{th}$ that rate (0.1cM/Mb). We observed more false positives in situations with a lower recombination rate (**Table 3.1**).

Previous studies have identified the introgressing Neanderthal population as a sister clade of the sequenced Altai Neanderthal[42,51]. We therefore tested how IBDmix would perform when the reference archaic genome is distantly related to the introgressing archaic. We simulated models with two Neanderthal lineages representing an introgressing lineage and a sampled reference lineage (non-introgressing lineage), and varied the split time between these two populations. We observed a small decrease in performance using the non-introgressing Neanderthal as the reference genome, but overall performance measures remained consistent. Predictably, performance became more comparable as the divergence time between these two lineages decreased (**Figure 3.4**).

In summary, IBDmix has higher power and lower FDR compared to the widely used existing method S*, and is robust to reference population biases. In the following, unless otherwise noted, we used a LOD score threshold of 4 and a minimum segment size of 50kb, which provides a reasonable tradeoff between power and false positive rate (**Figure 3.1B**).

3.1.2 IBDmix Reveals Substantial Amounts of Neanderthal Ancestry in Africans and Nearly Uniform Levels in Non-African Populations

We applied IBDmix to samples from the 1000 Genomes Project [68], collected from geographically diverse populations, and used the Altai Neanderthal reference genome [23] to identify introgressed Neanderthal sequence in these individuals. After filtering we identified 110.98 Gb of Neanderthal sequence among 2,504 modern individuals. When overlapping introgressed segments are merged, this equates to 1.29 Gb of unique Neanderthal sequence.

Because IBDmix does not use a putatively unadmixed modern reference population, we were able to robustly identify regions of Neanderthal ancestry in African populations for the first time **(Figure 3.5A).** Surprisingly, we identified on average 17 Mb of Neanderthal sequence per individual in the African samples analyzed, and this value was consistent across the mostly northern African subpopulations represented in the dataset (ranging from 16.4 Mb/individual in ESN to 18.0 Mb/individual in LWK; **Figure 3.5A; Table 3.2**). Furthermore, we observed a significant overlap of sequence identified in Africans with that in non-Africans (**Figure 3.5B**). Specifically, of the Neanderthal sequence identified in African samples, more than 94% was shared with non-Africans.

We also recovered a substantial amount of Neanderthal sequence in non-African samples across populations. Notably, we found similar levels of Neanderthal ancestry in Europeans (51 Mb/individual), East Asians (55 Mb/individual) and South Asians (55 Mb/individual) (**Figure 3.5A; Table 3.2**). Strikingly, we observed only a modest enrichment (8%) of Neanderthal ancestry in East Asian compared to European individuals. This contrasts with previous reports that have indicated ~20% enrichment of Neanderthal ancestry in East Asians compared to Europeans [31–34]. The observed level of East Asian enrichment was even smaller (~3%) when we were less conservative in our filtering methods (**Table 3.3**).

3.1.3 Neanderthal Haplotypes in Africans Derived from Historical Back-Migration with Non-Africans and Pre-Out Of Africa Human to Neanderthal Gene-Flow

Given the unexpectedly large amounts of Neanderthal sequence identified in African individuals, we sought to confirm that these were actual introgressed lineages (as opposed to false positives from shared ancestry) and to understand their origins. To rule out systematic biases, we also called Denisovan sequence in African individuals using IBDmix and only

identified 1.2 Mb/individual of Denisovan sequence in African samples (**Table 3.4**). Notably, the amount of Neanderthal sequence called in African individuals is considerably higher than the amount of Denisovan sequence identified by IBDmix. As well, preliminary testing with simulated data showed that the IBDmix signal for Neanderthal ancestry in Africans could not come from incomplete lineage sorting (ILS) alone, but was consistent with either non-African to African gene flow, or pre-Out of Africa (pre-OOA) human to Neanderthal gene-flow (**Figure 3.5C**) [92].

We therefore explicitly tested whether Neanderthal sequences identified in Africans were more likely to be explained by incomplete lineage sorting (ILS) or were introduced into Africans by back-migration of admixed non-Africans, or by pre-OOA human to Neanderthal gene-flow [51,92]. To differentiate between scenarios of ILS or back-migration, we compared the empirical data to simulated data, analyzing a variety of sequence characteristics **(Figure 3.6).** Specifically, we simulated genotype data under a series of demographic models that included Neanderthal admixture into non-Africans, increasing levels of back-migration from Europeans into Africans, and gene flow from a pre-OOA human lineage into Neanderthals at varying time points. We then identified introgressed sequence for these models using IBDmix. Under the null model, with no back-migration, any introgressed segments identified in Africans represent ILS. Models that include back-migration produce a combination of ILS and true-introgressed segments. Models with human to Neanderthal gene flow produce alternative false-positives distinct from those due to ILS. We compared the empirical and simulated data across features including introgressed segment length, frequency of introgressed segments in the African population, and the ratio of East Asian Neanderthal ancestry to European Neanderthal ancestry before and after masking Neanderthal sequence shared between Africans and non-Africans.

In the empirical data, segments identified in Africans (YRI) that are shared with non-Africans (EAS and EUR) have a distribution of segment sizes more similar to that of non-African calls, although with some enrichment of smaller sized segments, and also occur predominantly at high-frequency (>10%) in the African population **(Figure 3.6).** As noted previously, there is only a small enrichment (<10%) for Neanderthal ancestry in East Asians compared to Europeans without masking sequence shared with Africans. But, when shared sequence is masked, this enrichment increases to ~18% **(Figure 3.6).**

These features are not replicated in either models with back-migration or human to Neanderthal gene flow alone. While features like the distribution of segment lengths and the frequency of African segments in the African population are replicated in models with human to Neanderthal gene flow occurring at 100kya, only models with elevated back-migration ($5x10^{-5}$/generation) can replicate the enrichment of East Asian Neanderthal ancestry when masking shared African sequence. A model that combines both of these demographic events, elevated back migration and human to Neanderthal gene-flow, matches the empirical data best across all the features of the data. In summary, these data indicate that both pre-OOA human to Neanderthal gene-flow and elevated historic back-migration contribute to the archaic sequence identified in Africans.

To further confirm the role of back-migration in introducing Neanderthal sequence into African populations, we examined the rate of overlap between called Neanderthal segments and non-African ancestry tracks in African samples (**Figure 3.7A**). We hypothesized that if the Neanderthal sequence in Africans was introduced by back-migration from ancestors of contemporary Europeans, then there should be enrichment for overlap of Neanderthal segments and European ancestry segments in African samples. To test this hypothesis, we compared data from chromosome 1 for all 504 African samples in our analysis. For each individual, we identified tracks of European and East Asian ancestry using RFMix [93], and measured the rate of overlap with identified Neanderthal segments in the same individual (**Figure 3.7A**). We averaged these rates of overlap to calculate empirical rates of overlap for European ancestry and East Asian ancestry separately (**Figure 3.7B**). We found the rate of overlap with European ancestry to be highly significant (permutation $p < 0.0001$), while the rate of overlap with East Asian ancestry was not (permutation $p > 0.05$). These data are consistent with the hypothesis that Neanderthal segments in Africans are derived from back-migration. Furthermore, the data indicate that this back-migration came after the split of Europeans and East Asians, from a population related to the European lineage, which complement findings from simulations regarding the effect of back-migration on biases in S* ancestry estimates (**Figure 3.8**).

3.1.4 Previously Inferred Differences in Neanderthal Ancestry Between East Asians and Europeans Were Biased due to Unaccounted for Back-Migration

Previous methods that have relied on un-admixed modern reference populations, like S*, have reported >20% enrichment of Neanderthal sequence in East Asians compared to Europeans (**Figure 3.9A**). However, results from IBDmix show only 8% enrichment of Neanderthal sequence in East Asians compared to Europeans (**Figure 3.9A**). This level of enrichment is robust to changes in segment size cutoff (30kb, 40kb, 50kb) used for IBDmix calling (**Table 3.3**). To better understand the discrepancy between IBDmix and previous inferences, we first removed Neanderthal sequence called by IBDmix in Europeans and East Asians that was shared with Africans (YRI), and replicated an 18% enrichment of Neanderthal ancestry in East Asians compared to Europeans (**Figure 3.9A**). This result shows that our observation of similar levels of Neanderthal ancestry in Europeans and East Asians is due to no longer masking Neanderthal sequence shared with Africans.

In the IBDmix callset for Africans, Europeans, and East Asians, there is a large enrichment of Neanderthal sequence shared exclusively between Africans and Europeans compared with sequence shared exclusively between Africans and East Asians (**Figure 3.9B**). As a proportion of the total amount of Neanderthal sequence for each population, 7.2% of European sequence is shared exclusively with Africans, which is a substantially higher than the 2% of East Asian sequence shared exclusively with Africans (**Figure 3.9B**). This imbalance in the proportion of exclusively shared sequence between African and non-African populations directly contributes to the biased Neanderthal ancestry estimates in previous methods that use an African reference panel to mask shared sequence.

We also examined how the reference panel size for S*, and by extension the diversity of the panel, affects Neanderthal ancestry estimates. We bootstrap resampled Yoruba individuals from the 1000 Genome Project data to generate new reference panels and re-called Neanderthal sequence for European and East Asian individuals using the S*-pipeline. We compared the total S*-sequence called for each sample to the average amount of S*-sequence called for samples using a reference panel size of 1 (**Figure 3.9C)**. Increasing the reference panel size showed a significant reduction ($p < 2x10^{-16}$) in the amount of Neanderthal sequence called per individual. In addition, when comparing the amounts of Neanderthal sequence identified in Europeans and East Asians, increasing the reference panel size decreased the amount detected for both populations, but there was a greater loss in Europeans than East Asians. Using a reference sample size larger than 10 led to an apparent 20% enrichment of Neanderthal ancestry in East

Asians compared to Europeans, as previously reported (**Figure 3.9C**). Simulations of European to African back-migration using rates consistent with standard demographic models also generate a significant enrichment of Neanderthal ancestry in East Asians compared to Europeans when the data are analyzed with *S** ($p < 8x10-7$; **Figure 3.8**).

Collectively, these results show that Neanderthal ancestry estimates in East Asians and Europeans have been biased due to unaccounted for back-migrations from European ancestors into Africans.

3.1.5 IBDmix Reveals Novel Insights into Signatures of Adaptive Introgression

Admixture with Neanderthals may have provided a mechanism for modern humans to acquire novel adaptive variation. Previous analyses have reported population-specific high-frequency introgressed Neanderthal haplotypes, which may be instances of adaptive introgression [64,94,63,56]. We examined our IBDmix callset for similar findings. We leveraged population-level derived allele frequencies of variants that overlapped calls made by IBDmix and matched the Neanderthal allele, in order to detect Neanderthal haplotypes with unusually large differences in frequency between populations.

Specifically, for variants that intersected identified Neanderthal segments, we calculated the differences in the derived allele frequencies between Europeans and East Asians, Africans and Europeans, and Africans and East Asians. We then took an outlier approach to identify loci with allele frequency differences in the 99th percentile. We further filtered on loci where the derived allele matched the Neanderthal allele. Overall, we identified 38 non-African specific high-frequency haplotypes, and 13 African-specific high-frequency haplotypes, and 31 haplotypes exclusively shared by Europeans and Africans. (**Table 3.5**).

We compared these identified high-frequency haplotypes with previously identified high-frequency haplotypes [94], and the presence of previously reported GWAS SNPs. Of the 38 non-African specific high-frequency Neanderthal haplotypes we identified, 19 were previously reported [94], including well-known substrates of adaptive introgression such as *WDR88*, *POU2F3*, and *TLR1/6/10* (**Figure 3.10A,B**) Intriguingly, we also identified 31 high-frequency haplotypes shared by Africans and Europeans, including *TRIM55* (**Figure 3.10C; Table 3.5**). These haplotypes would have been undetected in previous methods that relied on un-admixed reference panels. Furthermore, we were for the first time able to detect African-specific high-

frequency Neanderthal haplotypes (**Figure 3.10D; Table 3.5**). The 13 African-specific high-frequency Neanderthal haplotypes we identified showed enrichment for genes involved in immunological function (e.g. *IL22RA1* and *IFNLR1*), and ultraviolet-radiation sensitivity (e.g. *DDB1* and *IL22RA1*) [95,96]. These novel findings provide insight into the evolutionary history of these populations, the selective pressures they faced, and current variation in health and disease.

3.1.6 IBDmix Refines Locations of Loci Depleted of Neanderthal Ancestry

Previous analyses have identified large (>10 Mb) autosomal regions of the genome that are significantly depleted of Neanderthal ancestry in all non-African populations [30,31,33,34]. These "deserts" of archaic introgressed sequence appear at frequencies greater than expected under neutral models [30]. Because archaic deserts have been a consistent feature of recent catalogues of genome-wide Neanderthal ancestry, we believed it was necessary to asses if they would be replicated by IBDmix, or if deserts result from biases in previous methods for detecting Neanderthal ancestry. Following protocols previously described to identify archaic deserts [30], we analyzed our IBDmix callset including both African and non-African samples. It is noteworthy that including all African samples, a subset (YRI), or none, does not dramatically change the distribution of desert frequencies. This is consistent with the hypothesis that African Neanderthal sequence is predominantly a subset of non-African segments, introduced through historic back-migration, and not an independent admixture event or ILS. We replicated 4 of the 6 previously reported deserts of Neanderthal sequence, including the regions that contain *FOXP2* (chr7) and *ROBO1* and *ROBO2* (chr3) (**Figure 3.11; Table 3.6**). Furthermore, all 4 of the replicated deserts overlap regions previously defined as significantly depleted of Denisovan ancestry as well. Due to our own ultra-conservative approach to filtering Denisovan calls made by IBDmix, in order to maximize our signal for Neanderthal ancestry detection, we did not compare the overlap of these regions with our own Denisovan callset.

3.2 Discussion

We developed a novel approach to identify introgressed hominin sequence that persists in the genomes of modern humans, and show that it performs well compared to existing methods.

The main novelty of IBDmix is that it does not use a non-admixed reference panel. As such, we were able to make more unbiased inferences about levels of Neanderthal ancestry in African populations, and reveal how back-migrations to Africa confounded previous estimates of variation in Neanderthal ancestry among non-African populations. Our data also confirm genomic regions significantly depleted of Neanderthal ancestry as well as putative targets of adaptive introgression, including several loci that were previously not detectable when using an African reference population.

It is important to note that IBDmix has several limitations. In particular, IBDmix requires an archaic reference genome and therefore is not suitable for discovering introgressed sequence from unknown or unsequenced hominin lineages. IBDmix also requires a sufficiently large sample size (simulations support a minimum of 10 individuals; **Table 3.7**) to robustly estimate population allele frequencies. Consequently, it will be difficult to apply IBDmix to individual genomes or ancient human samples, where the sample size is limited and estimates of allele frequencies are imprecise. Recombination rate heterogeneity across the genome and between populations can also influence IBDMix segment size cutoffs. Population specific recombination rate maps would be ideal for calibrating the IBDmix size cutoff in particular genomic regions and populations. As such, IBDmix compliments existing approaches for identifying introgressed sequences in modern humans.

Applying IBDmix to geographically diverse populations revealed two unexpected observations. First, we discovered a substantial amount of Neanderthal sequence among African individuals. Specifically, among the 1000 Genomes African populations, we identified approximately 17Mb of Neanderthal sequence per individual (**Figure 3.5A; Table 3.2**), whereas previous inferences found considerably less than a megabase (ranging from 0.026 Mb in Esan to 0.5 Mb in Luhya) [30]. Accordingly, African individuals have approximately 33% as much Neanderthal sequence compared to non-African individuals. Note, this is likely a conservative estimate as IBDmix is sensitive to population-specific recombination rate heterogeneity, and changing the minimum segment size to 40kb increases the average amount of Neanderthal sequence in Africans to 20 Mb (**Table 3.3**). Regardless of the exact level of Neanderthal ancestry in Africans, our data clearly indicate it is substantially higher than previously anticipated. The larger amount of Neanderthal sequence in African individuals is not entirely unexpected, as recent studies have indicated assumptions about Neanderthal ancestry in Africa may have led to

underestimates [38,97]. Moreover, even early estimates of Neanderthal ancestry in non-Africans noted there was likely some amount of Neanderthal sequence in Africa [22,98,99] albeit not at the magnitude we find.

Our results strongly support the hypothesis that Neanderthal sequences in Africans are a consequence of ancient back-migration from admixed European ancestors in combination with pre-OOA human to Neanderthal gene-flow. Interestingly, the African populations we analyzed had similar levels of Neanderthal ancestry, perhaps indicating such sequences were inherited before subsequent population splitting. Alternatively, admixture among ancestral African populations may have led to a uniform distribution of this sequence across the continent. It is important to note that we have only analyzed a small sample of mostly northern African populations, and our results are only a preliminary map of the variation in Neanderthal ancestry among African populations. One may speculate that Neanderthal ancestry would be much lower, or absent, in populations with more genetic isolation such as the Khoe-San, whether that be limited historical contact with Eurasian individuals or with other African groups [46,100,101]. Regarding this, it is increasingly recognized that gene flow occurred among structured populations across the African continent [102–104], as Eurasian ancestry is found across Africa [105], and even early diverging groups such as the Khoe-San have up to 30% ancestry from recent admixture with East Africans and Eurasians [101]. Therefore, it will not be surprising if Neanderthal ancestry is present across the African continent. A fine-scale map of Neanderthal ancestry across Africa could provide a complimentary approach for elucidating patterns of dispersal and admixture among contemporary African populations.

The second major insight afforded by IBDmix is that levels of Neanderthal ancestry among non-African populations are more uniform than previous estimates. Specifically, as opposed to the 20% enrichment of Neanderthal sequence previously found in East Asians compared to Europeans [13,39,41,87], we only find an approximately 8% enrichment (**Figure 3.9A; Table 3.2**). We show that the reason for this discrepancy is that previous inferences using an African reference population underestimated the amount of Neanderthal sequence in Europeans. Due to historical back-migrations preferentially from ancestral European populations, Neanderthal sequence has been disproportionately under-called in present day Europeans compared to East Asians. We believe the modest 8% enrichment of Neanderthal sequence in East Asians and Europeans found by IBDmix is most parsimoniously explained by a

single wave of Neanderthal admixture occurring shortly after the Out-of-Africa dispersal. Variation in Neanderthal ancestry could be attributable to dilution [39]. In particular, present day European populations are thought to be a mixture of three ancestral groups, one of which had ancestry from a Basal Eurasian lineage that had little or no Neanderthal ancestry [106]. Previous studies found that dilution could not explain Neanderthal ancestry differences as large as 20% [41,87], but can readily account for the modest differences we find.

Note, our data do not preclude the possibility of additional, population specific admixture events with Neanderthals, and numerous instances of admixture events are known from ancient human samples, even though these individuals did not contribute genetically to contemporary human populations [107,108]. Nonetheless, the vast majority of Neanderthal ancestry can likely be explained by a single wave of admixture in the population ancestral to all non-Africans.

We further show that simple models of back-migration primarily from the European lineage, even with modest levels of migration (5×10^{-5}/generation), can replicate features of the empirical data (**Figure 3.6**). Although there is strong genetic evidence for back-to-Africa dispersals within the past several millennia [105,109,110], our data emphasize that migrations must have also occurred earlier, since the split of European and East Asian lineages is estimated at ~30kya (**Figure 3.8**). Indeed, evidence of more ancient dispersals and subsequent admixture of Eurasians into Africa has been described [109,111–113], and [109] estimate that 4 to 7% of most African genomes can be attributed to Eurasian ancestry. Therefore, both Out-of-Africa and In-to-Africa dispersals must be accounted for when interpreting global patterns of genomic variation, and accurately estimating proportions of archaic hominin ancestry in contemporary populations. Furthermore, our data show a strong concordance with models that include pre-Out-of-Africa human to Neanderthal gene flow. In fact, models that incorporate both elevated back-migration and pre-OOA human to Neanderthal gene flow match patterns in the empirical data better than either feature does alone (**Figure 3.6**). This highlights how complex the admixture history was between human and Neanderthal populations, and suggests barriers between these populations were much lower than historically thought.

In summary, our data extend and refine our understanding of Neanderthal ancestry in African and non-African populations, clarify models of archaic hominin admixture, and refine catalogues of genomic regions where Neanderthal sequence was deleterious and advantageous. It is notable that Neanderthal sequences have been identified in every contemporary modern human

genome analyzed to date. Thus, the legacy of gene flow with Neanderthals likely exists in all modern humans, highlighting our shared history.

3.3 Materials and Methods

3.3.1 IBDmix Algorithm Overview

As an input, IBDmix requires format-converted genotype data from whole genome sequencing for one archaic reference individual and a group of modern humans as the target genome. IBDmix is distinct from previous methods because it does not use a modern human unadmixed reference population to control for ILS between the archaic and modern human populations.

Proceeding site-by-site, IBDmix operates on one pair of archaic and modern human genomes at a time. At each position that passes variant filtering, IBDmix estimates the probability of IBD between the archaic and modern sample based on allele frequencies and summarizes this as a LOD score (**Table 3.8**). In order to identify putatively introgressed archaic segments in the modern genome, IBDmix applies a scanning algorithm based on dynamic programming to maximize the sum of LOD scores across a region above a pre-set threshold. Variants are added consecutively, expanding the interval, until the sum of the LOD scores cannot be further increased. The region with the maximized LOD score is called as a putative introgressed segment in the modern individual.

At completion, the output from IBDmix is a list of putatively introgressed segments and the probability of IBD between the archaic and modern human sample summarized as a maximized LOD score. Greater positive LOD scores reflect a higher probability of IBD across the specified region.

3.3.2 IBDmix Performance

We used *msprime* [114] to simulate sequence data and to call introgressed segments in simulated European, East Asian, and African modern individuals. Our simulations comprised 100 replicates of 15 Mb, sampling 100 diploid genomes each for African, European, and East Asian lineages, and 1 Neanderthal diploid genome. We used the coalescent trees from the simulations to identify the true introgressed haplotypes in the human populations. We simulated

a mutation rate of 1.25 x 10^{-8} per bp per generation. We used a recombination rate of 10^{-8} per bp per generation (1cM/Mb). The parameters for our demographic model were based on published estimates and assume a generation time of 25 years and a haploid ancestral effective population size of 7310. The split between the ancestors of Neanderthals and modern humans was set to 28,000 generations ago. The Out-of-Africa human migration occurred 3,920 generations ago. The rate of migration between the African and Out-of-Africa populations was 2 x 10^{-4} haploid individuals per generation, which corresponds to a cumulative Eurasian admixture into Africa over 2,400 generations of 2.4%. The rate of back-migration from the modern European to the African population was 1.7 x 10^{-5} haploid individuals per generation. We allowed for Neanderthal introgression to occur between 2,200 to 2,230 generations ago at a rate of 0.1% per generation, for an overall admixture proportion of 3%. We allowed for rapid growth of ~2% per generation in all human populations starting 200 generations ago, simulating the development of agriculture. We also used a model with a higher migration rate (5 x 10^{-4}) between African and Eurasian lineages to evaluate IBDmix and *S** performance under different demographic scenarios.

We randomly introduced sequence error to the genotype data created from msprime and therefore allowed sequence errors in both archaic and modern human genotypes in the simulation model. We tuned the parameters for IBDmix (LOD cutoff, archaic sequence error, maximum sequence error in modern human, sequence error as a function of MAF in modern human) using the simulated data. We evaluated the performance of IBDmix on simulated data, assessing metrics such as false positive rate, power, false discovery rate, precision and recall.

We simulated models with higher mutation rates, 2x, 5x, and 10x the default value (1.25x10^{-8} per bp per generation). We evaluated IBDmix performance under these models (**Figure 3.3**). To investigate the impact of recombination rate on IBDmix calling, we also simulated null models using the genome-wide average (10^{-8} per bp per generation) and 1/10^{th} that rate (10^{-9} per bp per generation). These models did not include Neanderthal introgression. We evaluated FPR of IBDmix under these null models.

We simulated models with two Neanderthal lineages representing an introgressing lineage and a sampled lineage. We tested several models varying the split time between these two lineages (70kya, 100kya, 145kya; **Figure 3.4**). We called introgressed sequence using

IBDmix with the sampled Neanderthal lineage as the reference genome, rather than the introgressing Neanderthal.

Because determining the precise endpoints of introgressed segments for any method remains difficult, when calling introgressed segments by IBDmix, we required that it overlap a call made using the coalescent trees by >1bp in order to be determined a true positive. Any introgressed segment called by IBDmix that does not overlap a call from the coalescent trees is considered a false positive. We calculated power as: (count of true positives) / (count of true segments from coalescent trees). We calculated FDR as: (count of false positives) / (count of false postives + count of true positives). We calculated FPR as: (total size of false positives) / (15 Mb – total size of true segments from coalescent trees).

3.3.2 Simulations of Demographic Models with Back-Migration and pre-Out of Africa Gene-Flow

To analyze the effects of back-migration and pre-OOA human to Neanderthal gene flow on the level of Neanderthal ancestry in Africans we compared empirical data from IBDmix calls made on 1000 Genomes samples in EUR (n=503), EAS (n=504), and YRI (n=108) populations to simulated data from *msprime* [114]. Our simulations consisted of 1000 replicates of 15MB chromosomes with diploid sample sizes matching those of the empirical data and including a sampled Neanderthal lineage (n=1). We used a basic demographic model, in which modern human and Neanderthal lineages separated 28,000 generation ago, African and non-African lineages separated 3,000 generations ago, and European and East Asian lineages separated 920 generations ago. Effective populations sizes are allowed to vary, and at 200 generations ago there is rapid population growth in all populations, mirroring the development of agriculture. We specify a recombination rate of $1x10^{-8}$ per bp per generation, a mutation rate of $1.2x10^{-8}$ per bp per generation, and a generation time of 25yrs per generation. We included a single pulse of admixture from the Neanderthal into the non-African lineage 2,000 generations ago, at a level of 5% per generation for a single generation. As well, we included a single migration parameter from either the ancestral Eurasian population into Africans, which stopped after the split of Europeans and East Asians, or from Europeans into Africans after the split with East Asians until the present (**Figure 3.8**). We specified the migration to occur only in one direction (from non-Africans into Africans) and tested a range of migration rates that included levels established in

previous demographic models [78]. In models including pre-OOA gene flow, we specified an admixture level of 10% per generation for a single generation, and specified the admixture event at $4x10^3$, $6x10^3$, or $10x10^3$ generations ago.

Sequence data from the simulations were collected in vcf format and analyzed separately using IBDmix and the S* pipeline [30] in order to identify Neanderthal introgressed segments in simulated human individuals. As well, we collected the true introgressed segments from the simulated data using the coalescent trees. For IBDmix, we used a threshold of LOD > 4 and removed segments <50kb in order to create a final call set of introgressed segments. In order to identify introgressed segments using S*, we calculated S*-scores and Neanderthal match-percent in 50kb windows at 10kb overlapping steps. We determined statistically significant S*-scores and match-percent levels using 10,000 replicates of a null simulation. We required that windows have S* p-value < 0.01 and Neanderthal-match-percent p-value < 0.05 to be considered Neanderthal-introgressed. Overlapping statistically significant introgressed windows were merged to produce full Neanderthal introgressed segments.

3.3.3 Refining Neanderthal Callset by Using Denisovan Sequences as a Negative Control

We adopted an ultra-conservative approach to filtering our callset in order to maximize our signal of detected Neanderthal ancestry. After initially calling Neanderthal and Denisovan sequences using IBDmix, we refined the Neandertal callset by masking any regions that were called as Denisovan sequence in Africans and also present as Neanderthal sequence in any population. Such regions represent either ILS shared in all hominins from a deep coalescent event, or true Neanderthal sequence mis-assigned as Denisovan sequence. After filtering, the average amount of Neanderthal ancestry in each population decreased by several Mb, but maintained the same patterns and relative proportions as discussed in the paper (**Table 3.2**). Furthermore, we observed some regions with a high proportion of derived alleles in the Neanderthal genome that also shared an unusually high proportion of derived alleles in some or all modern human populations. These regions may contain exceptional local genetic features, and may exhibit more complex evolutionary and recombinant histories than other genomic regions. To be conservative, we also provide a callset removing regions where the proportion of derived alleles in the Neanderthal genome for a given window fell in the upper 99.9th-percentile. This

further reduced the amount of detected Neanderthal ancestry in all populations, however relative levels of Neanderthal ancestry for different populations were still robust (**Table 3.2**).

For our identification of Denisovan introgressed segments, we introduced additional filters to refine the initial callset. We masked any regions that were both detected as Neanderthal and Denisovan sequence for all populations, removing mis-assigned sequence and ILS. We further controlled for ILS by removing from all populations segments that were called as Denisovan in Africans at a frequency ≥ 30%, accounting for 10% of detected Denisovan segments in Africans. The average amounts of detected Denisovan sequence in all populations are reported in **Table 3.4**.

3.3.4 Replicating Regions Significantly Depleted of Neanderthal Introgressed Sequence

We have previously described a method for identifying regions significantly depleted of Neanderthal sequence identified by S* in non-African populations [30]. In summary, we break the genome into windows of varying size (5-15Mb) at 100kb overlapping steps, requiring that a window be composed of >70% unfiltered bases. We then determine, for a given window, the average number of Neanderthal introgressed bases across all individuals. We perform this measure for all windows that meet the filtering requirements in order to generate a distribution for the average level of Neanderthal ancestry across the genome. Windows that are in the lower 99th-percentile for average amount of introgressed sequence are considered significantly depleted and are merged with overlapping windows to define depleted regions. The final list of depleted regions is determined by merging the significant regions of all window sizes. We applied the same analysis to Neanderthal introgressed calls made with IBDmix and compared these sets of depletions to those identified using the S*-callset (**Figure 3.11**) [30].

3.3.5 Comparing Simulated Data to Empirical Data

In cases where we compared simulated data to empirical data (**Figure 3.6**) we filtered the simulated IBDmix calls to replicate filtering for empirical data, removing segments <50kb. To analyze the distribution of segment lengths for calls made in African and non-African populations, we used unmerged calls from all African individuals (LWK, GWD, MSL, YRI, ESN), and all non-African individuals, except for ASW and ACB. Calls made by IBDmix in African samples that overlapped any non-African call by 1bp were categorized as “African

shared calls" (n=95032), and those that did not overlap any non-African calls were categorized as "African unique calls" (n=900).

To analyze the frequency within the African population of segments identified as African and shared with non-Africans, we limited our analysis to calls made in YRI that overlapped by 1bp with calls made in Europeans or East Asians (n=19333). We then counted for each call the number of other African individuals who carried an overlapping call, and assigned each call as either "Below 10%", where < 11 YRI individuals carried an overlapping segment (n=2586), or "Above 10%", where ≥ 11 other YRI individuals carried an overlapping segment (n=16747). We measured the number of calls in each category as a proportion of the total number of calls in YRI that intersected calls made in Europeans or East Asians.

We measured the ratio of Neanderthal sequence in East Asians compared to Europeans with and without masking overlapping YRI calls. Eurasian calls were masked if they overlapped a YRI call by 1bp. We summed together the lengths of the calls in each data set to arrive at a total amount of sequence in each population.

3.3.6 Reference Panel Size Effect on S Admixture Estimates*

We examined how reference panel size for S* affects Neanderthal ancestry estimates by bootstrap resampling the Yoruba 1000 Genomes Project samples and reanalyzing chromosome 1 for Europeans and East Asians. We bootstrap sampled Yoruba (YRI, n=108) individuals from the 1000 Genomes Project to generate multiple reference panels of sizes n=[1, 2, 5, 10, 25, 50, 75, 108]. We then re-called Neanderthal introgressed sequence on chromosome 1 for European (n=503) and East Asian (n=504) individuals using the S*-pipeline [30] and the new reference panel, requiring S* p-value < 0.01 and Neanderthal match-percent p-value < 0.05. We performed 10 replicates of this analysis resampling the YRI reference panel for each replicate and calculated the mean level of S*-sequence identified per sample.

The mean S*-sequence called for each sample across the 10 replicates was compared to the average amount of S*-sequence called for samples using a reference panel of YRI=1. We used this normalized mean to test for significant difference (t-test) between the amount of S*-sequence called in EUR and EAS for different reference panel sizes. In addition, for each reference panel size, an average admixture proportion was calculated for each population across

replicates by dividing the mean S*-sequence for all 10 replicates by the total amount S*-queryable sequence [30].

3.3.7 Identifying High-Frequency Introgressed Haplotypes From IBDmix Data

We used derived allele frequencies calculated from 1000 Genomes Project to identify population specific high-frequency introgressed haplotypes. To do this, we identified sites that had extreme differences in derived allele frequency between populations, intersected Neanderthal segments identified by IBDmix, and matched the Altai Neanderthal reference alleles.

We began by removing 1000 Genomes Project variants that we masked during the IBDmix analysis. We then intersected the remaining variants with Neanderthal calls made by IBDmix in EUR, EAS, and AFR populations. For variants that intersected identified Neanderthal segments, we calculated the differences in the derived allele frequencies between EUR and EAS, and AFR and EUR and EAS. We identified the lower and upper 1% values for the differences in derived allele frequencies as part of an outlier approach. For example, in the comparison of EUR and EAS sites, we retained sites where the absolute difference in the derived allele frequency between EUR and EAS was >40%. We further filtered on the derived allele matching the Neanderthal allele, and in the case of EUR and EAS calls, that the AFR derived allele frequency was <1%. To maximize our ability to identify population-specific high-frequency haplotypes, we required that, for EUR-specific calls, the EUR derived allele frequency be >40% and the EAS derived allele frequency be <10%; for EAS-specific calls, the EUR derived allele frequency be <10% and the EAS derived allele frequency be >40%; for AFR-specific calls, the EUR and EAS derived allele frequencies both be <5%. We also required that for a given allele, the number of individuals in a population who carry the Neanderthal sequence at that locus be greater than 5. By intersecting the alleles that met these filtering criteria with the merged Neanderthal callsets for EUR and EAS combined and AFR, we identified a final set of distinct high-frequency introgressed haplotypes (**Table 3.5**). We compared these identified high-frequency haplotypes with previously identified high-frequency haplotypes [94], and the presence of previously reported GWAS SNPs pulled from UCSC Genome Browser with reported $p \leq 1x10\text{-}5$.

3.3.8 Rate of Overlap Between Neanderthal Segments and European Ancestry Tracks in African Samples

Under the model that back-migration from Europeans to Africans accounts for a substantial amount of Neanderthal ancestry in Africans, we hypothesized that we should find enrichment for Neanderthal ancestry in Africans at loci that also show evidence of European ancestry. To test this hypothesis, we compared for chromosome 1 the rate of overlap of Neanderthal segments identified by IBDmix with tracks of European and East Asian ancestry identified by RFMix [93] on a per-individual basis for all 504 African individuals analyzed in our study.

We began by taking the phased genotype data for chromosome 1 and processed this with vcftools [115] and custom scripts to retain only bi-allelic, completely phased sites that could be mapped to genomic coordinates. After processing, we retained 245,126 sites for analysis with RFMix.

We used RFMix to analyze the ancestry of each African individual separately. Specifically, we adopted a leave-one-out approach, in which each African individual was analyzed against a reference panel composed of the remaining 503 African samples, 503 European samples, and 504 East Asian samples. We recoded the ancestry tracks determined by RFMix from genomic positions into base-pair coordinates, and merged tracks of European or East Asian ancestry that were within 10kb of similar ancestry tracks. The median track length for European ancestry is 142kb, and for East Asian ancestry is 132kb. The average level of European and East Asian ancestry per individual is 2.2% and 0.45%, respectively.

Next, we compared the rate of overlap of Neanderthal calls with European or East Asian ancestry tracks on a per-individual basis,

$$r_{emp} = \frac{\text{\# of Neand segments overlapping EUR or EAS ancestry}}{\text{Total \# of Neand segments}}$$

and took the average across all 504 African individuals to calculate an empirical value for the average rate of overlap of Neanderthal sequence and European or East Asian ancestry. To test the significance of these empirical values, we performed permutation tests, analyzing an individuals Neanderthal calls against a random individuals European and East Asian ancestry tracks. We performed 10,000 replicates of this analysis, averaging the rate of overlap for all 504 Africans in each replicate. When we compared the empirical average rate of overlap for East Asian ancestry to the null distribution, we found 4495/10000 replicates equaled or exceeded the

empirical value. When we repeated this with the European ancestry data, we found 0/10000 replicates equaled or exceeded to empirical value. These results indicate an average rate of overlap for Neanderthal segments and tracks of European ancestry in African individuals greater than expected by chance. This is not the case for Neanderthal segments and tracks of East Asian ancestry.

3.3.9 Data and Software Availability

IBDmix software:
https://github.com/PrincetonUniversity/IBDmix

Segments of introgression detected in 1000 Genomes data using IBDmix:
https://drive.google.com/drive/folders/1mDQaDFS-j22Eim5_y7LAsTTNt5GWsoow?usp=sharing

3.4 Figures and Tables

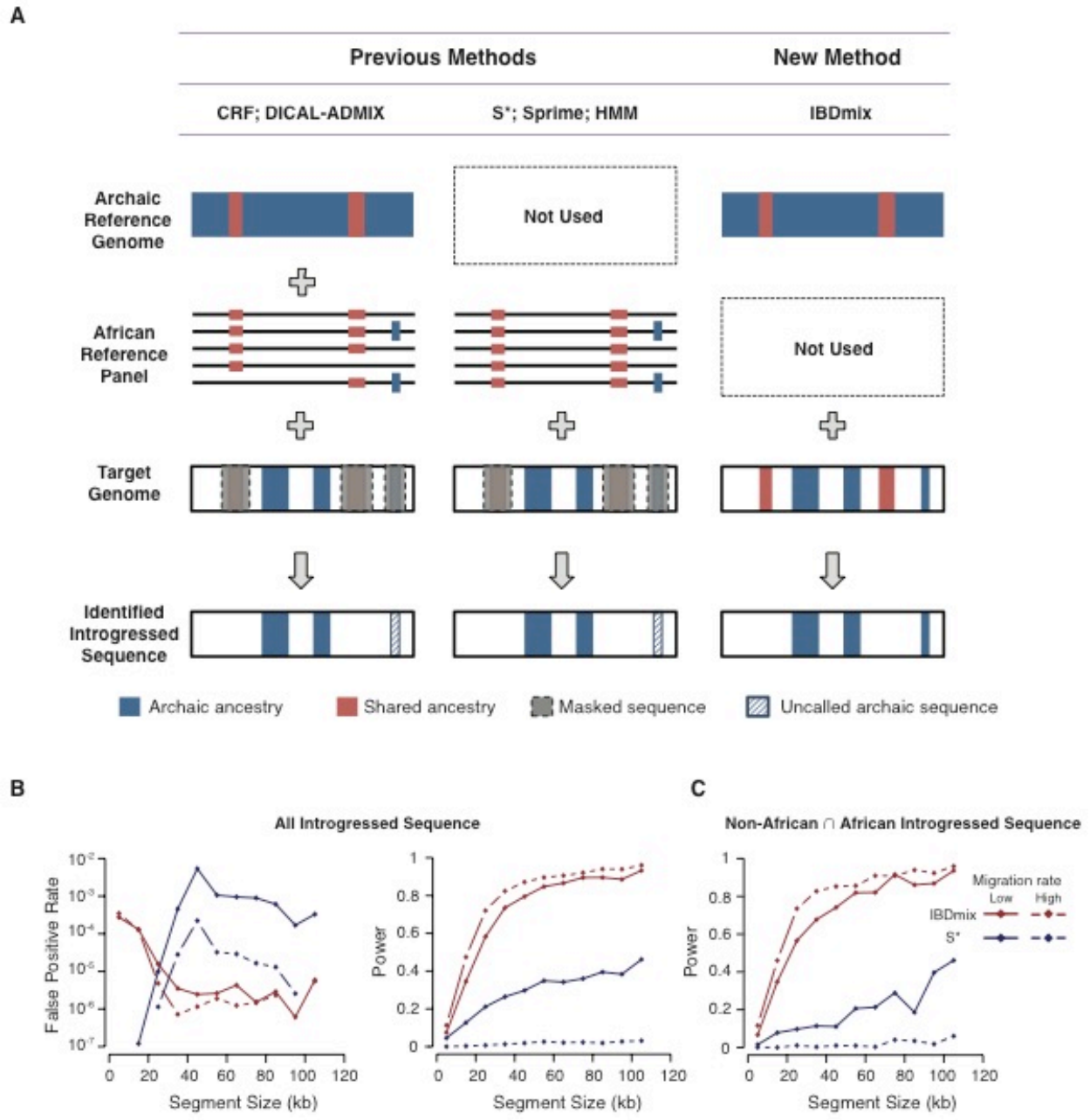


Figure 3.1. Evaluation of IBDmix performance and comparison to previous methods.
(A) Summary of IBDmix workflow compared to previous methods for identifying introgressed archaic sequences in modern human genomes.
(B and C) Comparison of IBDmix performance to *S** on simulated data generated from models with low back-migration rate (1.7×10^{-5}/generation) and high back-migration rate (5×10^{-4}/generation). In (B), power and false positive rate is calculated for all simulated Neanderthal segments in non-Africans whereas in (C) the power to detect a Neanderthal segment in non-Africans conditional on it also being present in Africans is shown.

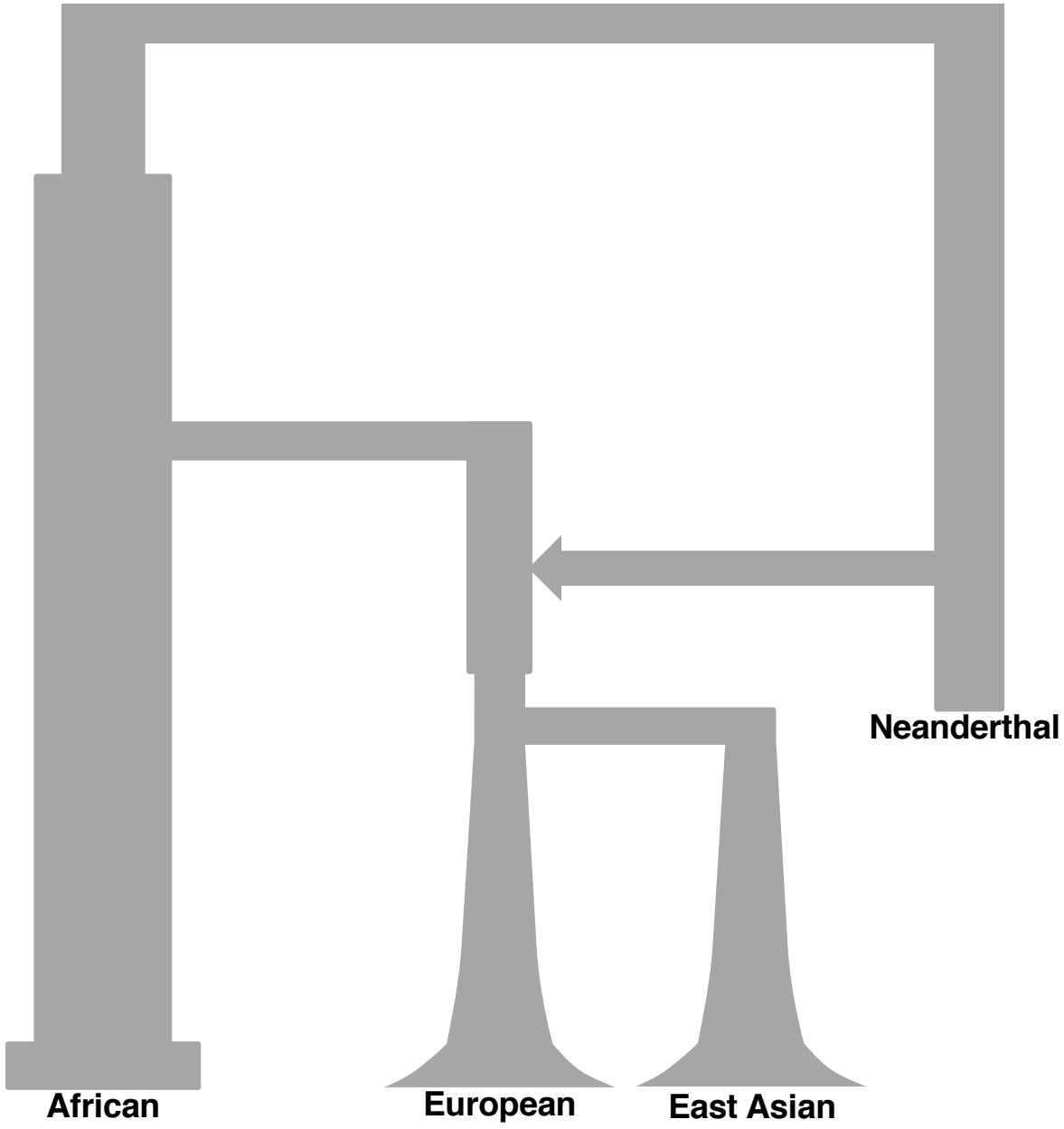


Figure 3.2. Simplified schematic of the demographic model used for simulations evaluating the performance of IBDmix.

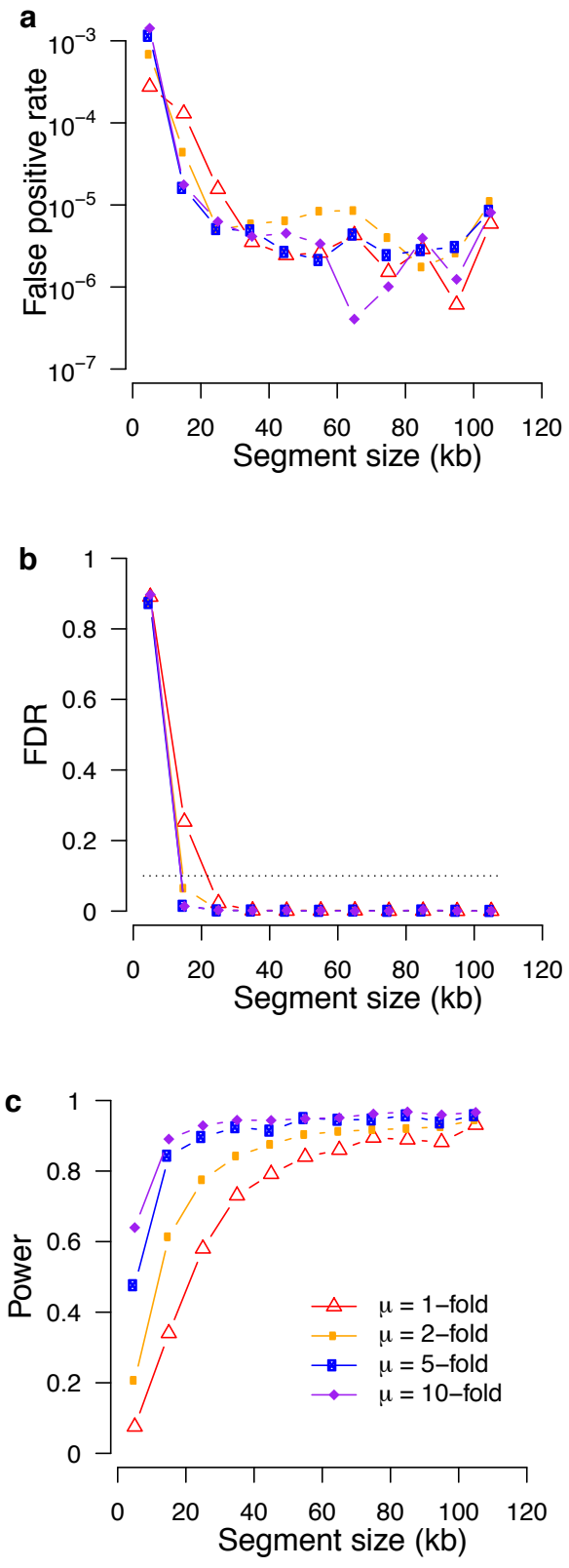


Figure 3.3. Effect of Mutation Rate on IBDmix Performance.

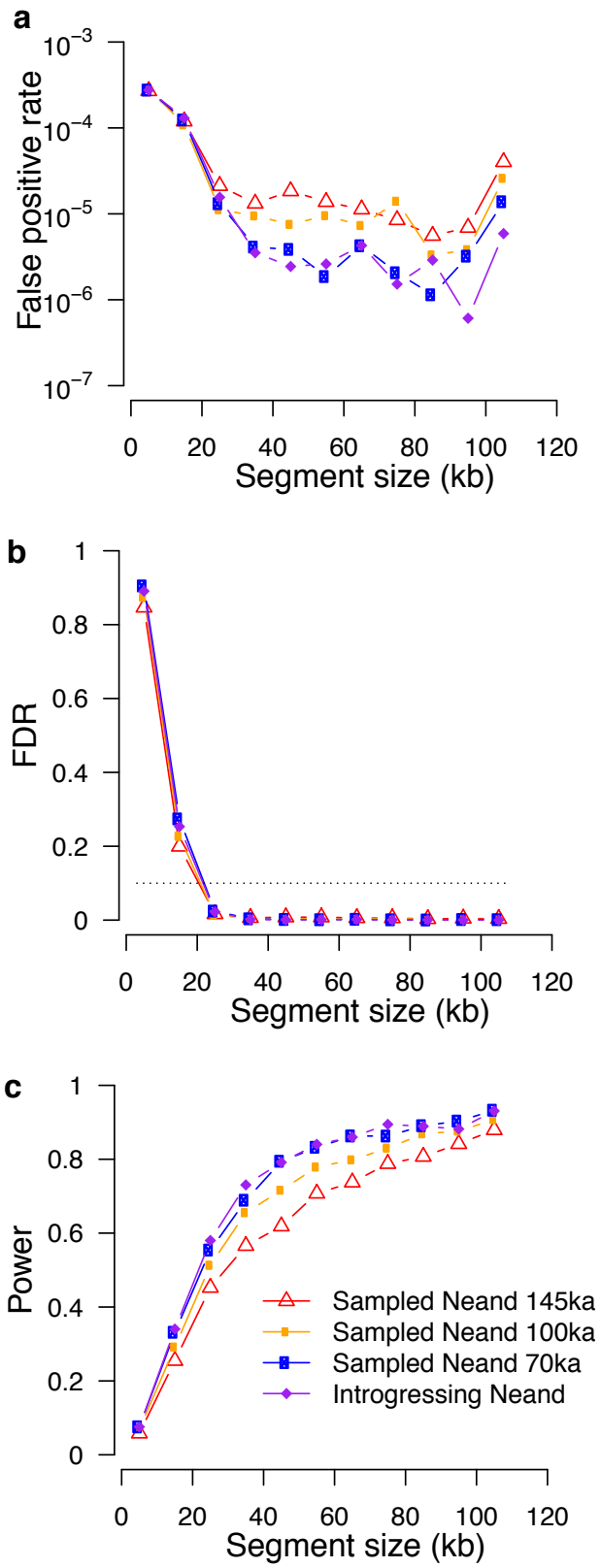


Figure 3.4. Effect of Reference Neanderthal Genome Split Time on IBDmix Performance.

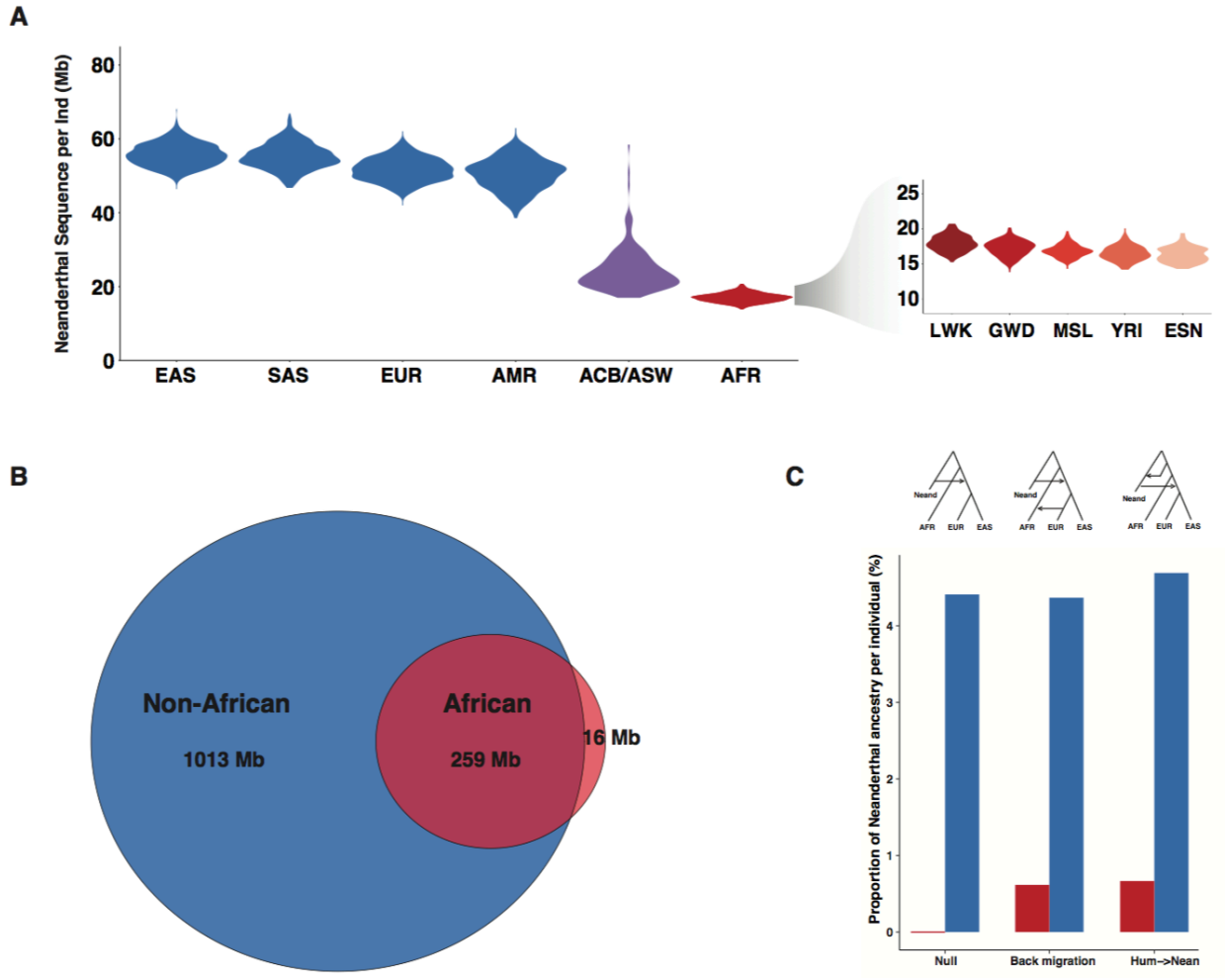


Figure 3.5. Neanderthal introgressed sequence detected in 1000 Genomes Project populations.
(A) Violin plots showing the amount of Neanderthal sequence called per individual across geographically diverse populations from the 1000 Genomes Project. Non-African, African admixed, and African populations are shown in blue, purple, and red, respectively. The inset figure shows the amount of Neanderthal sequence per individual for five African subpopulations.
(B) Venn diagram showing the amount overlap in identified Neanderthal sequence in non-African and African populations.
(C) Bar plots showing the proportion of Neanderthal ancestry detected in African (red) and non-African (blue) populations in simulations with 5% Neanderthal admixture and no gene-flow between modern populations (Null), gene-flow from the Europeans into Africans (Back migration), or gene-flow from a pre-OOA humans into Neanderthals (Hum→Nean).

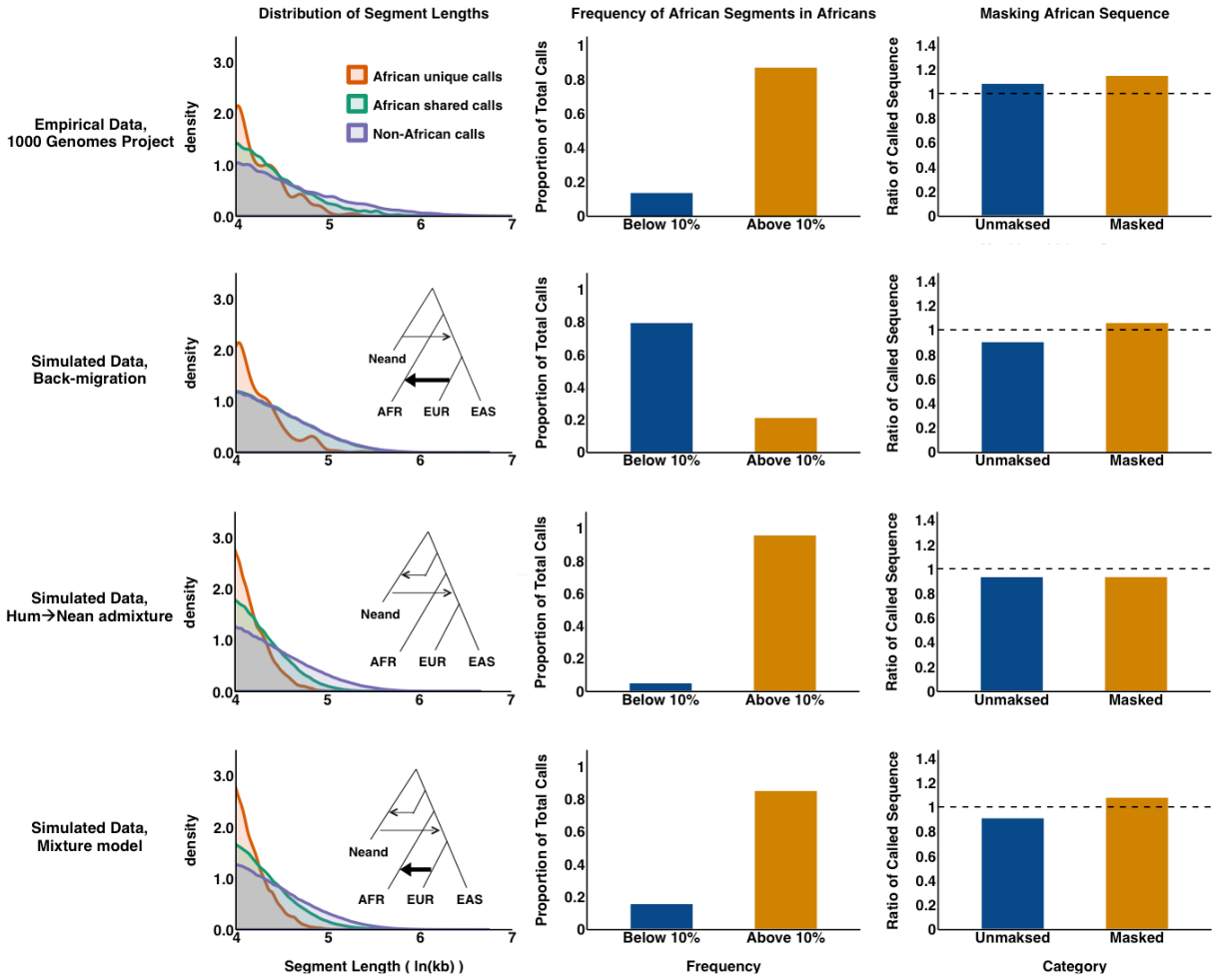


Figure 3.6. Neanderthal segments identified in Africans are a consequence of back-migration and pre-Out of Africa gene-flow.
Features of the empirical data were compared to simulated data with increased levels of back-migration (m=5x10^{-5}/gen), and pre-Out of Africa gene-flow (t=100kya, α=10%). From left to right, the distribution of Neanderthal segment lengths, frequency of segments in Africans that segregate in Africans and non-Africans, and the ratio of Neanderthal ancestry in East Asians compared to Europeans with and without masking of sequence shared with Africans.

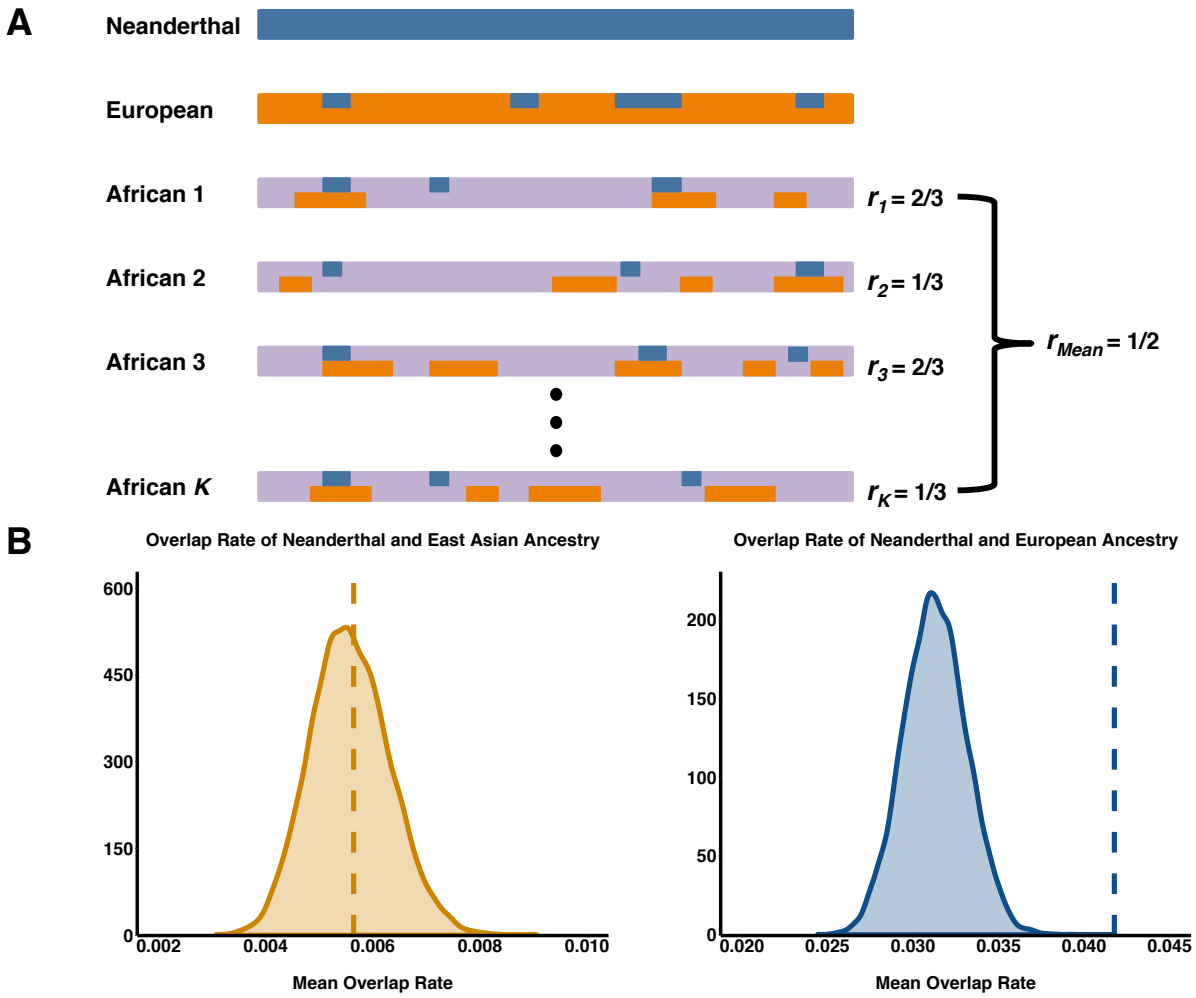


Figure 3.7. Enrichment in overlap of Neanderthal segments and European ancestry segments in African individuals.
(A) Schematic of how an enrichment for European ancestry overlap was assessed. For each African individual, sequence data from chromosome 1 were analyzed for tracks of Neanderthal, European, and East Asian ancestry. For each individual, the rate of overlap between Neanderthal segments and non-African segments was calculated, and the mean across all African individuals was taken as the empirical value.
(B) Distributions of the mean rate of overlap from permuted data, with the empirical value for Europeans and East Asians demarcated as dashed lines. The rate of overlap for European ancestry is highly significant ($p < 0.0001$), while the rate of overlap for East Asian ancestry is not ($p > 0.05$).

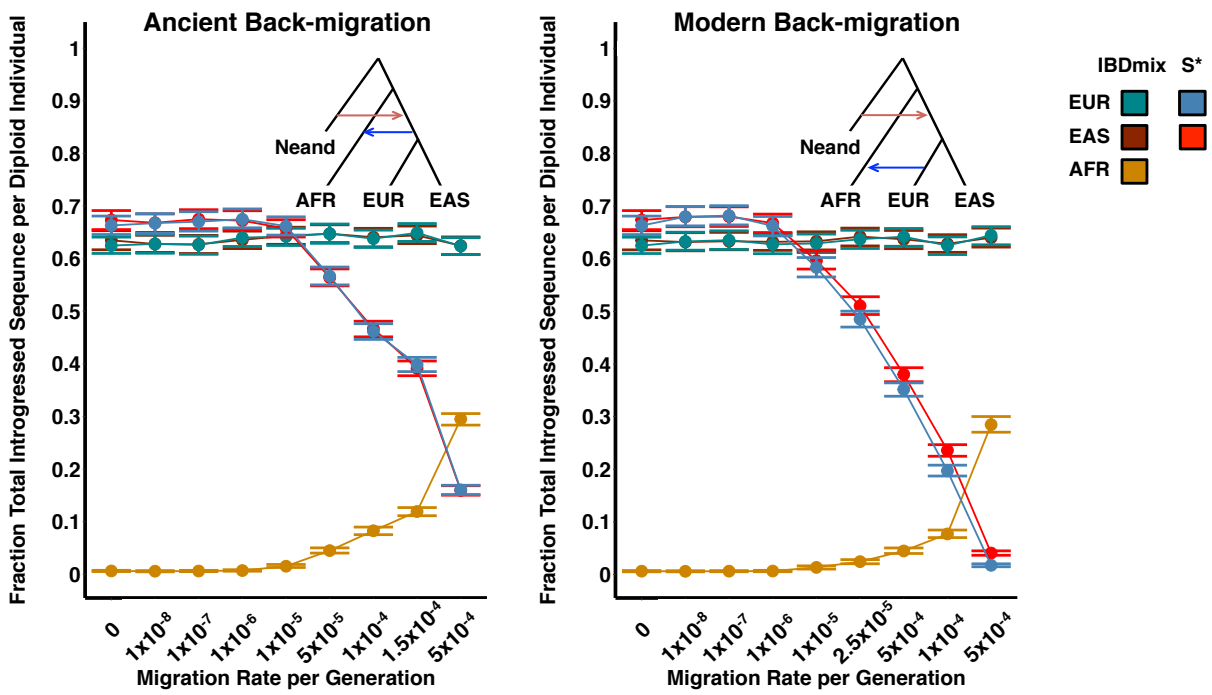


Figure 3.8. Migration back-to-Africa from ancestral Europeans affects estimates of Neanderthal ancestry.
The fraction of total sequence detected by *S** and IBDmix are shown from coalescent simulations for the models with increasing gene-flow into an African lineage from a non-African one. Neanderthal sequence was called in European (EUR), East Asian (EAS), and African (AFR) populations. Note, back-migration from European ancestors can produce a 20% enrichment in the amount Neanderthal sequence detected by *S** in East Asians compared to Europeans (right), while the amount of sequence detected by IBDmix is unaffected. Back-migration from ancestral Eurasians (left) reduces the amount of Neanderthal sequence recovered by *S**, but does not produce the apparent enrichment in East Asians when compared to Europeans.

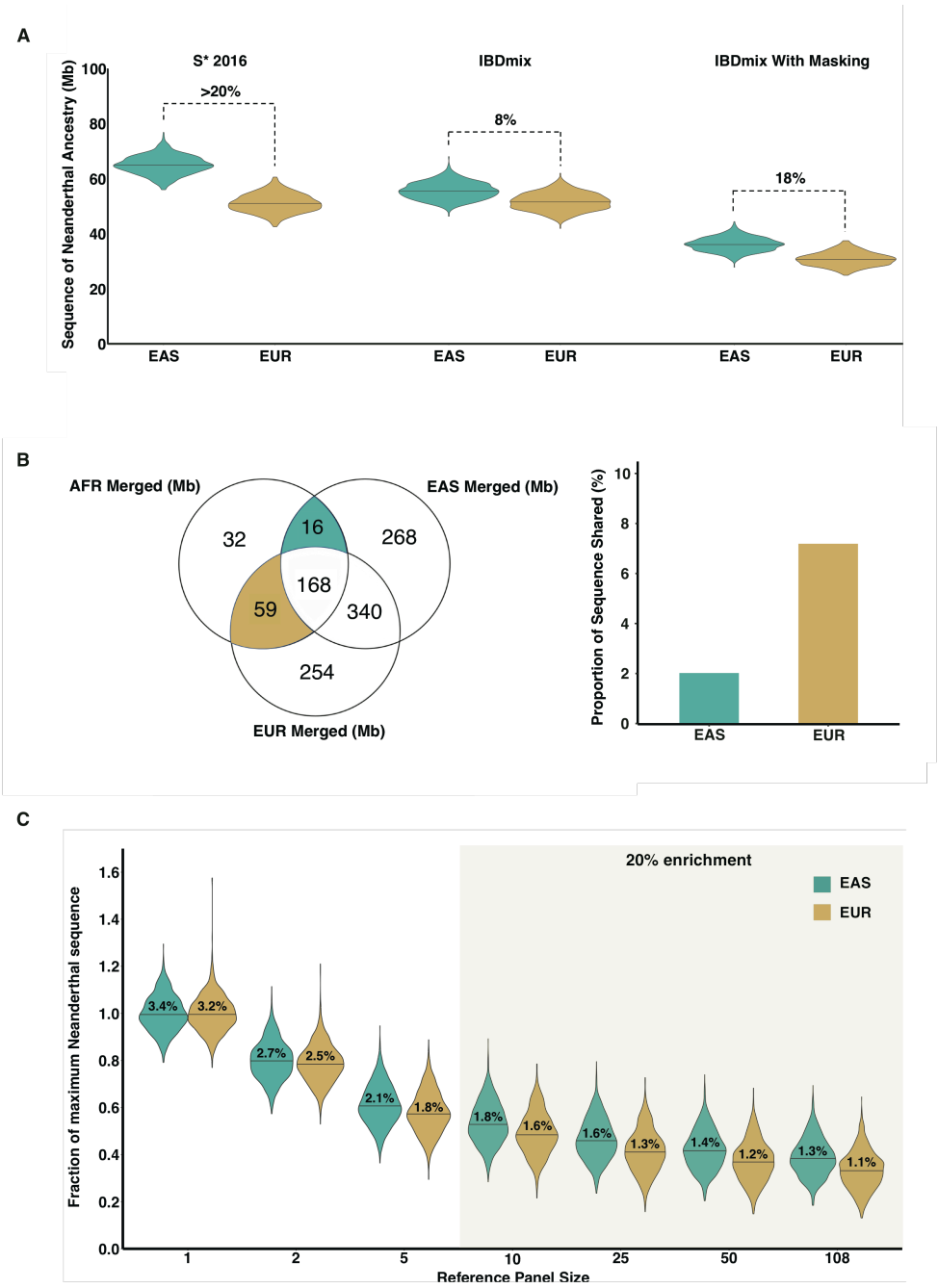


Figure 3.9. Disproportionate sharing of Neanderthal sequence differentially biases estimates of Neanderthal ancestry.
(A) Violin plots showing enrichment of Neanderthal ancestry in East Asians compared to Europeans for *S** and for IBDmix with and without masking Neanderthal sequence shared with Yoruba.
(B) Venn diagram illustrating the amount of sequence shared among Africans and non-Africans. The bar plot shows the amount of exclusively shared sequence between Africans and non-Africans as a proportion of the total amount of sequence for each population.
(C) Violin plot showing the decreasing amount of Neanderthal sequence identified in East Asian and European individuals by *S** with increasing African reference-panel size.

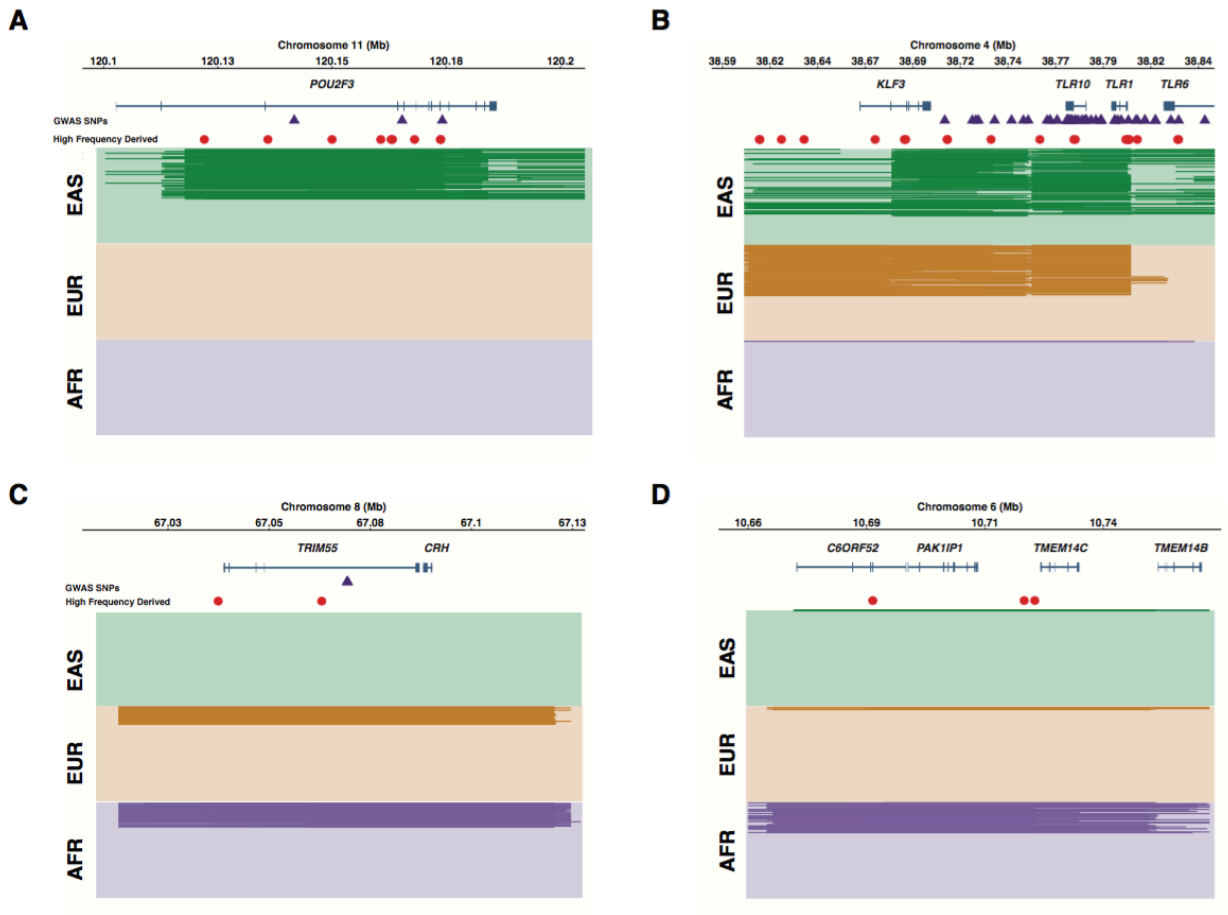


Figure 3.10. Population-specific high-frequency introgressed segments.
In all plots, each row is an individual and organized by population (EAS, EUR, and AFR for East Asians, Europeans, and Africans, respectively). Neanderthal segments called by IBDmix are plotted in dark green (EAS), orange (EUR), or purple (AFR). GWAS SNPs are shown as purple triangles and populations-specific high-frequency derived alleles (DAF > 40%) that match the Altai reference genome are shown as red circles.
In (A) and (B), examples of high-frequency introgressed segments detected in East Asian and European populations are shown for the *POU2F3* and the *TLR1|6|10* cluster.
(C) An example of a high-frequency Neanderthal segment shared between Europeans and Africans at *TRIM55*. This haplotype, identified by IBDmix, is missed by methods that mask sequence shared by African and non-African populations.
(D) Example of an African-specific high-frequency haplotype that spans multiple genes.

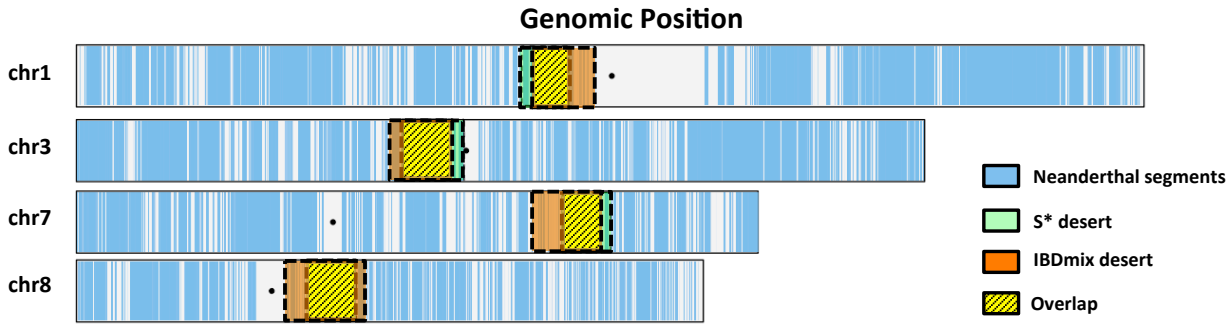


Figure 3.11. Visualization of *S and IBDmix identified desert regions and their overlap.**

Table 3.1. False Positive Rate of IBDmix under Models with Different Recombination Rate.

Segment Size (kb)	FPR (under model with 1cM/Mb)	FPR (under model with 0.1cM/Mb)
0-10	2.95×10^{-4}	3.96×10^{-4}
10-20	9.72×10^{-5}	6.00×10^{-4}
20-30	2.50×10^{-6}	2.38×10^{-4}
30-40	6.78×10^{-7}	1.10×10^{-4}
40-50	0.00	4.98×10^{-5}
50-60	0.00	3.12×10^{-5}
60-70	0.00	1.25×10^{-5}
70-80	0.00	3.97×10^{-6}
80-90	0.00	2.22×10^{-6}
90-100	0.00	1.03×10^{-5}
>=100	0.00	5.00×10^{-6}

Table 3.2. Average (and Standard Deviation) Amount of Neanderthal Sequence (Mb) for 1000 Genomes Project Populations.

Population code	Super population code	IBDmix calls	Filter1*	Filter2*
ACB	AFR	26.9 (4.7)	22.7 (4.1)	21.7 (4.2)
ASW	AFR	32.2 (7.9)	27.5 (7.3)	26.5 (7.3)
ESN	AFR	19.9 (1.2)	16.4 (1.1)	15.4 (1.1)
GWD	AFR	20.9 (1.3)	17.2 (1.2)	16.3 (1.1)
LWK	AFR	21.6 (1.3)	18.0 (1.2)	17.0 (1.2)
MSL	AFR	20.4 (1.1)	16.9 (1.0)	15.9 (0.9)
YRI	AFR	20.0 (1.2)	16.5 (1.1)	15.5 (1.1)
CLM	AMR	60.0 (4.0)	52.4 (3.7)	51.6 (3.6)
MXL	AMR	59.6 (4.0)	52.4 (3.3)	51.6 (3.3)
PEL	AMR	52.7 (4.4)	46.7 (3.8)	46.0 (3.6)
PUR	AMR	57.6 (4.0)	50.3 (3.6)	49.4 (3.7)
CDX	EAS	60.7 (3.1)	55.0 (2.8)	54.2 (2.9)
CHB	EAS	62.0 (3.1)	56.2 (3.0)	55.4 (3.0)
CHS	EAS	61.7 (3.2)	56.0 (3.0)	55.2 (3.0)
JPT	EAS	61.5 (3.3)	55.6 (3.1)	54.9 (3.1)
KHV	EAS	61.5 (3.7)	55.7 (3.4)	54.9 (3.4)
CEU	EUR	60.1 (3.2)	52.1 (2.9)	51.4 (2.9)
FIN	EUR	60.8 (3.1)	53.2 (2.7)	52.4 (2.7)
GBR	EUR	59.8 (3.7)	51.8 (3.1)	51.0 (3.1)
IBS	EUR	58.7 (3.6)	50.9 (3.3)	50.2 (3.3)
TSI	EUR	58.3 (3.2)	50.6 (3.0)	49.9 (3.0)
BEB	SAS	64.2 (3.7)	57.1 (3.6)	56.2 (3.6)
GIH	SAS	61.8 (3.3)	54.7 (2.9)	53.9 (2.9)
ITU	SAS	61.9 (4.1)	55.0 (3.8)	54.3 (3.7)
PJL	SAS	61.9 (4.1)	54.7 (3.7)	53.9 (3.6)
STU	SAS	61.8 (3.3)	54.9 (3.1)	54.2 (3.0)

*Filter1 is the average amount after filtering Denisovan sequence in Africans from the original callset of IBDmix using a 50kb threshold. Filter2 is the average amount after further filtering regions with high proportion of derived alleles in Neanderthal genome from callset of Filter1. All the sequence numbers are Mb, and shown as mean (sd).

Table 3.3. Average (and Standard Deviation) Amount of Neanderthal Sequence (Mb) for 1000 Genomes Project Populations Based on Different Segment Size Thresholds Using IBDmix.

Population code	Super population code	30kb Threshold		40kb Threshold		50kb Threshold	
		IBDmix calls	Filter1	IBDmix calls	Filter1	IBDmix calls	Filter1
ACB	AFR	37.5 (5.1)	29.8 (4.2)	31.4 (4.9)	25.9 (4.2)	26.9 (4.7)	22.7 (4.1)
ASW	AFR	43.2 (8.4)	34.8 (7.4)	36.8 (8.2)	30.7 (7.4)	32.2 (7.9)	27.5 (7.3)
ESN	AFR	29.4 (1.3)	22.9 (1.2)	23.9 (1.3)	19.2 (1.1)	19.9 (1.2)	16.4 (1.1)
GWD	AFR	30.5 (1.4)	23.7 (1.2)	25.0 (1.4)	20.1 (1.2)	20.9 (1.3)	17.2 (1.2)
LWK	AFR	31.7 (1.4)	24.7 (1.1)	25.9 (1.3)	21.0 (1.1)	21.6 (1.3)	18.0 (1.2)
MSL	AFR	30.2 (1.1)	23.5 (1.0)	24.5 (1.1)	19.8 (1.0)	20.4 (1.1)	16.9 (1.0)
YRI	AFR	29.8 (1.5)	23.2 (1.2)	24.1 (1.3)	19.4 (1.1)	20.0 (1.2)	16.5 (1.1)
CLM	AMR	72.6 (3.8)	60.4 (3.4)	65.8 (3.9)	56.3 (3.5)	60.0 (4.0)	52.4 (3.7)
MXL	AMR	71.6 (4.4)	60.0 (3.5)	65.0 (4.1)	56.0 (3.4)	59.6 (4.0)	52.4 (3.3)
PEL	AMR	64.5 (5.2)	54.4 (4.3)	58.2 (4.7)	50.5 (4.0)	52.7 (4.4)	46.7 (3.8)
PUR	AMR	70.4 (3.9)	58.4 (3.3)	63.4 (3.9)	54.1 (3.5)	57.6 (4.0)	50.3 (3.6)
CDX	EAS	72.9 (3.2)	62.9 (2.9)	66.3 (3.1)	58.8 (2.9)	60.7 (3.1)	55.0 (2.8)
CHB	EAS	74.4 (3.2)	64.1 (3.0)	67.7 (3.1)	60.0 (3.0)	62.0 (3.1)	56.2 (3.0)
CHS	EAS	74.0 (3.4)	63.8 (3.1)	67.3 (3.3)	59.8 (3.0)	61.7 (3.2)	56.0 (3.0)
JPT	EAS	73.8 (3.4)	63.4 (3.1)	67.2 (3.4)	59.5 (3.1)	61.5 (3.3)	55.6 (3.1)
KHV	EAS	73.9 (3.6)	63.7 (3.3)	67.1 (3.6)	59.5 (3.4)	61.5 (3.7)	55.7 (3.4)
CEU	EUR	72.1 (3.2)	59.5 (2.8)	65.5 (3.2)	55.6 (2.9)	60.1 (3.2)	52.1 (2.9)
FIN	EUR	72.8 (3.3)	60.6 (2.7)	66.3 (3.2)	56.8 (2.7)	60.8 (3.1)	53.2 (2.7)
GBR	EUR	71.6 (3.7)	59.0 (3.0)	65.1 (3.7)	55.2 (3.1)	59.8 (3.7)	51.8 (3.1)

IBS	EUR	70.7 (3.7)	58.2 (3.3)	64.2 (3.7)	54.4 (3.3)	58.7 (3.6)	50.9 (3.3)
TSI	EUR	70.4 (3.3)	58.0 (3.0)	63.8 (3.3)	54.1 (3.0)	58.3 (3.2)	50.6 (3.0)
BEB	SAS	77.3 (3.8)	65.4 (3.6)	70.1 (3.8)	61.0 (3.6)	64.2 (3.7)	57.1 (3.6)
GIH	SAS	74.5 (3.3)	62.8 (2.9)	67.5 (3.2)	58.4 (2.9)	61.8 (3.3)	54.7 (2.9)
ITU	SAS	74.8 (4.4)	63.2 (3.9)	67.7 (4.2)	58.9 (3.9)	61.9 (4.1)	55.0 (3.8)
PJL	SAS	74.4 (4.3)	62.6 (3.7)	67.5 (4.2)	58.4 (3.6)	61.9 (4.1)	54.7 (3.7)
STU	SAS	74.5 (3.5)	63.0 (3.1)	67.5 (3.4)	58.8 (3.0)	61.8 (3.3)	54.9 (3.1)
ACB	AFR	37.5 (5.1)	29.8 (4.2)	31.4 (4.9)	25.9 (4.2)	26.9 (4.7)	22.7 (4.1)

*Filter1 is the average amount after filtering Denisovan sequence in Africans from the original callset of IBDmix. All the sequence numbers are Mb, and shown as mean (sd).

Table 3.4. Average Amount of Denisovan Sequence (Mb) for 1000 Genomes Project Populations.

Population code	Super population code	Sequence*
ACB	AFR	1.3 (0.3)
ASW	AFR	1.3 (0.3)
ESN	AFR	1.2 (0.3)
GWD	AFR	1.2 (0.3)
LWK	AFR	1.3 (0.3)
MSL	AFR	1.3 (0.3)
YRI	AFR	1.2 (0.3)
CLM	AMR	1.0 (0.4)
MXL	AMR	1.2 (0.3)
PEL	AMR	1.2 (0.5)
PUR	AMR	1.3 (0.5)
CDX	EAS	1.3 (0.4)
CHB	EAS	1.5 (0.4)
CHS	EAS	1.4 (0.4)
JPT	EAS	1.3 (0.4)
KHV	EAS	1.4 (0.4)
CEU	EUR	0.9 (0.2)
FIN	EUR	1.0 (0.3)
GBR	EUR	0.9 (0.3)
IBS	EUR	0.9 (0.3)
TSI	EUR	1.0 (0.2)
BEB	SAS	1.4 (0.4)
GIH	SAS	1.3 (0.3)
ITU	SAS	1.5 (0.4)
PJL	SAS	1.4 (0.4)
STU	SAS	1.5 (0.5)

*All sequence numbers are Mb, and shown as mean (sd).

Table 3.5. Identified High-Frequency Neanderthal Haplotypes in Africans and Non-Africans.

Chromosome	Start	End	Population	Intersected Genes
1	24416980	24707234	African	*IL22RA1; IFNLR1; DQ597487; DQ601330; GRHL3; STPG1*
2	36292760	36575401	African	
4	64337344	64782205	African	
6	10661435	10758890	African	*C6orf52; PAK1IP1; TMEM14C; TMEM14B; SYCP2L*
7	145969388	146232193	African	*CNTNAP2*
8	36562358	36623582	African	
9	89380751	89505042	African	
10	131083147	131215166	African	
11	61084444	61236702	African	*DDB1; DAK; CYB561A3; TMEM138; TMEM216; CPSF7; SDHAF2*
13	64012971	64605115	African	*OR7E156P; AK127969; AK098560; AK057471; BC128161*
13	97650354	97770794	African	
19	20221202	20419091	African	*ZNF90; ZNF486*
20	45272036	45390945	African	*TP53RK; SLC2A10*
1	188604593	189609065	African-European	
2	71834241	71918700	African-European	*DYSF*
2	183690564	183959271	African-European	*FRZB;NCKAP1; DUSP19*
3	1781341	2029251	African-European	
3	5161136	5271025	African-European	*ARL8B;EDEM1*

3	73731951	73830221	African-European	
3	89587868	90134709	African-European	
4	8310419	8556248	African-European	*AX746755; ACOX3;TRMT44*
5	14378000	14430834	African-European	*TRIO*
5	21619181	22092619	African-European	*BC038535;CDH 12*
5	156931856	157147789	African-European	*ADAM19;SOX30 ; C5orf52*
5	180289312	180385894	African-European	*BTNL8*
6	75178778	75611224	African-European	*AF086303*
6	88821379	88978415	African-European	*CNR1*
6	93579793	93717561	African-European	
6	123980324	124110498	African-European	
8	67013268	67127559	African-European	*TRIM55;CRH*
9	92987159	93404340	African-European	*DQ584857; DQ597036;DQ601751;DQ579694; DQ585001; DQ593488; DIRAS2*
10	8440639	8607590	African-European	
10	92800188	93037716	African-European	*LINC00502; NUDT9P1;PCGF5*
12	52892007	52992241	African-European	*KRT5;KRT71; KRT74;KRT72*
13	31134256	31260771	African-European	*USPL1*
13	101402482	101526757	African-European	*NALCN-AS1*
14	32061112	32225793	African-European	*NUBPL*
15	77851227	77954689	African-	*LINGO1;*

			European	*BC038758*
15	98199342	98499010	African-European	*LINC00923; ARRDC4*
17	36051063	36230365	African-European	*HNF1B*
17	51656186	51797285	African-European	
17	53587616	53734332	African-European	
18	70613458	71023860	African-European	
20	25027516	25745794	African-European	*VSX1;AX747658 ; ENTPD6;PYGB; BC128043; ABHD12;GINS1; NINL;NANP; ZNF337;FAM18 2B*
1	208536469	209010804	non-African	
1	209723059	210633525	non-African	*CAMK1G ; LAMB3 ; MIR4260 ; G0S2 ; HSD11B1 ; TRAF3IP3 ; C1orf74 ; IRF6 ; DIEXF ; SYT14 ; SERTAD4-AS1 ; SERTAD4 ; HHAT*
1	232426559	232587793	non-African	*SIPA1L2*
2	159902551	160199938	non-African	*TANC1 ; WDSUB1 ; AK023566*
2	242790490	242946043	non-African	*PDCD1 ; CXXC11 ; AK097934 ; BC101234*
3	2107973	2577465	non-African	*CNTN4 ; CNTN4-AS2*
3	50194231	50400481	non-African	*SEMA3F ; GNAT1 ; SLC38A3 ; GNAI2 ; BC033528 ;*

				SEMA3B ; LSMEM2 ; IFRD2 ; HYAL3 ; NAT6 ; HYAL1 ; HYAL2 ; TUSC2 ; RASSF1 ; AB209621 ; ZMYND10 ; NPRL2 ; CYB561D2 ; TMEM115 ; CACNA2D2
3	190848688	191839248	non-African	*OSTN ; UTS2B ; CCDC50 ; PYDC2*
4	37982637	38245415	non-African	*TBC1D1*
4	38281886	38825449	non-African	*FLJ13197 ; KLF3 ; TLR10 ; TLR1*
4	129363292	130183094	non-African	*AK093416 ; PHF17 ; SCLT1 ; C4orf33*
4	130544530	131564821	non-African	*BC035172 ; BC041448*
4	167108469	167257857	non-African	
5	39351293	39412681	non-African	*C9 ; DAB2*
8	5595682	5969497	non-African	
8	6353897	6535411	non-African	*ANGPT2*
8	103201412	103774723	non-African	*RRM2B; AK095151; UBR5; ODF1; KLF10*
9	101800845	102132477	non-African	*COL15A1; TGFBR1 ; ALG2 ; SEC61B ; NAMA ; DQ673941 ; NAMA_1 ; NAMA_2*
9	112871891	113018578	non-African	*PALM2-AKAP2 ; AKAP2 ; C9orf152 ; TXN*
10	6292726	6388854	non-African	

10	6926250	7175095	non-African	
10	69137975	69564969	non-African	*CTNNA3 ; HI650153 ; DNAJC12*
10	87991148	88247502	non-African	*MIR346 ; WAPAL*
11	113195204	113275441	non-African	*TTC12 ; ANKK1*
11	119956224	120515929	non-African	*TRIM29 ; OAF ; POU2F3 ; TMEM136 ; ARHGEF12 ; GRIK4*
12	54092720	54333996	non-African	*CALCOCO1 ; HOXC-AS5 ; HOXC13*
12	102230076	103123339	non-African	*DRAM1 ; CCDC53 ; NUP37 ; PARPBP ; PMCH ; JX088243 ; IGF1*
12	114023983	114810778	non-African	*AK096932 ; BC007399 ; RBM19 ; TBX5*
12	125258201	125322018	non-African	*SCARB1 ; JB074994*
13	108086024	108734706	non-African	*FAM155A ; BC043519*
14	66324513	66480334	non-African	
14	86053923	86286227	non-African	*FLRT2*
14	86643788	86871982	non-African	
14	106241676	106308532	non-African	*IGH@; IGHE; DKFZp686O16217; IGHG1; IGHD; FLJ00382*
15	74467035	75255458	non-African	*ISLR; STRA6; HP11097; CCDC33 ; CYP11A1 ; BC013681 ; SEMA7A ; UBL7 ; UBL7-AS1 ; ARID3B ; AK095335 ; CLK3 ; EDC3 ;*

16	57015345	57320684	non-African	*CYP1A1 ; CYP1A2 ; CSK ; MIR4513 ; LMAN1L ; CPLX3 ; ULK3 ; SCAMP2 ; MPI ; FAM219B ; COX5A ; RPP25 CETP; NLRC5; CPNE2; FAM192A; RSPRY1;*
19	33091732	33743458	non-African	*ARL2BP; PLLP ANKRD27; RGS9BP; NUDT19; TDRD12; SLC7A9; JB050011; CEP89; C19orf40; RHPN2; GPATCH1;*
22	20766003	20945580	non-African	*WDR88; LRP3; SLC7A10 SCARF2; KLHL22; MED15*

Table 3.6. Regions Significantly Depleted of Neanderthal Sequence in S* and IBDmix.

Chromosome	S* Start	S* End	IBDmix Start	IBDmix End
1	102200000	114900000	105400000	120600000
2	201100000	211500000	--	--
3	76500000	90500000	74100000	89300000
7	106300000	124700000	106200000	123200000
8	53900000	66000000	49400000	66500000
18	25000000	41800000	--	--

Table 3.7. Average Amount of Neanderthal Ancestry (Mb) in CEU, CHB and YRI Detected by IBDmix Based on Different Population Panel Sizes.

Population size	CEU	CHB	YRI
10	49.2	50.6	16.4
20	54.6	56.9	18.0
30	58.6	58.7	18.6
50	58.5	60.2	19.1
70	59.5	61.6	19.5
90	60.1	61.9	19.8
Full size*	60.1	62.0	20.0

*The full size of CEU, CHB and YRI is 99, 103 and 108 respectively. All the sequence numbers are Mb.

Table 3.8. Approximate IBD and Non-IBD Likelihood.

GT_Archaic	***GT_Modern***	$\boldsymbol{P(\cdot\,|I)}$	$\boldsymbol{P(\cdot\,|nI)}$	$\boldsymbol{\frac{P(\cdot\,|I)}{P(\cdot\,|nI)}}$	$\boldsymbol{\frac{P_o(\cdot\,|I)}{P_o(\cdot\,|nI)}}$
AA	*AA*	p_A	${p_A}^2$	$\frac{1}{p_A}$	$\frac{1}{f_A}\frac{(1-\eta)(1-\varepsilon)+\eta\varepsilon}{1-\eta(1-\eta)}$
AA	*AB*	p_B	$2p_Ap_B$	$\frac{1}{2p_A}$	$\frac{1}{2f_B}\frac{(1-\eta)\varepsilon+\eta(1-\varepsilon)}{1-\eta(1-\eta)}$ $+\frac{1}{2f_A}\frac{(1-\eta)(1-\varepsilon)+\eta\varepsilon}{1-\eta(1-\eta)}$
AA	*BB*	*0*	${p_B}^2$	*0*	$\frac{1}{f_B}\frac{(1-\eta)\epsilon+\eta(1-\varepsilon)}{1-\eta(1-\eta)}$
AB	*AA*	p_A	${p_A}^2$	$\frac{1}{p_A}$	$\frac{1}{f_A[1+2\eta(1-\eta)]}$
AB	*AB*	*1*	$2p_Ap_B$	$\frac{1}{2p_Ap_B}$	$\frac{1}{2f_Af_B[1+2\eta(1-\eta)]}$
AB	*BB*	p_B	${p_B}^2$	$\frac{1}{p_B}$	$\frac{1}{f_B[1+2\eta(1-\eta)]}$
BB	*AA*	*0*	${p_A}^2$	*0*	$\frac{1}{f_A}\frac{(1-\eta)\varepsilon+\eta(1-\varepsilon)}{1-\eta(1-\eta)}$
BB	*AB*	p_A	$2p_Ap_B$	$\frac{1}{2p_B}$	$\frac{1}{2f_B}\frac{(1-\eta)(1-\varepsilon)+\eta\varepsilon}{1-\eta(1-\eta)}$ $+\frac{1}{2f_A}\frac{(1-\eta)\varepsilon+\eta(1-\varepsilon)}{1-\eta(1-\eta)}$
BB	*BB*	p_B	${p_B}^2$	$\frac{1}{p_B}$	$\frac{1}{f_B}\frac{(1-\eta)(1-\varepsilon)+\eta\varepsilon}{1-\eta(1-\eta)}$

Probabilities for a pair of genotypes having either 1 or 0 alleles shared due to identity by descent. Allele errors are independent and have probability $\boldsymbol{\eta}$ (a constant value; 0.01 for the analyses presented here) for archaic genome and $\boldsymbol{\varepsilon \geq 0}$ (described below) for modern human genomes. Genotypes have reference allele *A*, alternative allele *B*, true allele frequencies p_A and p_B, and error-added allele frequencies f_A and f_B. We estimate f_A and f_B with the observed allele frequencies, and we estimate p_B and $p_A = \mathbf{1} - p_B$ using the relationship $p_B = (\boldsymbol{f_B - \varepsilon})/(\mathbf{1 - 2\varepsilon})$.

Chapter 4. Extensive Modeling of Archaic Ancestry Deserts

Previous analyses of Neanderthal sequence in geographically diverse samples have described the heterogeneous distribution of archaic ancestry across the genome, and specifically the excess of large regions depleted for Neanderthal ancestry [30,31,33,34]. The frequency of these depletions, also referred to as “deserts”, is enriched compared to expectations based on models of neutral genetic drift that use standard estimates of human demographic history [30]. Desert regions also exhibit higher levels of background selection, human-Neanderthal sequence divergence, and are enriched for genes expressed in regions of the brain [30,33]. These initial results suggest a role for selection in the formation of desert regions. If deserts are produced by selection at specific loci, the human sequence at these loci could elucidate the genetic origin of exceptional human phenotypes.

However, complementary studies have also noted that acute demographic features can replicate specific aspects of the empirical data that correlate with deserts of archaic ancestry [34]. Specifically, models including intense bottlenecks around the time of admixture can produce levels of variation in Neanderthal ancestry across large regions that match patterns in the empirical data. These findings indicate that more complex demographic scenarios, which may not have a discernable effect on broad patterns of genetic diversity, can have acute effects on Neanderthal ancestry in contemporary populations.

In this analysis I use extensive modeling and coalescent simulations to explore a number of different demographic features that could affect the heterogeneous distribution of Neanderthal sequence across the genome. I investigate the effect of the admixing effective population size, the timing of bottlenecks around admixture, the presence of intermediary admixing populations, and the rate and frequency of admixture. Using results from these models, I produce a unified model and estimate parameters by applying Approximate Bayesian Computation and Bayesian Optimized Likelihood Free Inference. By extensively exploring demographic models, I aim to determine whether neutral processes alone are sufficient to explain the formation of archaic deserts, or if selection is a necessary component.

4.1 Results

4.1.1 Changes to Specific Demographic Features Provide Poor Fit to Empirical Data

I began by exploring 3 general structures for coalescent simulations, which include information about population size, growth rates, bottlenecks, and migrations. Importantly, all these models included a period of admixture from an archaic population into a non-African lineage (**Figure 4.1)**. Specifically, the 3 general structures I explored were a standard branching model to test the effect of the non-African population size, adapted from [78], a model that includes a very intense bottleneck around the time of admixture, adapted from [34], and a model that includes an early Out-of-Africa "Split Population" lineage that receives the admixture from the archaic lineage before re-merging with the main non-African lineage, to test the combined effect of population-size and bottleneck timing.

While certain features like the non-African effective populations size, lineage split times, and bottleneck start times vary within the models for each of the structures, several key parameters are fixed across all the models for the purpose of comparison. These fixed parameters include 1) an Out-of-Africa lineage that splits from the African lineage between 51-75 kya, 2) a single pulse of admixture from the archaic lineage into the non-African lineage, and 3) that the non-African lineage retains 2% archaic sequence in the modern samples (**Figure 4.1**).

For each simulation and set of parameters I generated 500 independent replicates of 10Mb chromosomes, for a total of 5Gb of sequence per haploid individual.

The performance of the models and the effects of their features were evaluated against empirical data from the Neanderthal callset from [30], using the distribution of desert sizes (**Table 4.1**), archaic ancestry in modern samples recovered from coalescent trees, and the estimated F4-ratio archaic ancestry proportion from sequence variants. While the measures of archaic ancestry are closely correlated, they can be sensitive to certain demographic features and provide complementary lines of analysis. During these simulations, I modulated the initial admixture level to control for a modern archaic ancestry level of ~2%, as reported in [29].

I investigated the effect of the non-African effective populations size (N_e) on desert distributions using population size estimates from several well established demographic models **(Figure 4.2A)** [78–81]. Models with smaller effective populations sizes are expected to result in more genetic drift and produce a greater frequency of large deserts. While this was the case for

the several values of N_e tested, none of the demographic models were able to replicate the full empirical distribution and all were deficient of the largest sized deserts (7-10Mb) when compared to the empirical data (**Figure 4.2A**).

In modeling an intense bottleneck around the time of admixture (T), I found that only instances where the bottleneck occurred shortly after admixture (50kya) was there a noticeable effect on the frequency of deserts (**Figure 4.2B**). This agrees with the findings from [34], and matches the expectation that placing a bottleneck closer to the time of admixture would limit the amount of recombination before introgressed segments were lost through drift, and would therefore produce larger deserts. While I found that certain models could replicate the empirical distribution for mid-sized deserts (5-8Mb; T=45kya), those that approached the empirical distribution for larger desert sizes also created an extreme excess of small and mid-sized deserts (**Figure 4.2B**). Therefore, though portions of the distribution could be matched for various bottleneck parameters, I was unable to match the full distribution.

Based on the behaviors of the previous models, I sought to create a synthesis of the two structures using parameters estimated from exome sequence data [78,79,91] and including an early Out-of-Africa population that receives the admixture from the archaic lineage before re-merging with the main non-African lineage (**Figure 4.1**). This "Split Population" acts as an intense bottleneck on the introgressed sequence, without affecting general patterns of population diversity. I evaluated the performance of this model, varying the effective population size (N_e) of the Split Population, with the expectation that reducing the size would lead to a more intense bottleneck and greater frequency of large sized deserts (**Figure 4.2C**). Again, while I found I could replicate the empirical distribution for small sized deserts (5-8Mb; N_e=100), any model that matched the empirical data for larger desert sizes showed an extreme excess of smaller desert sizes. Furthermore, because only a fraction of the Split Population was merged into the larger non-African lineage, the initial admixture level had to be set exceptionally high in order to guarantee an admixture proportion ~2% in modern samples. This is consistent with recent analyses that have posited initial admixture rates exceeded the current 2% archaic ancestry [35,38], and ancient human samples that have been recovered with higher proportions of Neanderthal ancestry [107,108].

4.1.2 Including Multiple Admixture Events Improves Desert Distributions But Distorts Admixture Proportions

Previous analyses have documented that levels of archaic sequence vary between global populations. In particular, East Asians have been documented to carry more Neanderthal ancestry than Europeans (EAS: 1.9%, EUR: 1.6%)[30–34,83], which some have interpreted as resulting from independent, population specific, admixture events [41,87,88]. Furthermore, analyses of Denisovan ancestry have demonstrated significantly greater overlap between deserts of Neanderthal and Denisovan sequence than expected by chance [30]. I therefore investigated how multiple admixture events could impact the distribution of deserts.

I began by picking the best fitting models from the previous analyses, and then incorporating an additional pulse of admixture, specific to the East Asian lineage. The timing of this pulse was consistent across all three model structures. Within each model, I varied the level of admixture for each of these pulses, with the first ranging from 0-20%, and the second from 0-10%. Models were evaluated on the distribution of the desert sizes and the proportion of archaic ancestry in the modern population, calculated from sampling the coalescent trees and analyzing the sequence data using the F4-ratio test [116].

For each simulation and set of parameters I generated 500 independent replicates of 10Mb chromosomes, for a total of 5Gb of sequence per haploid individual.

In order to examine the large number of model parameters, I evaluated the fit of the desert distribution for the simulated data to the empirical data by measuring the mean squared error (MSE) of the desert distributions. Using this framework, I identified parameter sets that minimized the MSE, and then compared their ancestry proportions to the empirical data.

For the Standard model, I found several parameters that minimized the mean squared error of the desert distribution (**Figure 4.3B)**, and when visualized showed good concordance with the empirical data for deserts 5-8Mb in size (**Figure 4.3C**). However, these models still showed a deficiency for deserts 9-10Mb (**Figure 4.3C**). Furthermore, the best fitting models yielded admixture proportions that varied dramatically from the empirical values. The F4-ratio estimated admixture proportions for Europeans and East Asians for the two best fitting models were 2.5% (EUR) and 4.6% (EAS), and 0.35% (EUR) and 9.7% (EAS)(**Figure 4.3D**).

Similarly, in the Bottleneck model, the empirical distribution for mid-sized deserts could be reproduced in the simulated data, but these models remained deficient for deserts 9-10Mb in

size (**Figure 4.4B,C)**. As well, the models with the best fitting desert distributions required admixture rates that gave exceptional modern admixture levels and relied heavily on a single population-specific pulse (**Figure 4.4D**). The F4-ratio test in the best fitting model yielded admixture estimates of 0.026% in EUR, and 1.26% in EAS, with the mean proportion of archaic sequence recovered from the coalescent trees only 0.4%.

Models utilizing a Split Population produced some of the closest fitting distributions for deserter sizes (**Figure 4.5B**). These distributions fell within the 95% CI for the empirical data (**Figure 4.5C**). Again, however, the necessary levels of admixture generated F4-ratio admixture proportions in modern samples that are inconsistent with empirical values (**Figure 4.5D**). The best fitting model required an initial admixture pulse of 3% and an EAS specific admixture pulse of 10%. These parameters resulted in modern admixture proportions of 0.23% (EUR) and 10.25% (EAS).

4.1.3 Combining Demographic Features and Admixture Rates to Estimate Best Fitting Models

The findings from the previous analyses made clear that independent changes to single parameters of the demographic models were insufficient to generate data that could match all the key features of the empirical data, namely the admixture level and desert distribution. I therefore aimed to combine the demographic effects and effects of admixture rates to estimate parameters. To do this, I utilized the structure from the Split Population model, which offers the greatest flexibility for incorporating major demographic features like the role of bottlenecks, independent admixture events, and changes in population size (**Figure 4.6**). Using this structure, I estimated key features of the model such as the admixture levels (n1 and n2), the fraction of the Split Population that merges back into the main non-African lineage (f), and the size of the Split Population (N_e).

To optimize these parameters and identify a best fitting model, I used the Bayesian Optimization for Likelihood-Free Inference (BOLFI) [117]. BOLFI is conceptually very similar to Approximate Bayesian Computation (ABC) [118–120], in that it involves generating data from a model given certain parameters, comparing the results to empirical data using a distance measure, and updating the model parameters to improve the fit of the simulated results to the empirical ones (**Figure 4.7**). However, BOLFI differs from ABC in that it models the conditional distribution of the distances given the model parameters and uses Bayesian approximation to find

regions of the parameter space where that distance tends to be smallest in a process of targeted acquisition. This dramatically reduces the number of simulation calls during parameter estimation, and speeds up the process of reaching a best-fitting model.

The fit of the simulated data to the empirical data was evaluated using summary statistics related to the desert frequency for window sizes 5-10Mb, the admixture level in modern samples estimated from calls using S*, and general population genetics measures π and F_{ST} (**Figure 4.7; Table 4.1**). Empirical values were calculated from 1000 Genomes Project Phase 3 data [68]. I evaluated desert and admixture proportions in simulated data using S* to match the process in the empirical data. General population genetics measures π and F_{ST} were included to set boundaries on demographic parameters.

For each simulation and set of parameters I generated 500 independent replicates of 10Mb chromosomes, for a total of 5Gb of sequence per haploid individual. I ran 400 initial simulations sampling parameter values from a uniform distribution (**Figure 4.8**). I then used these 400 sets of parameters and summary statistics to generate the BOLFI model, conditioned on the distances from these simulations. Based on this model, BOLFI sampled an additional 100 parameters sets, updating the model after every 10 new parameter sets.

The targeted acquisition of parameters using BOLFI quickly converged on a best model after only 10 evaluations. The parameters that BOLFI converged upon maximized values for n1, n2, *f,* and N_e (**Table 4.2**). Notably, the desert and admixture proportions (**Figure 4.9; Table 4.2**) do not match the empirical data well, with both metrics being reduced compared to the empirical data. These results suggest that the limits set on the model parameters are too restrictive for BOLFI to find the global optimum. Alternatively, the summary statistics I chose to evaluate the goodness-of-fit for models have constrained the parameter estimates. Specifically, including π and F_{ST} could be strongly restricting BOLFI's search of the parameter space, since these are the only summary statistics that show reasonable fit to the empirical data in the model identified by BOLFI.

4.1.4 Evaluating Fit Using Desert and Admixture Proportions Yields Extreme Demographic Models

As a parallel investigation, and to evaluate how different combinations of summary statistics could affect the parameter estimates, I analyzed the initial 400 simulations with

ABCreg [121], which implements ABC using a linear regression to approximate the posterior distribution of parameters from summary statistics. The same summary statistics were used for ABCreg as were used for BOLFI and simulated data were evaluated against the same empirical data as BOLFI (**Table 4.1**).

By using all summary statistics in the ABC fit, I found a strong boundary set on the estimation of n1, n2, *f*, and N_e (**Figure 4.10**). Specifically, while posterior estimates for N_e varied, n1 and n2 were fixed at their maximum (n1 = 0.5, n2= 0.1) and *f* was fixed at near 0 (**Table 4.3**). In such scenarios, the Split Population has little overall impact on the final sampled modern populations. Interestingly, the limitation of the Split Populations role seemed to be predominantly driven by its impact on general population statistics π and F_{ST}. When just these measures were included as summary statistics for ABCreg, again N_e varied independently, while n1, n2, and *f* were kept at near 0 (**Figure 4.10; Table 4.4**). This suggests that a significant bottleneck, the type necessary to cause large depletions of archaic sequence, could not occur and contribute substantially to modern populations without its impact being detected in measures like π and F_{ST}. Notably, parameters estimated using all the summary statistics showed the worst fit to the empirical data (**Table 4.3**). Distance measures dropped dramatically when estimates were based on π and F_{ST} alone (**Table 4.4**).

Due to the strong limiting effect of π and F_{ST} on the posterior distribution of model parameters, I repeated the ABCreg analysis only including summary statistics related to desert and admixture proportions. This resulted in greater variation among the posterior distributions for values of n1, n2, *f*, and N_e (**Figure 4.10**). I observed, however, a predominant trend towards low estimates for N_e, with a median value less than 100 haplotypes (**Figure 4.11**). This finding is consistent with the expectation that a strong bottleneck, or intense selection, is necessary for the expansive and heterogeneous removal of archaic sequence. There was also a strong inverse correlation between estimates of *f* and n1 (**Figure 4.10**), which is likely driven by conditioning on the admixture level in the final populations. Importantly, the majority of the posterior distributions for *f* and n1 appear as extreme values. The median estimate for *f* is 0.72 (**Figure 4.11**); indicating 72% of the final Eurasian population is derived from the Split Population. As well, the median estimate for n1 is 0.44 (**Figure 4.11**), indicating a required level of admixture 20x higher than that found in modern human samples.

Distance measures for models estimated from desert and admixture distributions alone were among the lowest evaluated (**Table 4.5**). As the desert proportions were a specific feature of interest, I examined these models more closely for fit to the empirical data. Best fitting models were selected from the posterior estimates based on Euclidean distance to the empirical data (**Figure 4.12; Table 4.2**). The best fitting model (n1 = 0.5, n2 = 0.0027, f = 0.25, N_e = 31) showed a strong fit with the empirical desert distribution, falling within the 95% CI for the empirical data and matching at mid- and large-sized deserts (**Figure 4.9**). Compared to empirical admixture proportions, this model matched closely the European admixture proportion, but was reduced compared to the empirical East Asian admixture proportion. There was a clear trade-off in the fit of the models between the admixture and desert proportions. Top fitting models showed either increased admixture proportions and reduced desert proportions, or reduced admixture proportions and increased desert proportions (**Figure 4.9; Table 4.2**). Overall, the results of fitting using ABCreg and desert and admixture proportions show a strong requirement to minimize the effective population size of the Split Population, and demonstrate that with extreme values it is possible to match the empirical data across numerous measures of fit.

4.2 Discussion

The regions significantly depleted of archaic ancestry are reported to be enriched in modern humans compared to simulations of neutral models of human demographic history [30]. The fact that these regions also exhibit higher levels of background selection, human-Neanderthal sequence divergence, and are enriched for genes expressed in regions of the brain [30,33] has promoted the interpretation that these regions are functionally relevant and shaped by selection. However, some studies have also noted that acute demographic features, like intense bottlenecks, can replicate specific aspects of these deserts of archaic ancestry [34].

In order to clarify the effect of demographic processes on the frequency of archaic deserts, I explored more complex model structures that vary the admixing effective population size, the timing of bottlenecks around admixture, the presence of intermediary admixing populations, and the rate and frequency of admixture. Initial results, comparing simulated data to empirical data over just a few key features like the distribution of deserts and admixture proportion, supported the conclusion that neutral processes could not replicate the empirical data.

However, comprehensive searches of the parameter space with ABC and BOLFI identified certain sets of parameters that could generate data to match the empirical results across features like genetic diversity, admixture proportion, and desert distributions. The parameters these methods estimate are, however, at the extreme end of what would be expected. The closest fitting data are generated by a model with an effective population size of 16 diploid individuals for the human admixed population and an initial Neanderthal admixture level of 50%. This level of admixture is 5x to 10x higher than that found in any ancient human sample to date [107,108]. It is important to note, however, that these models represent simplifications of complex demographic processes, and so estimated parameters should not be interpreted as historical fact.

Another interesting feature of the parameters estimated by BOLFI and ABC is that they consistently produce admixture levels that are higher in European samples than in East Asian ones. This is contrary to all previous estimates of Neanderthal ancestry in modern populations [30–34]. Analyses using statistical and simulation approaches have indicated that models incorporating multiple pulses of admixture are necessary to explain the ~20% enrichment of Neanderthal ancestry in East Asians [30,33,40–42]. Though models explored in BOLFI and ABC analyses are allowed to include a second admixture event specific to East Asians, the second admixture level is consistently kept below 1%. This may reflect the strong limitations recurrent independent admixture events place on the formation of deserts through neutral processes.

The fact that best fitting models regularly support a single, strong, admixture event is consistent with recent analyses using reference-free methods of archaic sequence detection (see chapter 3). Using new methods for detecting archaic sequence, we found the enrichment of Neanderthal ancestry in East Asians compared to Europeans to be less than 10%. This modest enrichment can be parsimoniously explained by a single wave of Neanderthal admixture occurring shortly after the Out-of-Africa dispersal, followed by dilution from a non-admixed Basal Eurasian lineage [39,106]. This type of demographic feature was not explored in this current study, but could be readily incorporated in future models.

Notably, the extremely small effective population size estimated from these simulations could be analogous to models of strong selection against introgressed sequence. Several studies suggest the rapid loss of Neanderthal ancestry overall was due to the accumulation of weakly deleterious alleles in the Neanderthal genome [35,38]. Modeling the loss of Neanderthal

sequence as a neutral process may therefore be inaccurate, and instead should include selection. What exactly this selection should look like at desert loci—weak selection across many variants or strong selection at a few loci—remains unclear.

It is also important to consider the role for structural variation, like large inversions, to prevent introgression at desert loci by suppressing recombination. Inversions could be present on either the human or archaic lineages. Considering the overlap of Neanderthal and Denisovan deserts [30], large inversions seem unlikely to explain all of the archaic depletions found to date, but remain a formal possibility.

Overall, the results of my analyses indicate the possibility for neutral demographic features to generate large depletions of archaic sequence consistent with those found in the empirical data. Importantly, the formation of these deserts appears strongly limited by recurrent independent admixture events. These results have important implications about admixture between humans and Neanderthals overall, supporting a single admixture event followed by dilution in Europeans. This model has not been formally explored here, however. Conversely, there is a significantly greater overlap of regions depleted of Neanderthal ancestry and Denisovan ancestry [30], which has also not been formally explored in these models. Future work should necessarily investigate these additional demographic features, as well as the role for weak selection in the formation of deserts. Deserts of archaic sequence represent especially interesting loci for studying human evolutionary history. Fully understanding the etiology of these features can shape our understanding of human-Neanderthal admixture broadly, and could elucidate the genetic origin of exceptional human phenotypes.

4.3 Materials and Methods

4.3.1 Coalescent simulations to model neutral demographic histories

To model different demographic histories, I used the recently developed coalescent simulator *msprime* [114], which improves over the original *ms* simulator [122] by utilizing sparse trees and coalescence records, enabling exact simulation of the coalescent with recombination over chromosome-sized regions for large sample sizes. Significantly, this method improves upon sequentially Markov coalescent approximation simulators, like MaCS [77], which do away with long rage linkage information and therefore do a poor job modeling features

such as the admixture block length. As well, *msprime* is designed to be used through its Python API, which makes coding demographic models, generating data, and analyzing results more easily integrated into a single pipeline.

For demographic models, I investigated 3 general structures, which I refer to as 1) a "Standard" branching model, 2) a "Bottleneck" model, and 3) a "Split Population" model. The Standard branching model is adopted from [78,79,91] and is distinguished by a rapid population expansion in modern human lineages around 5kya to simulate the effect of agriculture, as well as a historically smaller effective population size in East Asians compared to Europeans, and a larger effective population size in Africans to capture greater genetic diversity. The Bottleneck model is adapted from [34], and specifies constant and equal effective population sizes for all modern populations, except for a short and intense bottleneck when the Eurasian lineage is reduced to $1/100^{th}$ its initial size for 20 generations. The Split Population model is a synthesis of these two models, primarily using the structure and population sizes defined in [78,79,91], but also including a lineage which separates from the Out-of-Africa population 1ky before admixture with Neanderthals, receives the admixture pulse from Neanderthals, and then remerges with the main Eurasian lineage 1ky later at a proportion less than or equal to 1. Varying the size and rate at which this Split Population merges with the Eurasian lineage affects the amount of drift acting upon introgressed Neanderthal sequence, similar to an intense bottleneck.

Across all model structures, several features remained consistent. All models included African/Yoruba (AFR), European (EUR), and East Asian (EAS) modern populations, which were sampled at levels to match their representation in the 1000 Genome Project Phase 3 [68] (EUR=503 diploid genomes, EAS=504 diploid genomes, AFR=108 diploid genomes). As well, the models included 1 Chimpanzee lineage, 1 Denisovan lineage, and 2 Neanderthal lineages. The 2 Neanderthal lineages represent an "introgressing" Neanderthal and a "reference" Neanderthal, to account for differences in accumulation of mutations between the actual source of Neanderthal admixture in modern populations and the reference genome used for analyses. These archaic lineages are sampled at 55kya for the introgressing Neanderthal (Neand_i), and 125kya for the reference Neanderthal (Neand_r).

I used genome-wide averages for features like mutation rate and recombination rate, $1.2x10^{-8}$bp/generation and 1cM/Mb respectively. The generation time was defined as 25yrs/generation.

From simulations I collected sequence data in vcf format, EIGENSTRAT format for use with AdmixTools [116], and "true calls" of introgressed segments identified using the coalescent trees. While *msprime* simulates haploid genomes by default, these were combined into diploid genomes within populations when generating vcfs for ease of analysis with archaic sequence detection methods like S*.

For ease of use and analysis, I combined the *msprime* simulations with custom scripts into a pipeline for running admixture models, generating output data in an array of formats, calling introgressed sequence using S*, estimating admixture proportions with F4-ratio test, and collecting general population genetics statistics like π and F_{ST}. Scripts and documentation can be found here:
https://github.com/abwolf/msprime_scripts

4.3.2 S-pipeline analyzes simulated data*

The method for identifying introgressed sequence in simulated data using S* is a two-stage approach to first identify candidate introgressed sequences using the statistic S* [32,33] and then refine this set of haplotypes by calculating a p-value to quantify whether a putatively introgressed haplotype matches an archaic sequence more than expected by chance.

The significance of S* scores is assigned based on a null distribution generated by simulating sequence data under a standard demographic model of non-African populations without archaic admixture. In previous implementations of S*, these simulations were carried out under a grid of recombination rates and population diversity (represented by the number of segregating sites in a 50kb window for one non-African individual and the 108 Yoruba reference individuals), and a generalized linear model was built to the grid of S* quantiles [30,33].

Because I was experimenting with multiple different models, and was frequently changing model parameters, which require the development of new null datasets, I sought to simplify and speed up the process of generating a null distribution for S* scores. Specifically, I analyzed vcf data for EUR and EAS samples from null simulations using the S* software, assigning S*-scores to 50kb windows at overlapping 10kb steps. I then binned these windows based on the source population and the number of segregating sites for the target non-African individual and the 108 African reference individuals. Note, this set of segregating sites is not the same as the sites used to calculate the actual S*-score, since variants that segregate in Africans

are masked during S* scoring. However, S*-snps are a subset of these segregating sites, and the two values are correlated. I generated separate empirical cumulative distribution functions (ECDFs) of the S*-scores for each of these bins, and estimated the S*-score equivalent to a p-value < 0.01 for matching 50kb windows. When analyzing data from admixture models, 50kb windows are matched to the null S* ECDF based on the count of segregating sites in that window, and are retained if the S* score is in the 99th percentile of the null distributions.

Along with assigning a significance to the S* score for a given 50kb window, the null simulations are also used to assign a significance to the sequence match between a 50kb window in a target sample and the archaic reference genome. The calculation for match percentage also operates on 50kb windows at overlapping 10kb steps. For a given window, I calculate an "informative site count", which is the number of sites that segregate in the target populations (i.e. EUR and EAS) and where the archaic reference genome carries the derived allele. Importantly, I do not mask sites that segregate in the African population like I do for S*, since this leads to too a highly discretized distribution with a high probability of windows with few variants being non-significant. The target samples are then analyzed one haploid genome at a time, and a match percentage is assigned to the given 50kb window for that haplotype. Windows from the null distribution are binned together based the their informative site count and combined into a database which can be referenced for assigning an empirical p-value to windows from admixture models with matching informative sites counts.

After the null distributions for S*-scores and match-percentages are generated, it is possible to analyze the S*-scores for data from admixture models. Analyzed in 50b windows at overlapping 10kb steps, the admixture data is assigned S*-scores, then assigned empirical p-values based on the null distributions. In this analysis, windows were considered significant with S* p-value < 0.01 and match-percent p-value < 0.05. Significant windows are then merged together at the haplotype level to generate bedfiles of introgressed segments for each individual.

Scripts and documentation can be found here:
https://github.com/abwolf/msprime_scripts
https://github.com/abwolf/matchpval_scripts
https://github.com/lparsons/archaic_match

4.3.4 Calculating level of archaic ancestry depletion

I compared the results of the coalescent simulations to empirical data from [30] (**Table 4.1**), specifically the calls of Neanderthal ancestry in East Asian and European samples. I used the same approach to identify desert regions as described in the methods of that paper. Specifically, I calculated the average percent of introgressed bases across all European and East Asian individuals in overlapping windows of size 5-10Mb at 100kb steps. I additionally required that at least 90% of the sequence in each window was callable given the filters described in the primary analysis. For each window size, I counted the number of windows with an average percent introgression less than $10^{-3.5}$, representing the 99th-percentile. As well, I bootstrap-resampled 5000 times from the empirical data for each window size to estimate 95% confidence intervals around the empirical distribution.

For simulated data, each parameter set generated 500 replicates of 10Mb chromosomes, equivalent to 5Gb, and sampled 216 AFR, 1006 European, and 1008 East haploid genomes, the equivalent of 503 (EUR) and 504 (EAS) diploid genomes. I then measured the average percent introgressed bases for a given window size (5-10Mb) using calls made in EUR and EAS simulated individuals and counting the proportion of windows with an average percent introgression less than $10^{-3.5}$. However, when varying the window size, only one window per simulated chromosome was used so that each window represented an independent simulation of a given model. In other words, having a 10Mb chromosome, if I wanted to look at the percent introgression in only 1Mb windows, I took the first 1Mb of the original 10Mb simulation.

4.3.5 Calculating Admixture Proportion From Sequence Data

For simulated data, archaic admixture proportion was estimated using 1) the introgressed segments identified from the coalescent trees with 100% power, or 2) the S*-calls, or 3) from the sequence data using AdmixTools F4-ratio test [116]. For the F4-ratio test, I estimated the admixture proportion (α) as:

$$\frac{f4(\mathbf{Denisovan,Chimp:EUR,AFR})}{\mathbf{f4(Denisovan,Chimp:Neand\_i,AFR)}} \text{ or } \frac{f4(\mathbf{Denisovan,Chimp:EAS,AFR})}{\mathbf{f4(Denisovan,Chimp:Neand\_i,AFR)}}$$

for European and East Asian samples respectively . The sequence data for the F4-ratio test was generated separately from the main simulation, and consisted of 2000 independent replicates of 500kb chromosomes. The α was calculated for sets of 20 chromosomes, and these were then averaged for a given simulation to provide the estimated α for the given parameter set.

4.3.3 BOLFI pipeline estimates model parameters

I adopted the structure from the Split Population model to estimate best fitting parameters because it offered the greatest flexibility for incorporating major demographic features like the role of bottlenecks, independent admixture events, and changes in population size. For example, the standard branching model from [78] can be replicated in the Split Population model by setting the Split Population N_e to be the same size as the non-African lineage, and setting the proportion with which the Split Population merged back into the main lineage to 1. Using this structure, I estimated key features of the model such as the admixture level (n1 and n2), the fraction of the split-population that merged back into the main non-African lineage (f), and the size of the split-population (N_e) (**Figure 4.6**).

To estimate parameters for a best fitting model I used the Bayesian Optimization for Likelihood-Free Inference (BOLFI) [117,123]. BOLFI is conceptually very similar to Approximate Bayesian Computation (ABC), in that it involves generating data from a model conditional on certain parameters. It then compares the results to empirical data over a series of summary statistics using a distance measure (Euclidean), and updates the model parameters to improve the fit of the simulated results to the empirical ones. BOLFI differs from ABC in that it uses only a limited number of initial simulations before generating a model of the conditional distribution of the distances given the initial simulation parameters. It then uses Bayesian approximation to find regions of the parameter space where that distance tends to be smallest. This dramatically reduces the number of simulation calls during parameter estimation, and speeds up the process of reaching a best-fitting model.

The fit of the simulated data to the empirical data was evaluated using summary statistics related to the desert proportion for window sizes of 5-10Mb, the admixture level in modern samples estimated from calls using S*, and general population genetics measures π and F_{ST} (**Table 4.1**). I evaluated the desert and admixture proportions using S* calls made on simulated vcf data to match the process for calling deserts in the empirical data. General population statistics π and F_{ST} were included to set boundaries on demographic parameters and were calculated from the same vcf data used to make S* calls (1000 Genome Project Phase 3 [68]). As we had seen in previous results, extreme demographic models could replicate certain features of the desert distributions, but would not match other patterns of population diversity and divergence.

Each simulation and set of parameters generates 10Mb of sequence and 500 independent replicates. BOLFI sampled 400 parameter sets to generate its model of the conditional distribution of the distances. The initial simulations sampled parameters for n1, n2, f, and N_e from uniform distributions with bounds [0.0-0.50], [0.0-0.1], [0.0-1.0], and [1-2758] respectively. After the initial 400 parameter sets, BOLFI generated a model of the conditional distribution of the distances and began a targeted acquisition of parameters from the prior to minimize these distances. BOLFI updated this model at every additional 10 parameter sets.

4.3.4 ABCreg estimates posterior values for parameters from simulated data

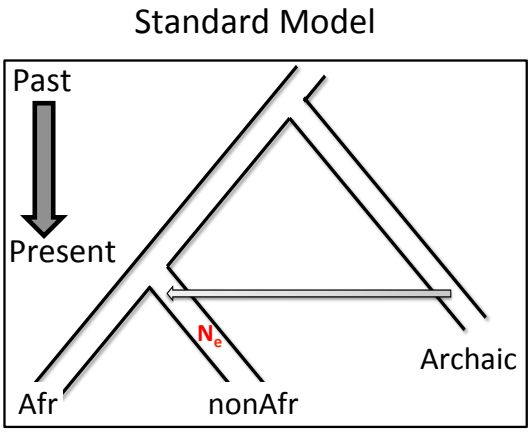
I analyzed the initial 400 simulations used for BOLFI, where parameters were sampled from uniform distributions, with ABCreg [121], which implements ABC using a linear regression to approximate the posterior distribution of parameters from summary statistics. In brief, a specified proportion of random simulations are taken from a prior distribution, and simulations that generate summary statistics that minimize the distance to the empirical data are "accepted" at some tolerance. The parameters from these accepted simulations are used to generate a regression of the distance between the simulated and observed summary statistics. Regression-adjusted parameter values are then back-transformed from the accepted prior values as estimates of the posterior. The main advantage of regression ABC is speed, as opposed to rejection-sampling where a finite number of parameters are chosen to minimize the distance measure and therefore a large number of simulations must be run to obtain reasonable parameter estimates. Furthermore, the flexibility of regression ABC allows one to explore the effect of combinations of summary statistics, which can bias estimations in subtle ways.

The same summary statistics were used for ABCreg as were used for BOLFI, including the desert frequency for window sizes of 5-10Mb, the admixture level in modern samples estimated from calls using S*, and general population genetics measures π and F_{ST}. I also analyzed the data using just π and F_{ST} or just the desert and admixture proportions separately. The simulated data were evaluated against the same empirical data as BOLFI (**Table 4.1**).

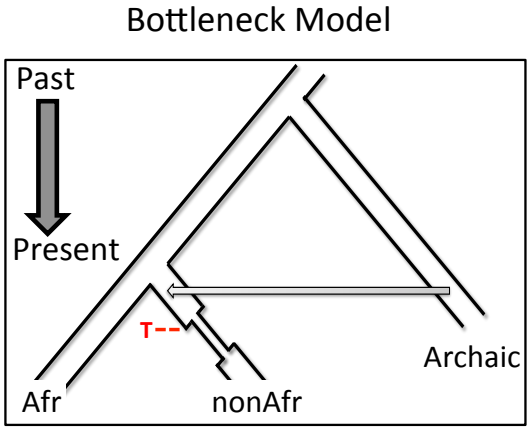
ABCreg transforms the parameters simulated from the prior using a tangent transformation to assure that the posterior distribution is contained within the bounds of the prior. When running ACBreg, I experimented using un-normalized summary statistics, and normalizing them by the natural-log. I specified the tolerance for acceptance at 10%, which

represents the fraction of draws from the prior to accept, and from which the regression is calculated.

4.4 Figures and Tables



J.A. Tennessen, et al. **Science**. (2012)



S. Sankararaman, et al. **Nature.** (2014)

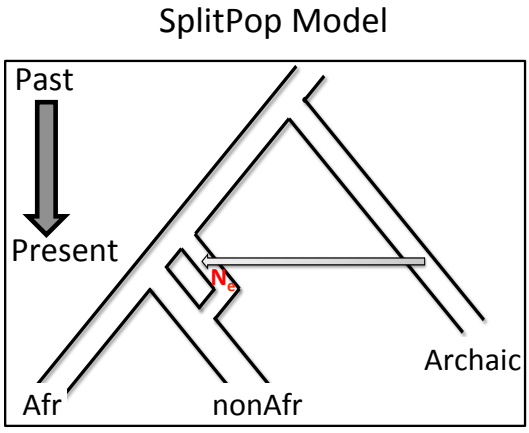


Figure 4.1. Basic structures of simulated demographic models. The basic structure for each of the simulated demographic models is show, highlighting the key variable parameter in each. In the Standard model [78] the non-African effective population size (N_e), in the Bottleneck model [34] the timing of the bottleneck relative to admixture (T), and in the Split Population model the effective population size of the Split Population.

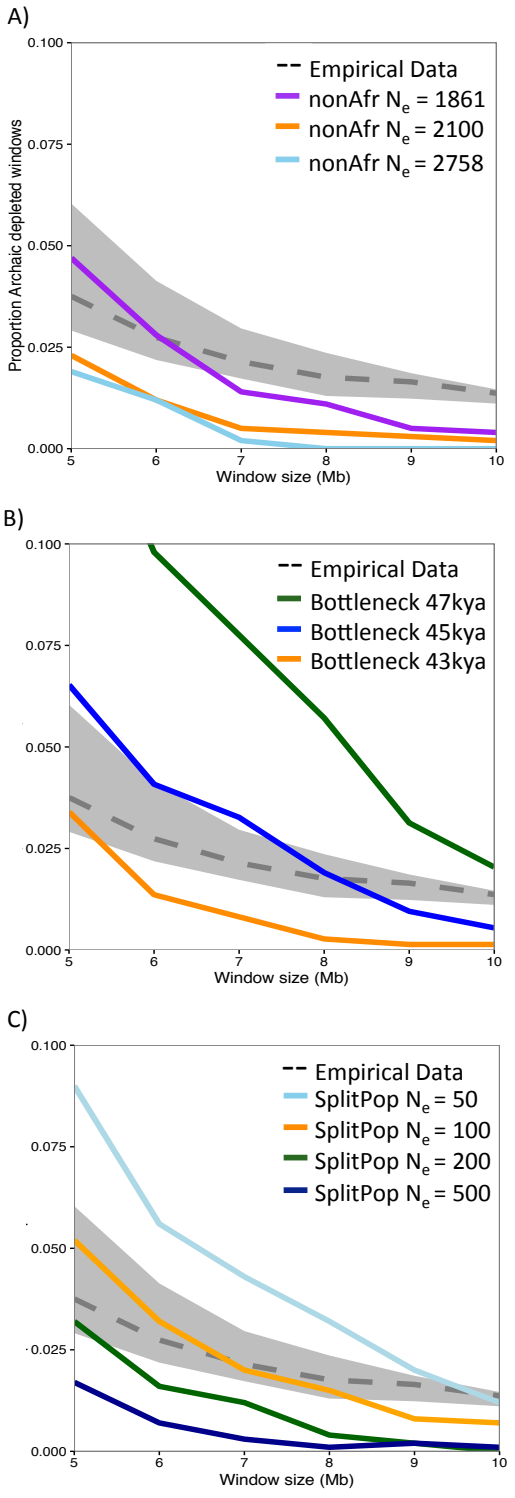
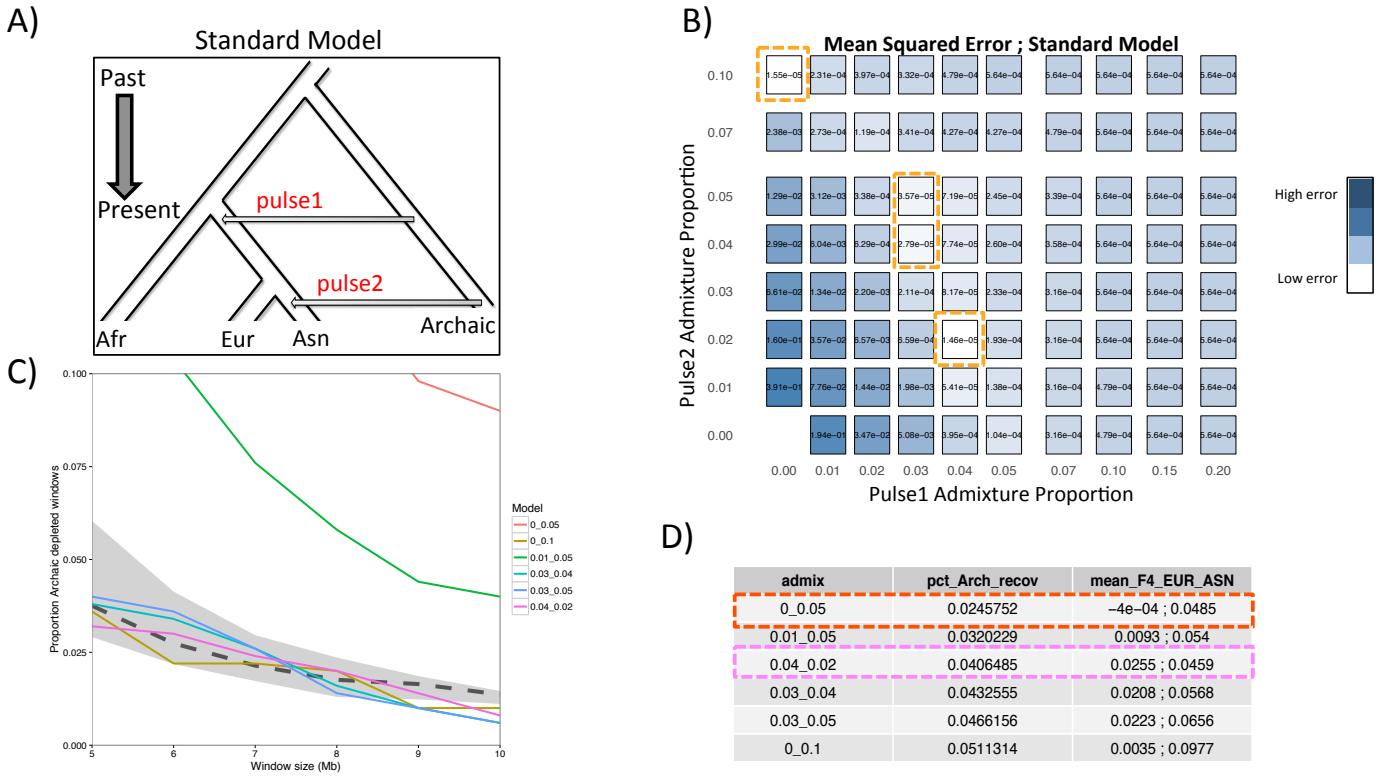
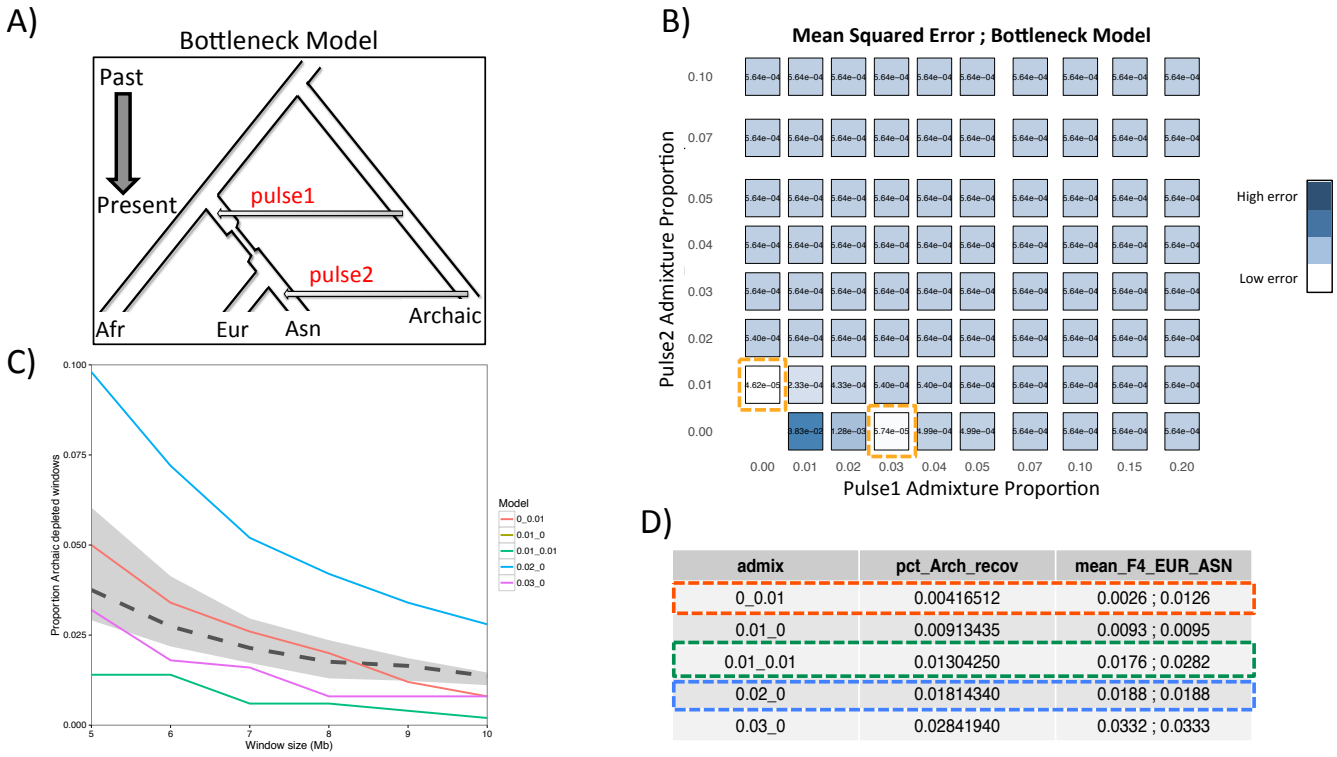


Figure 4.2. Desert distributions for simulations compared to empirical data. The desert distributions for best fitting parameters for each of the model structures. (A) Standard Model, (B) Bottleneck Model, (C) Split Population Model. Colored lines correspond to simulated data. The dashed black line represents empirical data, with bootstrapped 95% CI.



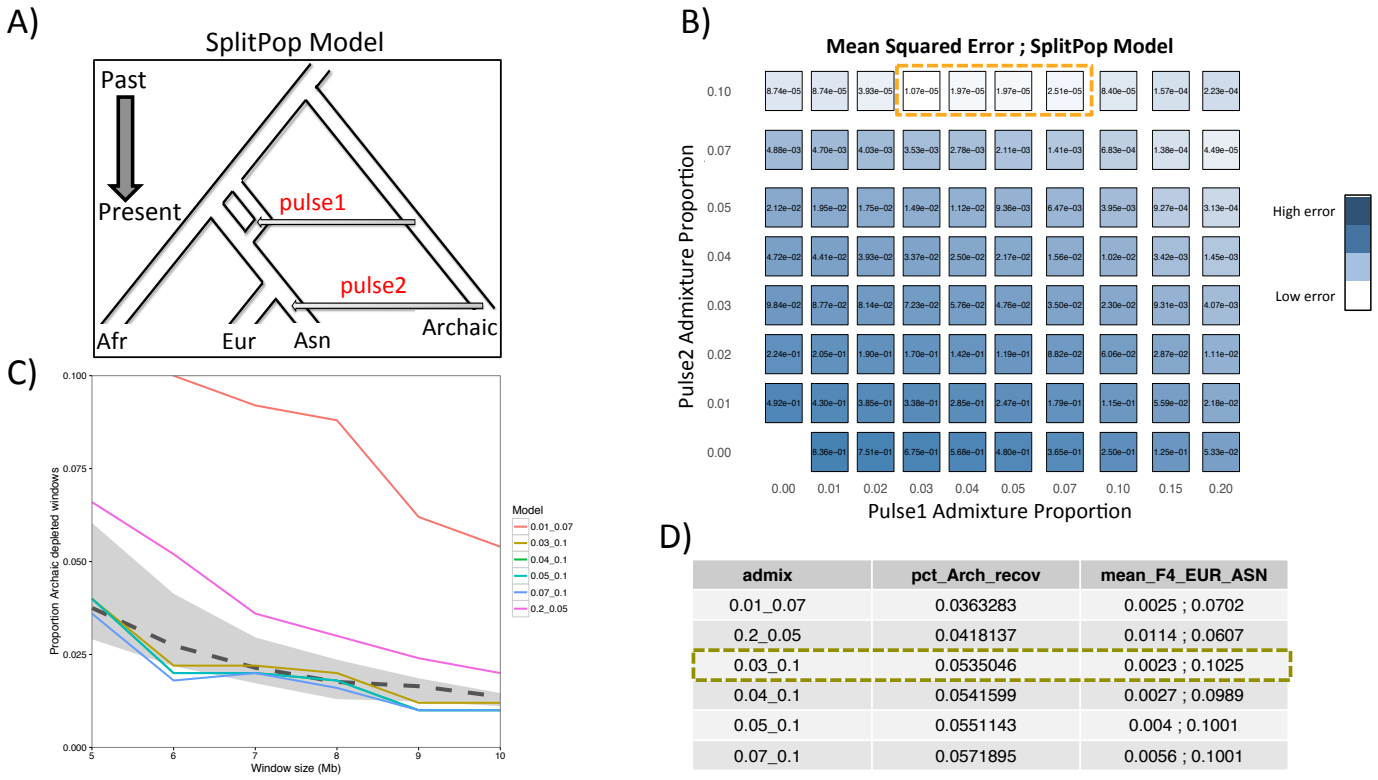
admix	pct_Arch_recov	mean_F4_EUR_ASN
0_0.05	0.0245752	−4e−04 ; 0.0485
0.01_0.05	0.0320229	0.0093 ; 0.054
0.04_0.02	0.0406485	0.0255 ; 0.0459
0.03_0.04	0.0432555	0.0208 ; 0.0568
0.03_0.05	0.0466156	0.0223 ; 0.0656
0_0.1	0.0511314	0.0035 ; 0.0977

Figure 4.3. Fitting Standard Model based on desert distribution requires imbalanced admixture levels.
(A) Standard Model structure, with 2 independent pulses of admixture.
(B) Heat map of Mean Squared Error calculated for desert distributions comparing empirical data to simulated data with 2 independent pulses of admixture of varying levels. Lowest MSE values are highlighted.
(C) Plots of desert distributions for models with lowest MSE values. The dashed black line represents empirical data, with bootstrapped 95% CI.
(D) Table listing simulation parameters (admix1_admix2), the percent archaic sequence recovered from the coalescent trees, and the calculated mean F4-ratio alpha for simulated EUR and ASN individuals. Best fitting and worst fitting model from (C) are highlighted.



admix	pct_Arch_recov	mean_F4_EUR_ASN
0_0.01	0.00416512	0.0026 ; 0.0126
0.01_0	0.00913435	0.0093 ; 0.0095
0.01_0.01	0.01304250	0.0176 ; 0.0282
0.02_0	0.01814340	0.0188 ; 0.0188
0.03_0	0.02841940	0.0332 ; 0.0333

Figure 4.4. Intense bottleneck replicates empirical desert distribution with single-population specific admixture pulse.
(A) Bottleneck structure, with 2 independent pulses of admixture.
(B) Heat map of Mean Squared Error calculated for desert distributions comparing empirical data to simulated data with 2 independent pulses of admixture of varying levels. Lowest MSE values are highlighted.
(C) Plots of desert distributions for models with lowest MSE values (red, purple) or admixture levels closest to empirical data (blue, green). The dashed black line represents empirical data, with bootstrapped 95% CI.
(D) Table listing simulation parameters (admix1_admix2), the percent archaic sequence recovered from the coalescent trees, and the calculated mean F4-ratio alpha for simulated EUR and ASN individuals.



D)

admix	pct_Arch_recov	mean_F4_EUR_ASN
0.01_0.07	0.0363283	0.0025 ; 0.0702
0.2_0.05	0.0418137	0.0114 ; 0.0607
0.03_0.1	0.0535046	0.0023 ; 0.1025
0.04_0.1	0.0541599	0.0027 ; 0.0989
0.05_0.1	0.0551143	0.004 ; 0.1001
0.07_0.1	0.0571895	0.0056 ; 0.1001

Figure 4.5. Split Population model replicates the empirical desert distribution with balance of multiple admixture pulses.
(A) Split Population structure, with 2 independent pulses of admixture.
(B) Heat map of Mean Squared Error calculated for desert distributions comparing empirical data to simulated data with 2 independent pulses of admixture of varying levels. Lowest MSE values are highlighted.
(C) Plots of desert distributions for models with lowest MSE values. The dashed black line represents empirical data, with bootstrapped 95% CI.
(D) Table listing simulation parameters (admix1_admix2), the percent archaic sequence recovered from the coalescent trees, and the calculated mean F4-ratio alpha for simulated EUR and ASN individuals. Values for best fitting model are highlighted.

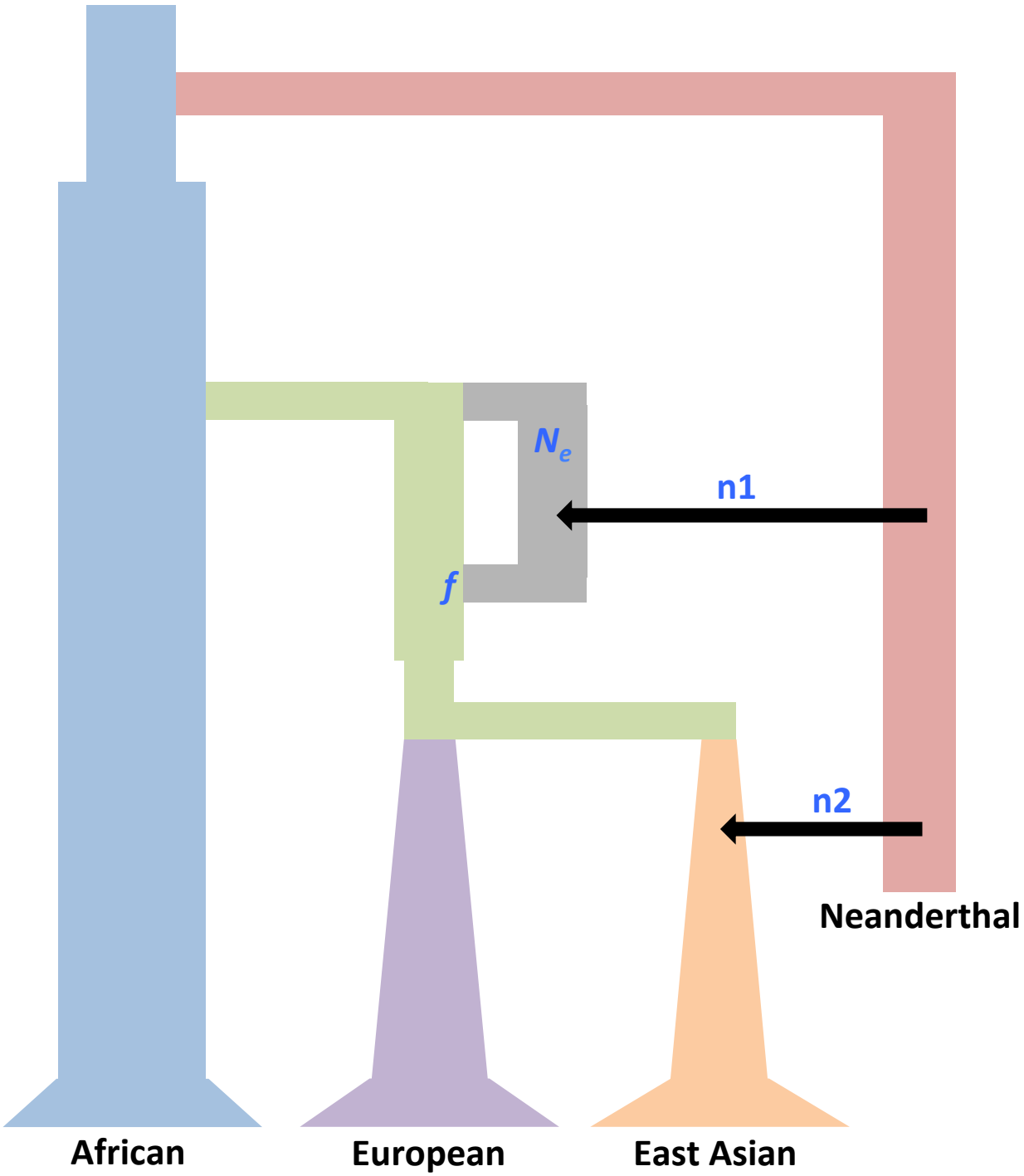


Figure 4.6. **Simplified schematic of the demographic model used for BOLFI simulation.** The model is based on the Split Population structure, with variable parameters for the admixture proportion (n1 and n2), effective population size of the Split Population (N_e), and the fraction of the Split Population that remerges (*f*).

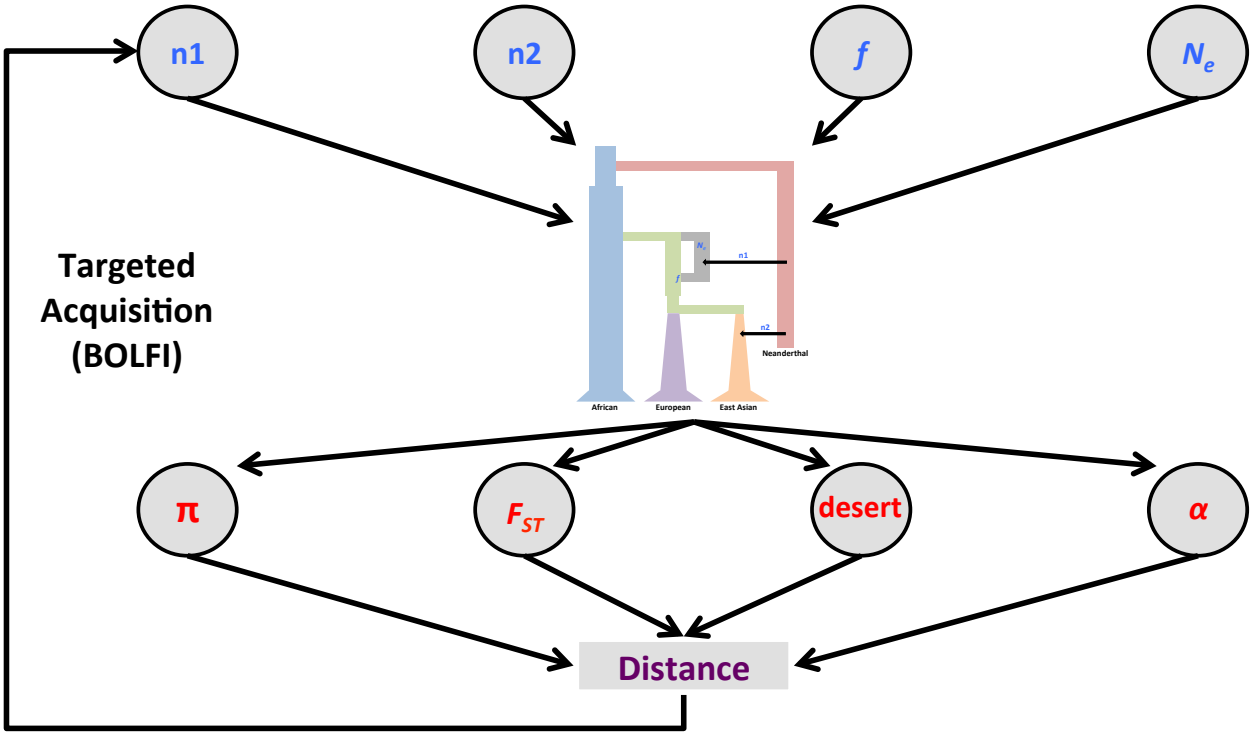


Figure 4.7. Simplified diagram of BOLFI optimization. Values for variable parameters (blue) are sampled from a prior distribution and used in the simulator to generate sequence data, which is then reported as summary statistics (red). Summary statistics are compared to the empirical data to calculate a distance measure. At later stages, BOLFI replaces the simulation step with its own model of the simulator, and updates its sampling of the parameters from the prior to minimize the distance measure in a process of targeted acquisition.

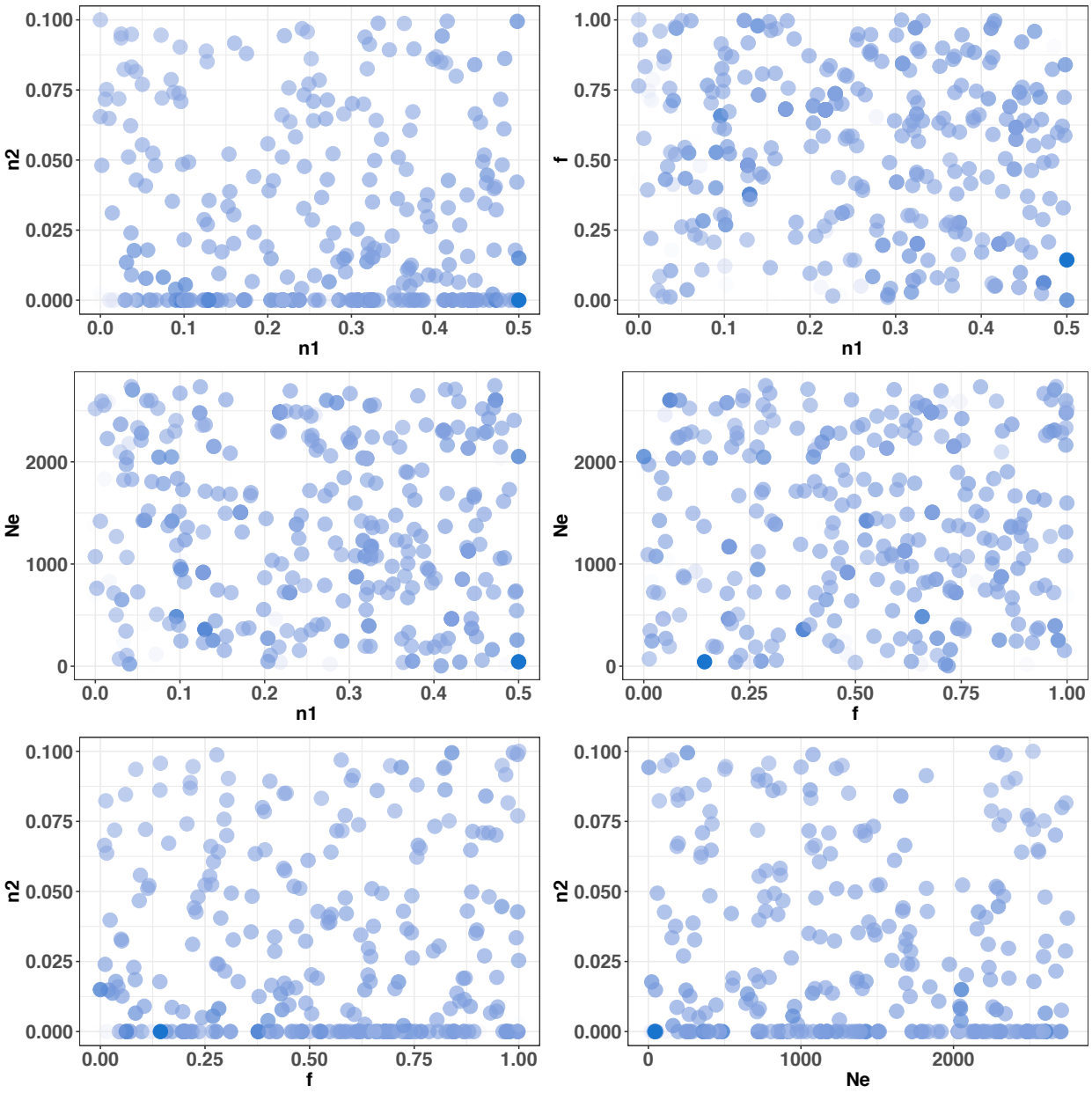


Figure 4.8. Scatter plots of prior parameters selected for BOLFI and ABCreg. Each dot represents a set of simulated model parameters, projected in reduced two-dimension space. Parameters for BOLFI and ABCreg were sampled from uniform distributions for n1 [0.0-0.5], n2 [0.0-0.1], *f* [0.0-1.0], and N_e [1-2758]. ABCreg was performed with priors from the uniform distributions, while BOLFI used the initial sampling to construct a model of the simulator and begin targeted acquisition.

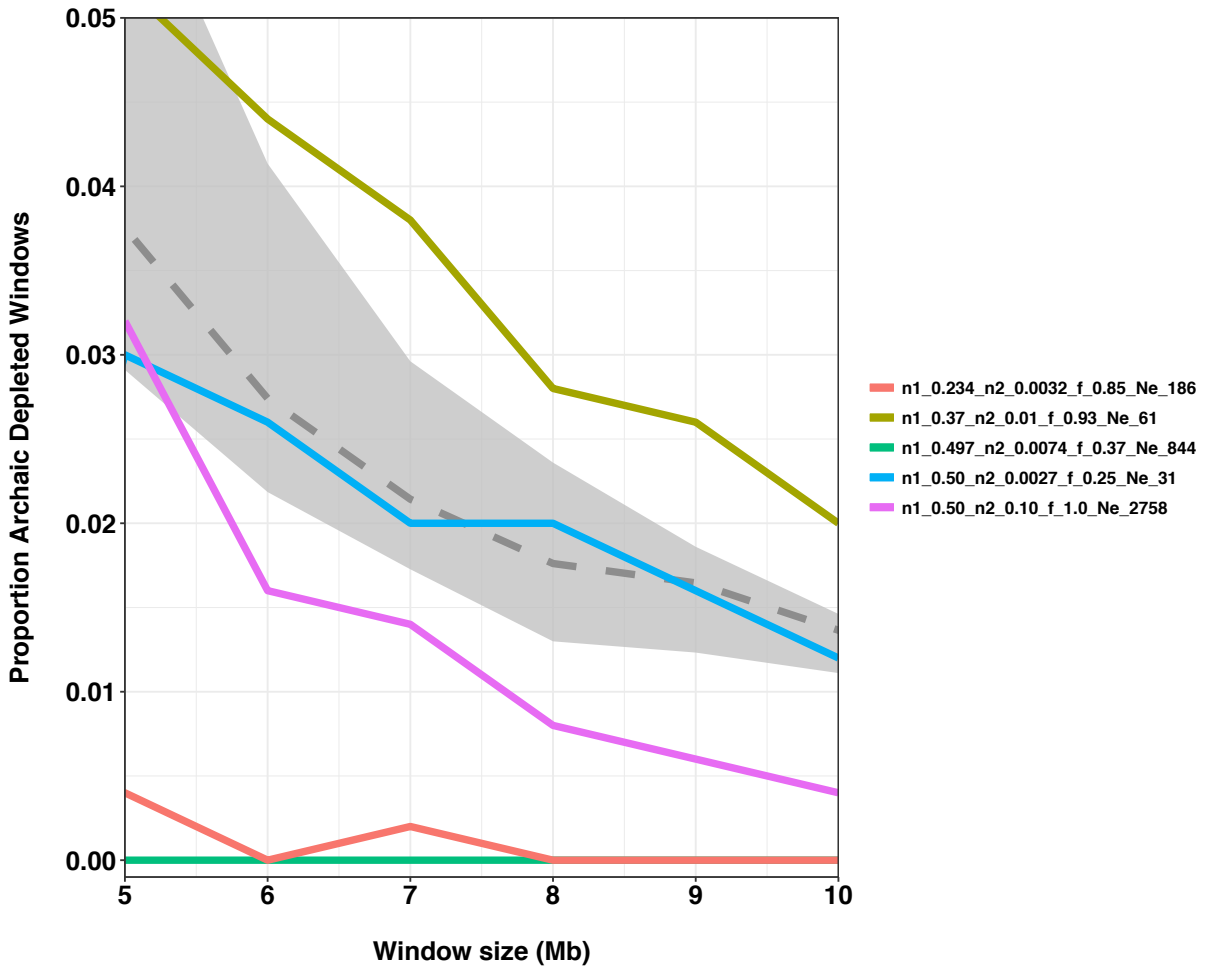


Figure 4.9. Distribution of desert proportions for low distance models from ABCreg and BOLFI. The desert distributions for models with low distance (Euclidean) to empirical data from BOLFI and ABCreg (**Table 4.2**) are displayed. The dashed line represents the empirical distribution, with bootstrapped 95% CI.

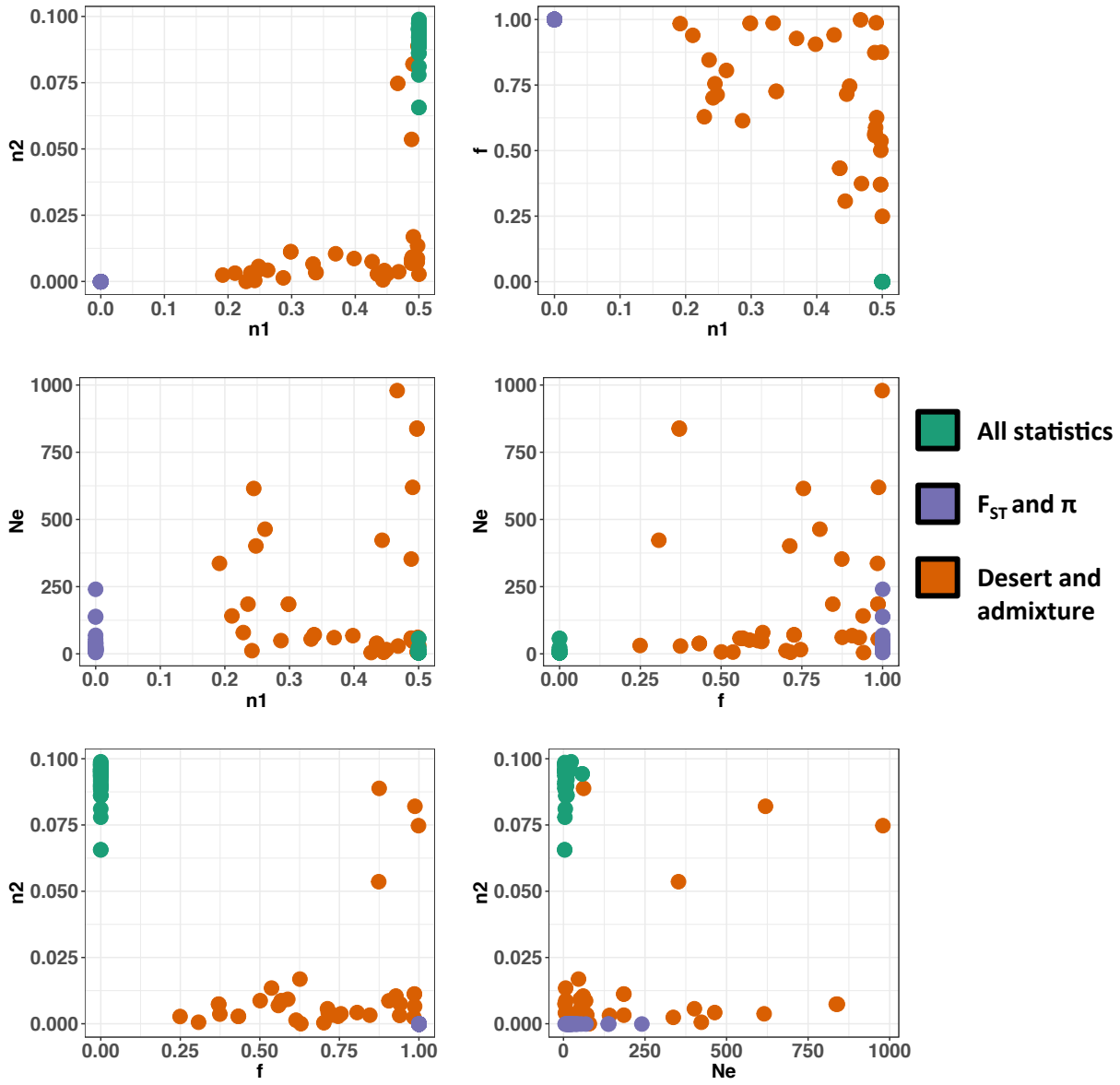


Figure 4.10. Scatter plots of posterior parameters estimated by ABCreg. Each dot corresponds to a single set of estimated posterior parameters, and is projected in a reduced two-dimension space. Posterior estimates in ABCreg used all the summary statistics (Green), just F_{ST} and π (Purple), or just the desert and admixture proportions (Orange).

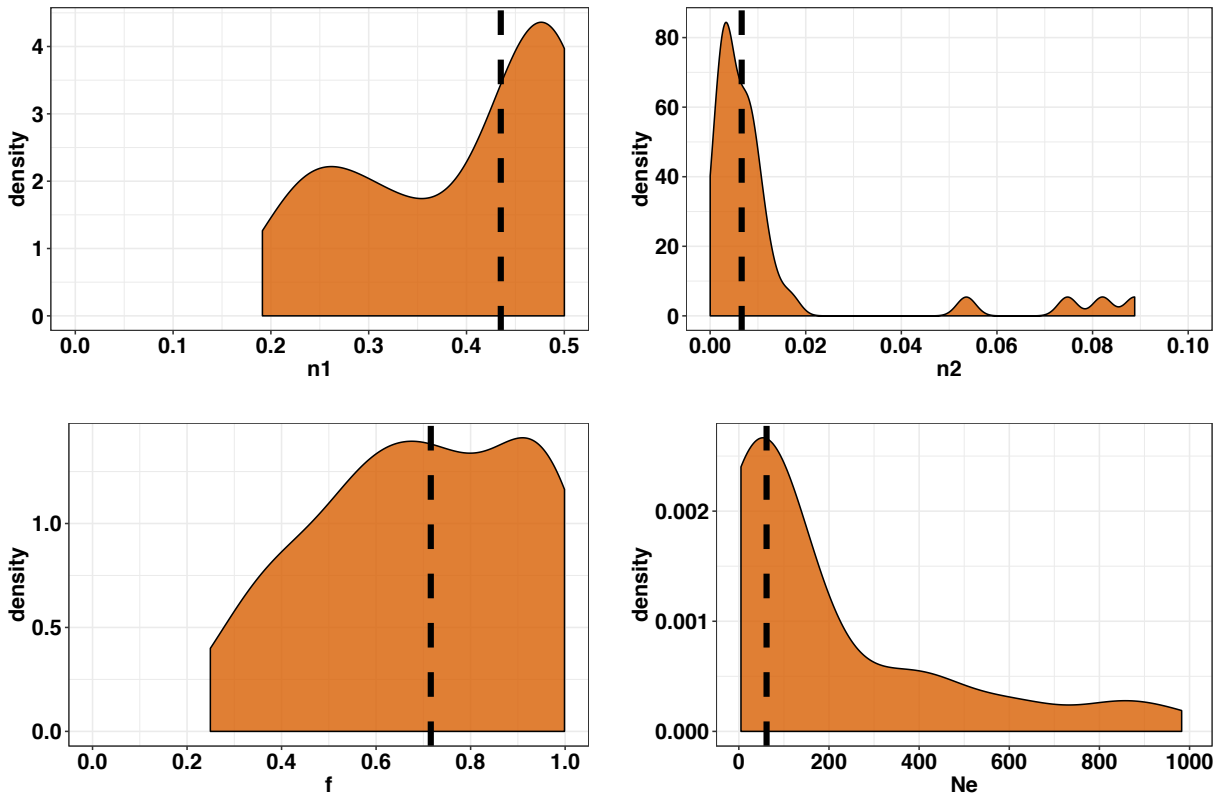


Figure 4.11. Density distributions of posterior parameters estimated by ABCreg based on desert and admixture proportions. Density distributions of the estimated posterior values for model parameters n1, n2, *f,* and N_e using ABCreg and desert and admixture proportions as summary statistics. The dashed line demarcates the median for each distribution.

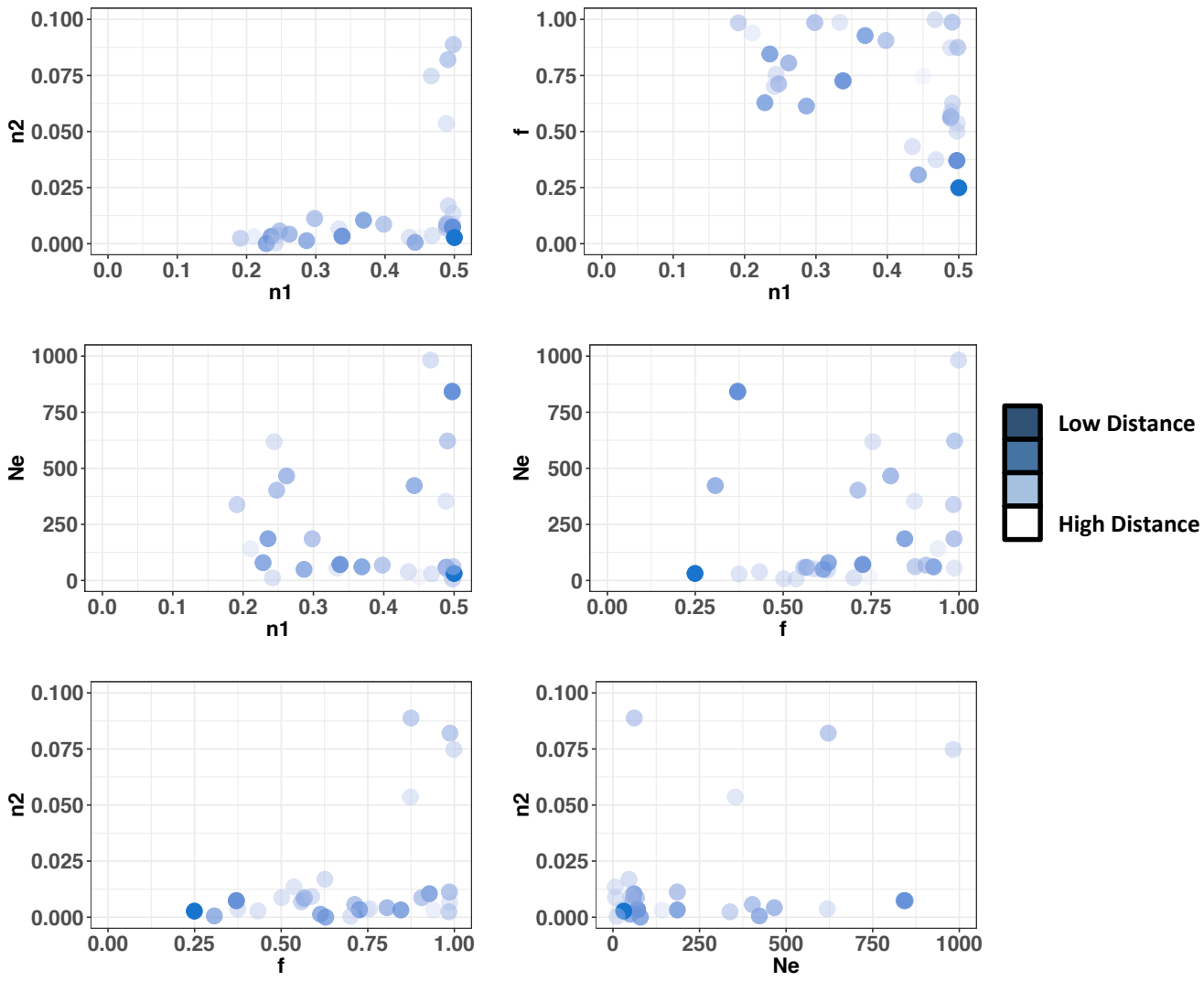


Figure 4.12. Heat map of distance measures for posterior parameters estimated by ABCreg from desert and admixture proportions. Each dot corresponds to a single set of estimated posterior parameters, and is projected in a reduced two-dimension space. Parameters were estimated using ABCreg and the desert and admixture proportions. Dots are colored based on the Euclidean distance calculated between their summary statistics and the empirical data (darker = low distance, lighter = high distance).

Table 4.1. Empirical values evaluated against in BOLFI and ABC.

5Mb desert rate	6Mb desert rate	7Mb desert rate	8Mb desert rate	9Mb desert rate	10Mb desert rate	pi AFR	pi EUR	pi EAS	Fst AFR-EUR	Fst AFR-EAS	Fst EUR-EAS	S* admix EAS	S* admix EUR
0.0375339	0.0274105	0.0214289	0.0176096	0.0164607	0.0136497	0.291761	0.328705	0.335145	0.12985	0.156539	0.092154	0.019	0.016

Table 4.2. Best Fitting Simulated Model Parameters from BOLFI and ABCreg.

n1	n2	*f*	N_e	European Neanderthal Ancestry (S*)	East Asian Neanderthal Ancestry (S*)	Distance (Euclidean)	Estimation Method
0.50	0.00273	0.25	31	0.0167787	0.016032	0.0093033	ABCreg
0.497	0.00739	0.37	844	0.0232726	0.022160	0.0140492	ABCreg
0.234	0.00325	0.85	186	0.0195113	0.018252	0.0147008	ABCreg
0.369	0.01047	0.93	61	0.0121707	0.013693	0.0150296	ABCreg
0.50	0.10000	1.0	2758	0.0082306	0.010583	0.1018369	BOLFI

Table 4.3. Posterior Values Estimated from All Summary Statistics with ABCreg.

n1	n2	*f*	N_e	Distance (Euclidean)
0.500089	0.0947013	0.0	60.4723	1.6664
0.500089	0.0947013	0.0	60.7513	1.6664
0.499742	0.0901596	0.0	5.69007	1.75842
0.499996	0.0675815	0.0	3.68374	1.77929
0.499996	0.0675815	0.0	3.6846	1.77929
0.50004	0.0939211	0.0	8.93365	1.83321
0.500073	0.0989382	0.0	24.7973	1.85501
0.499991	0.0986796	0.0	4.11175	1.86432
0.499952	0.0979907	0.0	4.58284	1.86555
0.499932	0.0934214	0.0	7.97809	1.86955
0.499778	0.0961712	0.0	21.9444	1.87298
0.500043	0.0954296	0.0	5.37579	1.87332
0.499239	0.0954899	0.0	10.3443	1.87416
0.500077	0.0891094	0.0	5.97186	1.87541
0.499727	0.0974629	0.0	6.78462	1.87691
0.500046	0.098051	0.0	4.21397	1.87742
0.500022	0.0963489	0.0	4.43447	1.88065
0.500022	0.0963489	0.0	4.43561	1.88065
0.500046	0.0941249	0.0	9.33635	1.89391
0.500052	0.0962551	0.0	11.4486	1.89539
0.50008	0.0920978	0.0	4.8411	1.89651
0.500078	0.0971764	0.0	4.84169	1.89708
0.499736	0.0870759	0.0	10.8597	1.89715
0.499736	0.0870759	0.0	10.8791	1.89715
0.500088	0.0956843	0.0	4.19655	1.89742
0.500087	0.0823055	0.0	5.69082	1.89872
0.500065	0.0906671	0.0	6.84522	1.89986
0.500092	0.0903888	0.0	9.81084	1.90285
0.500083	0.0869288	0.0	7.18619	1.90303
0.50009	0.0911607	0.0	4.02553	1.9031
0.500053	0.0895962	0.0	4.41929	1.90325
0.499833	0.079385	0.0	4.50424	1.90919
0.499722	0.0927259	0.0	11.1004	1.91065
0.500091	0.0940421	0.0	4.07258	1.91352
0.500089	0.0956329	0.0	4.11134	1.91886
0.499938	0.0953654	0.0	3.69552	1.92655
0.50002	0.097953	0.0	4.93008	1.93503

Table 4.4. Posterior Values Estimated from F_{ST} and π with ABCreg.

n1	n2	*f*	N_e	Distance (Euclidean)
0.0	0.0	1.0	22.2804	0.0298901
0.0	0.0	1.0	14.3777	0.0298912
0.0	0.0	1.0	43.3737	0.0298914
0.0	0.0	1.0	5.1756	0.0298917
0.0	0.0	1.0	16.1199	0.0298919
0.0	0.0	1.0	12.2055	0.0298924
0.0	0.0	1.0	12.217	0.0298924
0.0	0.0	1.0	14.7399	0.0298926
0.0	0.0	1.0	11.4952	0.0298932
0.0	0.0	1.0	138.844	0.0298933
0.0	0.0	1.0	30.3788	0.0298939
0.0	0.0	1.0	10.9788	0.0298943
0.0	0.0	1.0	25.5874	0.0298944
0.0	0.0	1.0	16.3042	0.0298946
0.0	0.0	1.0	16.3178	0.0298946
0.0	0.0	1.0	19.7817	0.0298946
0.0	0.0	1.0	22.7117	0.029895
0.0	0.0	1.0	7.874	0.029895
0.0	0.0	1.0	7.87808	0.029895
0.0	0.0	1.0	240.262	0.0298961
0.0	0.0	1.0	23.5756	0.0298963
0.0	0.0	1.0	17.065	0.0298965
0.0	0.0	1.0	54.2347	0.0298965
0.0	0.0	1.0	17.5635	0.0298975
0.0	0.0	1.0	25.6664	0.029899
0.0	0.0	1.0	68.8639	0.0299002
0.0	0.0	1.0	14.68	0.0299017
0.0	0.0	1.0	40.8479	0.0299023
0.0	0.0	1.0	14.759	0.0299062
0.0	0.0	1.0	20.6542	0.0299066
0.0	0.0	1.0	137.146	0.0299078
0.0	0.0	1.0	17.2385	0.0299085
0.0	0.0	1.0	6.65448	0.0299114
0.0	0.0	1.0	36.6283	0.0299136
0.0	0.0	1.0	14.8819	0.0299142
0.0	0.0	1.0	25.0713	0.0299181
0.0	0.0	1.0	37.5768	0.0299254

Table 4.5. Posterior Values Estimated from Desert and Admixture Proportions with ABCreg.

n1	n2	*f*	N_e	Distance (Euclidean)
0.500052	0.00272596	0.249281	31.2041	0.00930325
0.497342	0.00739504	0.370596	840.339	0.0140492
0.497342	0.00739504	0.370596	843.828	0.0140492
0.235429	0.00324614	0.845567	185.883	0.0147008
0.368885	0.0104708	0.927781	60.7113	0.0150296
0.228416	1.54861e-05	0.628488	79.1743	0.0152975
0.337855	0.0033609	0.726031	71.0351	0.0153551
0.337855	0.0033609	0.726031	71.21	0.0153551
0.286647	0.00134805	0.61357	49.2362	0.016166
0.44335	0.000569761	0.306984	422.936	0.0165929
0.261789	0.00424016	0.805684	465.935	0.0183691
0.247739	0.00567103	0.712441	402.427	0.0206612
0.488885	0.00871582	0.567068	58.7208	0.0210037
0.398	0.00866455	0.905744	68.1134	0.0220958
0.490691	0.082059	0.987433	621.681	0.0225521
0.498567	0.0888093	0.874889	61.5429	0.0241656
0.488484	0.00689339	0.558583	58.3335	0.0254769
0.191305	0.0024085	0.984189	337.896	0.0261843
0.298275	0.011232	0.985766	185.611	0.028131
0.298275	0.011232	0.985766	186.078	0.028131
0.488796	0.00801054	0.56527	58.0605	0.0291753
0.491257	0.0168886	0.626209	46.3842	0.0309209
0.489748	0.00925148	0.588005	51.4753	0.0315642
0.466662	0.0747555	0.998663	982.466	0.0321297
0.241726	0.0003778	0.700806	11.794	0.03325
0.244068	0.00375337	0.754671	618.613	0.0344948
0.497663	0.00873208	0.501399	7.21819	0.0346094
0.468174	0.0036471	0.374495	29.1235	0.0351895
0.497977	0.0134858	0.537074	6.74761	0.035972
0.488517	0.0535965	0.87348	353.193	0.0365879
0.333394	0.00660095	0.986492	55.2104	0.0381184
0.210833	0.00314704	0.939631	141.51	0.0458903
0.43504	0.00279141	0.432825	38.6993	0.0460419
0.43504	0.00279141	0.432825	38.9951	0.0460419
0.450125	0.00282167	0.746131	15.7356	0.0598967
0.426498	0.00750256	0.941274	4.74308	0.139661
0.445749	0.00410578	0.715741	5.86883	0.154642

Chapter 5. Conclusion

Altogether, this work elucidates some of the complexities of archaic and modern human admixture and answers several of the outstanding questions in this field. I have discussed results from multiple different methods for detecting archaic introgressed sequence in the modern human genome, which all demonstrate the universal presence of archaic sequence across human populations. Furthermore, while archaic sequence is present in every human population examined to date, the distribution of this sequence across the genome is heterogeneous, with several large regions significantly depleted of it. Here, I return to the questions outlined in my Introduction and summarize my main findings as they relate to them. I also suggest future avenues of exploration.

5.1 Initial Levels of Archaic-Modern Human Admixture

Results from analyses with S* and IBDmix are consistent in estimating Neanderthal ancestry across global non-African populations as ~2%, and among African populations as <1%. Less conservative filtering approaches with IBDmix do increase the amount of Neanderthal ancestry detected in both African and non-African populations. However, this increase is only on the order of less than 1%. Recent reports using the high quality Vindija Neanderthal reference genome, which is purported to be more closely related to the admixing Neanderthal population than the Altai reference sample, also discover marginally higher levels of Neanderthal ancestry in modern populations, again only on the order of an additional 1% [51]. Alternatively, simulations modeling admixture history conditional on the formation of large deserts estimate initial admixture levels close to 50%, exceptionally higher than the 2-3% initially estimated [34], or found in any ancient human samples to date [107,108].

It appears that the discovery of substantially more Neanderthal ancestry in any modern population is unlikely. Instead, future research in modern populations can turn to two alternative paths: 1) To improve methods for detecting Neanderthal ancestry tracks to provide the most precise catalogues of these loci at the individual and population level, or 2) Improve methods for detecting archaic sequence generally in samples, without the use of an archaic reference genome, in order to identify new archaic species through their genetic legacy. IBDmix is well designed to

approach the first of these options. By not relying on a "non-admixed" reference population, IBDmix is powered to accurately detect archaic sequence in any population, regardless of its admixture history, so long as sample sizes are appropriate and an archaic reference genome is available. Continued analyses with IBDmix can provide novel insights about population admixture histories and population specific adaptive introgression. We do find IBDmix is sensitive to heterogeneous local recombination-rates, and so future modifications to IBDmix should focus on integrating population-specific recombination-rate maps to further improve performance in terms of FDR and power.

Alternatively, S* offers one of the better approaches for identifying archaic sequence in samples without the use of an archaic reference genome, as in [124]. However, the performance of S* in this capacity is still highly sensitive to the admixture history of the "un-admixed" reference population it uses. In addition, without the secondary step of comparing putative archaic sequence to a reference genome, S*'s FDR is difficult to control [33]. One could leverage IBDmix results when picking a non-admixed reference population for use in S*, to find the population or collection of samples with the lowest rates of admixture, but this detection would still be limited to archaic sequences for which a reference genome is available.

Ultimately, answering questions regarding the initial admixture level will depend on continued collection and analysis of ancient human and Neanderthal samples closer to the time of admixture. In this situation, where sample sizes are very limited for each population, IBDmix will be underpowered due to its reliance on accurate allele frequency estimates. However, IBDmix could be used as a complementary tool in these cases. Our results demonstrate the complex admixture histories among human and archaic populations. As recent analyses have highlighted the sensitivity of certain admixture statistics to assumptions about demographic histories [38], a fine-scale map of archaic ancestry across continents could be used to detail patterns of dispersal and admixture among populations. This knowledge could improve the application of other admixture statistics to ancient samples.

5.2 Frequency of Admixture with Neanderthals

Previous analyses found that levels of Neanderthal ancestry varied among populations [29–31,33,34]. Analyses of introgressed Neanderthal sequence using the high-quality Altai

reference genome [23] noted more regions of Neanderthal origin in Asian and American populations than European ones [32], as well as higher levels of Neanderthal ancestry in East Asian populations compared to European populations [30,31]. The differences in these Neanderthal ancestry proportions manifest as an ~20% enrichment of Neanderthal ancestry in East Asians compared to Europeans. Analyses using statistical and simulation approaches have indicated that models incorporating multiple pulses of admixture explain these patterns better than a single pulse of admixture [30,33,40–42].

However, our analyses of these same samples and populations using IBDmix show near uniform levels of Neanderthal ancestry across non-African populations. While we still observe some enrichment of Neanderthal ancestry in East Asians compared to Europeans, this enrichment is less than 10%. This modest enrichment can be most parsimoniously explained by a single wave of Neanderthal admixture occurring shortly after the Out-of-Africa dispersal. Variation in Neanderthal ancestry could be attributed to dilution afterward [39]. In particular, present day European populations are thought to be a mixture of three ancestral groups, one of which had ancestry from a Basal Eurasian lineage that had little or no Neanderthal ancestry [106]. Previous studies found that dilution could not explain Neanderthal ancestry differences as large as 20% [41,87], but can readily account for the modest differences we find.

While the total recovered amount of Neanderthal sequence can be informative of the number of admixture pulses, strong negative selection is believed to have affected introgressed Neanderthal sequence [35,38] and may have repeatedly driven archaic ancestry to a uniform level. A more informative mode of analysis would be to investigate the sequence diversity of recovered Neanderthal segments across populations. In this regard, IBDmix could be well powered to provide the most accurate catalogue of Neanderthal introgressed loci at the individual and population level. Furthermore, we show IBDmix to be robust in its performance, regardless of the genetic distance of the archaic reference genome to the actual introgressing archaic population. Maximizing the detection of Neanderthal ancestry would provide the necessary starting point for comparing divergence between these haplotypes within and between populations. We could then discern whether the Neanderthal sequence across populations are derived from a single archaic population or several.

5.3 Archaic Admixture in Africa

Studies of archaic admixture in African populations have been limited, despite the fact that numerous archaic hominin lineages are known to have existed in Africa [44], and may have overlapped in time and space with modern humans [45]. Several studies have investigated the likelihood of archaic admixture in African populations [46–48], and have been necessarily cautious in their conclusion that admixture occurred between an unidentified archaic hominin and several African populations.

Part of the challenge for identifying introgressed sequence in African populations has been the issue of controlling for ILS. While S* may be capable of detecting introgressed sequence from an unknown archaic based on strong LD between highly divergent variants, the reliance on a non-admixed reference population to control for ILS can introduce serious biases. This would be especially true if the admixture happened in Africa before the split of African and non-African populations, and is therefore present in both lineages. IBDmix is also poorly suited to this application, as its reliance on an archaic reference genome means it cannot be applied in situations where the admixing archaic hominin is unknown. The recovery of an African archaic hominin genome is complicated by the combined effects of the greater age of archaic samples in that region and the environmental impediments to recovering ancient DNA.

Importantly, our analyses with IBDmix have shown that the presence of Neanderthal sequence in African samples is likely not due to additional and independent admixture events. The detected Neanderthal sequence in Africans is almost entirely a subset of the sequence present in non-Africans, and shows many characteristics to suggest it is derived from back-migration by already admixed populations. Our additional finding that some of this sequence is a result of older human to Neanderthal gene-flow indicates that barriers to admixture between modern and archaic hominins were not as high as previously thought, and that admixture has been a recurrent feature of our history. As such, it would be surprising if similar instances of admixture did not occur among archaic and modern humans in Africa. Future progress will certainly depend on the recovery and analysis of additional archaic samples, and additional ancient human samples dated closer to the time of admixture.

5.4 Gene-flow from Modern Humans into Neanderthals

Several studies have confidently inferred signals of human to Neanderthal gene-flow at a low level. The timing of this event remains unclear, however, with some estimates suggesting 100kya [50], others closer to 150kya [51], and some potentially as early as 300kya [52]. Our own results from analyses with IBDmix confirm a signal for pre-OOA gene-flow from humans to Neanderthals. Based on simulation studies, we determine the origin of our signal is an admixture event approximately 100kya. IBDmix does not detect admixture events older than 100kya in these simulations. However, this may be a function of the stringent filtering threshold that we apply to IBDmix calls. We expect older admixture events to leave short segments, and sequences from these events may not exceed our cutoff threshold. Distinguishing these signals from ILS and false-positives could be especially difficult.

One possible approach to improve estimates of human introgressed sequence in Neanderthals would be to apply a reference-free method like IBDmix, using ancient human genomes as the "archaic" and Neanderthal genomes as the "target". Though IBDmix performs best when target sample sizes are large and therefore provide better estimates of allele frequencies, the continued collection and sequencing of additional Neanderthal samples [66] could make applying IBDmix to Neanderthal genomes possible. However, we would need to explore first how the historically small effective population size and isolated population structure of Neanderthal lineages [125] complicate allele frequency estimates in these populations. While we are unlikely to ever collect a sufficient number of archaic samples to fully replicate studies of adaptive introgression or significant depletions across Neanderthal populations, documenting the distribution of human sequence across the Neanderthal genome, even roughly, would be informative about the selective forces shaping their evolution, especially where it differs from modern humans.

5.5 Deserts of Archaic Sequence

On the autosomes, the largest deserts span multiple megabases, with a handful extending up to 10Mb in length [33,34]. The mechanisms responsible for the heterogeneous distribution of Neanderthal sequence across the autosomes are not yet fully understood, and several may act in

combination. These mechanisms may include neutral demographic processes, selection against introgressed sequence at specific loci, or genomic features like structural variation that reduce or prevent recombination.

In analyses using S* [30], I examined the likelihood of large depletions of archaic sequence forming under simple neutral models. The results of these simulations suggested neutral processes alone were unable to generate deserts at a sufficient size and frequency to match the empirical data.

During later analyses (chapter 4) I explored more complex model structures that include intense bottlenecks during or after the time of admixture. It initially seemed unlikely that neutral processes could replicate the empirical data across multiple features, such as the distribution of deserts and admixture proportion. However, comprehensive searches of the parameter space with ABC and BOLFI identified certain sets of parameters that match the empirical data when conditioning on features like genetic diversity, admixture proportion, and desert distributions. The parameters these methods estimate are extreme, especially the necessary effective population size and initial admixture rates. However, these models represent simplifications of complex demographic processes, and so estimated parameters should not be interpreted as historical fact.

Interestingly, the extremely small effective population size estimated is analogous to models of strong selection against introgressed sequence. Several studies have indicated the rapid loss of Neanderthal ancestry overall was due to the accumulation of weakly deleterious alleles in the Neanderthal genome [35,38]. This suggests that modeling the loss of Neanderthal sequence generally as a neutral process may be inaccurate, and instead any investigation of deserts should model weak selection. Using forward-in-time simulators to generate these demographic models and include selection parameters would be an important next step. However, what exactly this selection should look like at these loci—weak selection across many variants or strong selection at a few loci—remains unclear. Additionally, environmental differences between modern humans and Neanderthals mean that the strength and direction of selection on variants may have changed when archaic alleles entered the human population. What was advantageous or neutral in a Neanderthal population could have been deleterious in a human one.

It is also important to consider the role for structural variation, like large inversions, to prevent introgression at these loci by suppressing recombination. These inversions could be present on either the human or archaic lineages. Considering the overlap of Neanderthal and Denisovan deserts [30], large inversions seem unlikely to explain all of the archaic depletions found to date, but remain a formal possibility. Unfortunately, identifying potential lineage-specific inversions is incredibly difficult given short sequencing read lengths for ancient DNA.

In addition to modeling the etiology of desert loci, they should also be examined at the functional level. We found these regions exhibit higher levels of background selection, human-Neanderthal sequence divergence, and are enriched for genes expressed in regions of the brain [30,33]. Despite these findings, detailed testing of the functional differences between Neanderthal and humans at these loci has yet to be carried out. The continued development of gene synthesis technologies could eventually make it financially feasible to perform extensive comparison of the human and Neanderthal sequences at these loci in massively-parallel reporter assays like STARR-Seq [126]. Developments in the field of organoids could even allow for testing the effect of Neanderthal variants in complex tissue structures. In short, the phenotypic consequences of human or Neanderthal sequence at these loci have not been extensively explored. If desert regions are depleted due to selection, the human sequence at these loci could elucidate the genetic origin of exceptional human phenotypes.

5.6 Concluding Remarks

The discovery that all non-African populations carry ~2% Neanderthal ancestry was a significant breakthrough in anthropology and paleogenomics. The research detailed above examines some of the complexities of archaic and modern human admixture and answers several of the outstanding questions in the field. Specifically, we've demonstrated the universal presence of archaic sequence across human populations. The legacy of gene flow with Neanderthals likely exists in all modern humans, highlighting our shared history.

Future progress will certainly depend on the recovery and analysis of additional archaic samples—Neanderthals, Denisovans, and others—and additional ancient human samples dated closer to the time of admixture. As well, analyses using improved methods to detect introgressed archaic sequence in existing data remain critical. The research described above highlights how

complex the admixture history was between human and Neanderthal populations, and suggests barriers between these populations were much lower than historically thought.

REFERENCES

1. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, et al. (2003) Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature 423: 742–747.
2. McDougall I, Brown FH, Fleagle JG (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. Nature 433: 733–736.
3. Hublin J-J, Ben-Ncer A, Bailey SE, Freidline SE, Neubauer S, Skinner MM, et al. (2017) New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. Nature 546: 289–292.
4. Richter D, Grün R, Joannes-Boyau R, Steele TE, Amani F, Rué M, et al. (2017) The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. Nature 546: 293–296.
5. Finlayson C, Giles Pacheco F, Rodríguez-Vidal J, Fa DA, María Gutierrez López J, Santiago Pérez A, et al. (2006) Late survival of Neanderthals at the southernmost extreme of Europe. Nature 443: 850–853.
6. Higham T, Douka K, Wood R, Ramsey CB, Brock F, Basell L, et al. (2014) The timing and spatiotemporal patterning of Neanderthal disappearance. Nature 512: 306–309.
7. Brown P, Sutikna T, Morwood MJ, Soejono RP, Jatmiko, Wayhu Saptomo E, et al. (2004) A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. Nature 431: 1055–1061.
8. Morwood MJ, Brown P, Jatmiko, Sutikna T, Wahyu Saptomo E, Westaway KE, et al. (2005) Further evidence for small-bodied hominins from the Late Pleistocene of Flores, Indonesia. Nature 437: 1012–1017.
9. Hublin JJ (2009) The origin of Neandertals. Proc Natl Acad Sci 106: 16022–16027.
10. Langley MC, Clarkson C, Ulm S (2008) Behavioural Complexity in Eurasian Neanderthal Populations: a Chronological Examination of the Archaeological Evidence. Cambridge Archaeol J 18: 289–307.
11. Hayden B (1993) The cultural capacities of Neandertals: a review and re-evaluation. J Hum Evol 24: 113–146.
12. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053–1060.

13. Meyer M, Kircher M, Gansauge M, Li H, Racimo F, Mallick S, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222–226.
14. Slon V, Hopfe C, Weiß CL, Mafessoni F, de la Rasilla M, Lalueza-Fox C, et al. (2017) Neandertal and Denisovan DNA from Pleistocene sediments. Science 356: 605–608.
15. Trinkaus E (2007) European early modern humans and the fate of the Neandertals. Proc Natl Acad Sci USA 104: 7367–7372.
16. Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M, Paabo S (1997) Neandertal DNA Sequences and the Origin of Modern Humans. Cell 90: 19–30.
17. Krings M, Geisert H, Schmitz RW, Krainitzki H, Pääbo S (1999) DNA sequence of the mitochondrial hypervariable region II from the neandertal type specimen. Proc Natl Acad Sci U S A 96: 5581–5585.
18. Ovchinnikov I V., Gotherstrom A, Romanova GP, Kharitonov VM, Liden K, Goodwin W (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature 404: 490–493.
19. Green RE, Malaspinas A-S, Krause J, Briggs AW, Johnson PLF, Uhler C, et al. (2008) A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. Cell 134: 416–426.
20. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, et al. (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113–1118.
21. Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, et al. (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444: 330–336.
22. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. (2010) A draft sequence of the Neandertal genome. Science 328: 710–722.
23. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43–49.
24. Kelso J, Prüfer K (2014) Ancient humans and the origin of modern humans. Curr Opin Genet Dev 29: 133–138.
25. Pääbo S (2014) The human condition-a molecular approach. Cell 157: 216–226.
26. Vattathil S, Akey JM (2015) Small Amounts of Archaic Admixture Provide Big Insights into Human History. Cell 163: 281–284.

27. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541: 302–310.
28. Wall JD, Brandt DYC (2016) Archaic admixture in human history. Curr Opin Genet Dev 41: 93–97.
29. Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538: 201–206.
30. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. (2016) Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science 352: 235–239.
31. Sankararaman S, Mallick S, Patterson N, Reich D (2016) The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. Curr Biol 26: 1241–1247.
32. Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, et al. (2013) Higher levels of neanderthal ancestry in East Asians than in Europeans. Genetics 194: 199–209.
33. Vernot B, Akey JM (2014) Resurrecting Surviving Neandertal Linages from Modern Human Genomes. Science 343: 1017–1021.
34. Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, et al. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507: 354–357.
35. Harris K, Nielsen R (2016) The genetic cost of neanderthal introgression. Genetics 203: 881–891.
36. Juric I, Aeschbacher S, Coop G (2016) The Strength of Selection against Neanderthal Introgression. PLoS Genet 12: e1006340.
37. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, et al. (2016) The genetic history of Ice Age Europe. Nature 534: 200–205.
38. Petr M, Pääbo S, Kelso J, Vernot B (2019) Limits of long-term selection against Neandertal introgression. Proc Natl Acad Sci 116: 1639–1644.
39. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. (2016) Genomic insights into the origin of farming in the ancient Near East. Nature 536: 419–424.
40. Kim BY, Lohmueller KE (2015) Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations.

Am J Hum Genet 96: 454–61.

41. Vernot B, Akey JM (2015) Complex History of Admixture between Modern Humans and Neandertals. Am J Hum Genet 96: 448–453.
42. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, et al. (2016) A genomic history of Aboriginal Australian. Nature 538: 207–214.
43. Kim BY, Lohmueller KE (2015) Selection and Reduced Population Size Cannot Explain Higher Amounts of Neandertal Ancestry in East Asian than in European Human Populations. Am J Hum Genet 96: 454–461.
44. Brauer G (2008) The Origin of Modern Anatomy: By Speciation or Intraspecific Evolution? Evol Anthropol 17: 22–37.
45. Dirks PH, Roberts EM, Hilbert-Wolf H, Kramers JD, Hawks J, Dosseto A, et al. (2017) The age of Homo naledi and associated sediments in the Rising Star Cave, South Africa. Elife 6: e24231.
46. Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo J-M, et al. (2012) Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. Cell 150: 457–469.
47. Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, et al. (2016) Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Res 26: 291–300.
48. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD (2011) Genetic evidence for archaic admixture in Africa. Proc Natl Acad Sci U S A 108: 15123–15128.
49. Xu D, Pavlidis P, Taskent RO, Alachiotis N, Flanagan C, DeGiorgio M, et al. (2017) Archaic Hominin Introgression in Africa Contributes to Functional Salivary MUC7 Genetic Variation. Mol Biol Evol 34: 2704–2715.
50. Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, et al. (2016) Ancient gene flow from early modern humans into Eastern Neanderthals. Nature 530: 429–433.
51. Prüfer K, Filippo C de, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. (2017) A high-coverage Neandertal genome from Vindija Cave in Croatia. Science 358: 655–658.
52. Posth C, Wißing C, Kitagawa K, Pagani L, van Holstein L, Racimo F, et al. (2017) Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene

flow into Neanderthals. Nat Commun 8: 16046.

53. Posth C, Wißing C, Kitagawa K, Pagani L, Holstein L van, Racimo F, et al. (2017) Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. Nat Commun 8: 16046.
54. Enard W, Przeworski M, Fisher SE, Lai CSL, Wiebe V, Kitano T, et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418: 869–872.
55. Dannemann M, Kelso J (2017) The Contribution of Neanderthals to Phenotypic Variation in Modern Humans. Am J Hum Genet 101: 578–589.
56. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, et al. (2016) The phenotypic legacy of admixture between modern humans and Neandertals. Science 351: 737–741.
57. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM, et al. (2016) Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Curr Biol 26: 3375–3382.
58. Mendez FL, Watkins JC, Hammer MF (2012) A Haplotype at STAT2 Introgressed from Neanderthals and Serves as a Candidate of Positive Selection in Papua New Guinea. Am J Hum Genet 91: 265–274.
59. Mendez FL, Watkins JC, Hammer MF (2012) Global genetic variation at OAS1 provides evidence of archaic admixture in Melanesian populations. Mol Biol Evol 29: 1513–1520.
60. Dannemann M, Andrés AM, Kelso J (2016) Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. Am J Hum Genet 98: 22–33.
61. Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, et al. (2016) Archaic adaptive introgression in TBX15/WARS2. Mol Biol Evol 34: 509–524.
62. Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature 512: 194–197.
63. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. Nat Rev Genet 16: 359–371.
64. Dannemann M, Prüfer K, Kelso J (2017) Functional implications of Neandertal introgression in modern humans. Genome Biol 18: 61.

65. McCoy RC, Wakefield J, Akey JM (2017) Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell 168: 916–927.
66. Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, Grote S, et al. (2018) Reconstructing the genetic history of late Neanderthals. Nature .
67. Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, et al. (2018) The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature 561: 113–116.
68. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. (2015) A global reference for human genetic variation. Nature 526: 68–74.
69. Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. PLoS Genet 2: e105.
70. Maricic T, Günther V, Georgiev O, Gehre S, Ćurlin M, Schreiweis C, et al. (2013) A Recent Evolutionary Change Affects a Regulatory Element in the Human FOXP2 Gene. Mol Biol Evol 30: 844–852.
71. Borrell V, Cárdenas A, Ciceri G, Galcerán J, Flames N, Pla R, et al. (2012) Slit/Robo signaling modulates the proliferation of central nervous system progenitors. Neuron 76: 338–52.
72. Evans TA, Santiago C, Arbeille E, Bashaw GJ (2015) Robo2 acts in trans to inhibit Slit-Robo1 repulsion in pre-crossing commissural axons. Elife 4: e08407.
73. Kidd T, Bland KS, Goodman CS (1999) Slit Is the Midline Repellent for the Robo Receptor in Drosophila. Cell 96: 785–794.
74. Bates TC, Luciano M, Medland SE, Montgomery GW, Wright MJ, Martin NG (2011) Genetic Variance in a Component of the Language Acquisition Device: ROBO1 Polymorphisms Associated with Phonological Buffer Deficits. Behav Genet 41: 50–57.
75. Hannula-Jouppi K, Kaminen-Ahola N, Taipale M, Eklund R, Nopola-Hemmi J, Kääriäinen H, et al. (2005) The Axon Guidance Receptor Gene ROBO1 Is a Candidate Gene for Developmental Dyslexia. PLoS Genet 1: e50.
76. Wang S, Lachance J, Tishkoff SA, Hey J, Xing J (2013) Apparent Variation in Neanderthal Admixture among African Populations is Consistent with Gene Flow from Non-African Populations. Genome Biol Evol 5: 2075–2081.
77. Chen GK, Marjoram P, Wall JD (2008) Fast and flexible simulation of DNA sequence data. Genome Res 19: 136–142.

78. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. (2012) Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science (80-) 337: 64–69.
79. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, et al. (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108: 11983–11988.
80. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5.
81. Schaffner S, Foo C, Gabriel S (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome … : 1576–1583.
82. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. (2016) Suppl: Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science .
83. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E (2017) Tracing the peopling of the world through genomics. Nature 541: 302–310.
84. Vernot B, Pääbo S (2018) The Predecessors Within . . . Cell 173: 6–7.
85. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM (2018) Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell 173: 1–9.
86. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. (2016) The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 538: 201–206.
87. Kim BY, Lohmueller KE (2015) Selection and Reduced Population Size Cannot Explain Higher Amounts of Neandertal Ancestry in East Asian than in European Human Populations. Am J Hum Genet 96: 454–461.
88. Villanea FA, Schraiber JG (2019) Multiple episodes of interbreeding between Neanderthal and modern humans. Nat Ecol Evol 3: 39–44.
89. Steinrücken M, Spence JP, Kamm JA, Wieczorek E, Song YS (2018) Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. Mol Ecol 27: 3873–3888.
90. Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, et al. (2018) Detecting

archaic introgression using an unadmixed outgroup. Racimo F, editor. PLOS Genet 14: e1007641.

91. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493: 216–220.
92. Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, et al. (2016) Ancient gene flow from early modern humans into Eastern Neanderthals. Nature 530: 429–433.
93. Maples BK, Gravel S, Kenny EE, Bustamante CD (2013) RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am J Hum Genet 93: 278–288.
94. Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM (2016) Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments. Curr Biol 26: 3375–3382.
95. Keeney S, Chang GJ, Linn S (1993) Characterization of a human DNA damage binding protein implicated in xeroderma pigmentosum E. J Biol Chem 268: 21293–300.
96. Kim Y, Lee J, Kim J, Choi CW, Hwang Y-I, Kang JS, et al. (2017) The pathogenic role of interleukin-22 and its receptor during UVB-induced skin inflammation. Blumenberg M, editor. PLoS One 12: e0178567.
97. Lorente-Galdos B, Lao O, Serra-Vidal G, Santpere G, K Kuderna LF, Arauna LR, et al. Whole-genome sequence analysis of a Pan African set of samples reveals archaic gene flow from an extinct basal population of modern humans into sub-Saharan populations.
98. Wang S, Lachance J, Tishkoff SA, Hey J, Xing J (2013) Apparent Variation in Neanderthal Admixture among African Populations is Consistent with Gene Flow from Non-African Populations. Genome Biol Evol 5: 2075–2081.
99. Sánchez-Quinto F, Botigué LR, Civit S, Arenas C, Ávila-Arcos MC, Bustamante CD, et al. (2012) North African Populations Carry the Signature of Admixture with Neandertals. Caramelli D, editor. PLoS One 7: e47765.
100. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. (2012) The genetic prehistory of southern Africa. Nat Commun 3: 1143.
101. Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. (2017)

Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science (80-) : eaao6266.

102. Scerri EML, Thomas MG, Manica A, Gunz P, Stock JT, Stringer C, et al. (2018) Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? Trends Ecol Evol 33: 582–594.
103. Schlebusch CM, Skoglund P, Sjodin P, Gattepaille LM, Hernandez D, Jay F, et al. (2012) Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and Complex African History. Science (80-) 338: 374–379.
104. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, et al. (2017) Reconstructing Prehistoric African Population Structure. Cell 171: 59–71.e21.
105. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. (2014) Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A 111: 2632–7.
106. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513: 409–413.
107. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, et al. (2015) An early modern human from Romania with a recent Neanderthal ancestor. Nature 524: 216–219.
108. Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, et al. (2017) 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. Curr Biol 27: 3202–3208.e9.
109. Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, et al. (2015) Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. Science 350: 820–2.
110. Henn BM, Botigué LR, Gravel S, Wang W, Brisbin A, Byrnes JK, et al. (2012) Genomic Ancestry of North Africans Supports Back-to-Africa Migrations. Schierup MH, editor. PLoS Genet 8: e1002397.
111. Hervella M, Svensson EM, Alberdi A, Günther T, Izagirre N, Munters AR, et al. (2016) The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa. Sci Rep 6: 25501.

112. Hodgson JA, Mulligan CJ, Al-Meeri A, Raaum RL (2014) Early Back-to-Africa Migration into the Horn of Africa. Williams SM, editor. PLoS Genet 10: e1004393.
113. van de Loosdrecht M, Bouzouggar A, Humphrey L, Posth C, Barton N, Aximu-Petri A, et al. (2018) Pleistocene North African genomes link Near Eastern and sub-Saharan African human populations. Science 360: 548–552.
114. Kelleher J, Etheridge AM, McVean G (2016) Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. Song YS, editor. PLOS Comput Biol 12: e1004842.
115. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. (2011) The variant call format and VCFtools. Bioinformatics 27: 2156–2158.
116. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. (2012) Ancient admixture in human history. Genetics 192: 1065–1093.
117. Gutmann MU, Cor J, er (2016) Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. J Mach Learn Res 17: 1–47.
118. Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–18.
119. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16: 1791–8.
120. Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian Computation in Population Genetics. Genetics 162.
121. Thornton KR (2009) Automating approximate Bayesian computation by local linear regression. BMC Genet 10: 35.
122. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.
123. Lintusaari J, Vuollekoski H, Kangasrääsiö A, Skytén K, Järvenpää M, Marttinen P, et al. (2017) ELFI: Engine for Likelihood-Free Inference.
124. Tucci S, Vohr SH, McCoy RC, Vernot B, Robinson MR, Barbieri C, et al. (2018) Evolutionary history and adaptation of a human pygmy population of Flores Island, Indonesia. Science 361: 511–516.
125. Castellano S, Parra G, Sánchez-Quinto FA, Racimo F, Kuhlwilm M, Kircher M, et al.

(2014) Patterns of coding variation in the complete exomes of three Neandertals. Proc Natl Acad Sci U S A 111: 6666–71.

126. Muerdter F, Boryń ŁM, Arnold CD (2015) STARR-seq - principles and applications. Genomics 106: 145–50.

VITA

Aaron Wolf grew up in Arlington, Massachusetts. From a young age he was enamored with origin stories—theological, mythological, or otherwise. Despite having little interest in the "dead rocks" of old cities, he was always interested in the stories of the people who had lived there. He pursued this passion to a B.A. in Biology and Anthropology, *summa cum laude*, from Oberlin College in 2011. As an undergraduate, he dug in the hills of eastern Italy, excavating pre-Roman dwellings, and studied the bones of early Bronze Age urbanites from Jordan. He started his PhD work in Genome Sciences at the University of Washington in 2014. His graduate work with Joshua Akey focuses on the history of admixture between modern and archaic humans, in order to better understand the evolutionary origins of human behaviors. He plans to follow his passion for science and story telling into the world of science communication and policy, with maybe a detour in Hollywood.