INFERRING THE POPULATION HISTORY OF ANCIENT HOMININS THROUGH USE OF THE ANCESTRAL RECOMBINATION GRAPH

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Melissa Jane Hubisz

August 2019

INFERRING THE POPULATION HISTORY OF ANCIENT HOMININS THROUGH USE OF THE ANCESTRAL RECOMBINATION GRAPH

Melissa Jane Hubisz, Ph.D.
Cornell University 2019

This dissertation was inspired by two exciting developments in the field of genomics. The first is the high-quality genome sequencing of ancient archaic individuals, including two Neanderthals and one Denisovan, which has made possible many new insights about human and archaic hominin evolution over the past half million years. Previous studies have demonstrated strong evidence for multiple interbreeding events between these groups, as well as with other unsequenced hominins.

The second development is a new method, called ARGweaver, which infers ancestral recombination graphs (ARGs) from the genome sequences of multiple individuals. The ARG describes the genetic relationships between these individuals along the genome, in the form of local trees with branch lengths describing times to the most recent common ancestor.

In the first chapter, I provide an introduction to ARGweaver and describe several new features that make it applicable to a wider range of data, including integration over phase, accounting for ancient sampling dates, correcting for low-quality genomes, and sampling under the more accurate SMC' model.

In the second chapter, I show how ARGweaver was used to provide strong evidence in favor of a migration event from ancient humans out of Africa over a hundred thousand years ago. These humans likely encountered Neanderthals and admixed with them, leaving segments of their DNA in the Neanderthal

genome.

In the final chapter, I introduce an extended version of ARGweaver that can sample ARGs conditional on a generic demographic model that may include population divergences and migrations. Once ARGs are inferred under this model, the posterior probability of introgression can be computed along the genome for any migration event. I apply this method to human and archaic hominins, and classify 3% of the Neanderthal genome as potentially introgressed from humans. The properties of these segments suggest that this admixture occurred roughly 250 thousand years ago, and there are no signs that natural selection acted against these regions. I also detect lower levels of introgression from an unknown archaic hominin in the Denisovan genome, and possible traces of the same type of introgression in the Neanderthal genome.

BIOGRAPHICAL SKETCH

Melissa Jane Hubisz grew up in Toms River, NJ, graduating from Toms River High School South in 1998. She then attended the California Institute of Technology, earning a Bachelors of Science in Engineering and Applied Science in 2002. After graduation, she moved to Ithaca, NY and joined Dr. Rasmus Nielsen's lab in Biological Statistics and Computational Biology (BSCB) at Cornell University, working as a Programmer/Analyst. It was at this job that she first started studying genetics and evolution. While working there she co-authored several papers and also earned a Masters Degree in Biometry. In 2005 she enrolled in the Human Genetics program at the University of Chicago, eventually joining the lab of Dr. Jonathan Pritchard and working on extensions to the STRUCTURE algorithm. In 2008, she made a difficult personal decision to leave Chicago with a Masters degree and move back to Ithaca. She rejoined BSCB as a programmer/analyst in Adam Siepel's lab, where she has been involved in numerous studies of comparative evolution and population genetics in a wide array of species, from humans and Neanderthals to Streptococcus. In 2015 she decided it was time to complete her PhD, and she enrolled as a graduate student in Computational Biology at Cornell, under Adam Siepel. She was awarded a National Science Foundation Graduate Research Fellowship in 2016.

Outside of her scientific pursuits, Melissa plays violin and piano, and enjoys running and exploring the trails around Ithaca with her dog Colin. Most importantly, she is a devoted mother to her two children, Riley and Maple, and loves spending time with them and sharing her passion for music, books, and the outdoors with them, along with her partner Jay.

To my family, for their love and support.

And to the archaic hominins within all of us, for their mysterious influences.

ACKNOWLEDGEMENTS

It has been a very long road to my PhD, but also an enjoyable and productive one. I need to thank my earlier advisors, Rasmus Nielsen and Jonathan Pritchard, who have been exceedingly supportive and generous to me. I have met so many amazing scientists over the years who have influenced me in various ways, much too many to mention. I especially enjoyed sharing an office with John Novembre, Charles Danko, Ilan Gronau, and Lenore Pipes. Piper Below was an amazing friend during my time at Chicago and taught me so much about the importance of friendship and family, and keeping a healthy perspective in life.

Thank you of course to Adam Siepel, who has supported and encouraged me for quite a long time now, and whose ideas underlie most of the work in this thesis. Matt Rasmussen originally wrote and developed ARGweaver and provided much of the inspiration for this work. And Amy Williams has been an effective local sounding board for me since Adam departed Ithaca.

Thank you to my other committee members, Jim Booth and Andy Clark. Andy Clark has been a steady presence since the very beginning who seems to pop up at crucial junctures in my career with wise advice.

None of this would have been possible without my husband Jay, who really surpasses all reasonable expectations in his support for my career and our family. Thank you also to "Abuela and Pappy", who moved to Ithaca and have provided endless free child care and delicious salads these past two years. And none of this would have been worthwhile without the joy that Riley and Maple bring to my life every day.

Finally, this work was made possible by funding from NSF Graduate Fellowship (grant DGE-1650441) and NIH/NIGMS R35 GM127070.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

ARGWEAVER: OVERVIEW, IMPROVEMENTS, AND EXTENSIONS

1.1 ARGweaver overview

ARGweaver is software originally written and published by Matt Rasmussen as the product of his postdoctoral work in the Siepel lab at Cornell [1]. It was the end of his academic career (at least for now, as he moved to industry), and formed the beginning of my doctoral research. All of the work in this dissertation is based upon using, improving, and extending ARGweaver, and interpreting the resulting ARGs. I shall therefore start this chapter with an introduction to ARGs and ARGweaver. Section 1.1 is adapted from a chapter about ARGweaver, written by Melissa Hubisz and Adam Siepel, in the book "Statistical Population Genomics" and which will be published open-source by Springer.

In the remainder of this chapter, I will describe and demonstrate some of the major modifications that I have made to ARGweaver. The section on phase integration was originally published as part of the Supplementary material for [2].

1.1.1 Introduction

The ARG can rightly be considered the holy grail of statistical population genetics. The ARG represents the history of a collection of related genome sequences, in terms of the *coalescence* events by which segments of genomes

trace to common ancestral segments and the historical *recombination* events that cause patterns of ancestry to differ from one genomic site to the next. Provided the sequences under study are orthologous and co-linear—meaning that they trace to a common ancestral sequence without genomic duplications or rearrangements—the ARG is a complete description of their evolutionary relationships. Moreover, in statistical terms, the ARG provides a highly compact and precise description of the correlation structure of such a collection of sequences. Importantly, the ARG naturally defines a set of recombination breakpoints, a set of haplotypes, and a genealogy for each non-recombining interval in the genome—all objects that are useful starting points for countless population genetic analyses.

Many questions in applied population genetics can be reframed as questions about ARG structure. For example:

- *Recombination rate estimation.* Recombination rates can be estimated by simply counting recombination events and dividing by the total branch-length of the ARG.
- *Estimation of allele ages or mutation rates.* Mutation events can easily be mapped to branches within the ARG by maximum parsimony, enabling straightforward estimation of allele ages and mutation rates.
- *Local ancestry inference.* The local ancestry structure of an admixed individual (i.e., which genomic segments derive from which distinct source populations) can be determined by tracing the individual's two diploid lineages in the ARG and identifying the source population with which each genomic segment clusters, as well as the recombination events that terminate these segments.

- *Demography inference.* More general information about demographic history (such as population sizes, migration rates, and divergence times) is also embedded in the ARG. A demographic model can fairly easily be estimated from a known ARG by making use of the counts of coalescence events within and between populations.
- *Detection of sequences under selection.* Natural selection can be detected by identifying local distortions in the ARG, for example, unusual clusters of coalescence events or extremely deep times to most recent common ancestry. Recent progress using this approach has recently been published [3]

In practice, however, the true ARG is impossible to know with certainty. The "ARG space," consisting of every possible ancestral history of a set of genomes, is astronomically large, and the information in genome sequences is insufficient to choose a specific ARG above all others. But, given a model of coalescence, recombination, and nucleotide substitution, it is possible to compute the probability of an observed data set under a particular ARG, and it will generally be true that some ARGs are much more likely to have produced the data than others. The approach taken by ARGweaver is to sample from the posterior distribution of ARGs, given a collection of genome sequence data and a reasonable set of modeling assumptions. This approach is computationally expensive, and it has the drawback of producing a complex and unwieldy output—a collection of potential ARGs, none of which is exactly correct, but which, in the aggregate, reflect certain properties of the true ARG. Nevertheless, this approach can be extremely powerful, potentially providing insights into the structure of the data and the evolutionary history of the sample that are not easily obtained using simpler methods.

1.1.2 What is an ARG?

An ARG represents all ancestral relationships among a collection of genomes (see Figure 1.1). If n is the number of (haploid) genomes under study (usually from $\frac{n}{2}$ diploid individuals), then at the present day, there are n lineages in the ARG. As we trace these lineages back in time at a particular genomic location, we will find that distinct lineages gradually *coalesce* into shared ancestral lineages, until all n lineages have found a single most recent common ancestor. These coalescence events define a tree known as a *genealogy* that fully describes the evolutionary relationships among the present-day genomes at the locus in question.

However, *recombination* events in the history of the sample can cause the genealogy to change from one genomic location to the next. Looking backward in time, a recombination at a particular genomic location has the effect of splitting a lineage into two, with one path representing the evolutionary history to one side of the breakpoint and another path representing the history to the other side. The ARG captures these recombination events together with the coalescence events. As one follows a lineage upward in the ARG, that lineage may either merge with another lineage, representing a coalescence event, or it may split into two lineages, representing a recombination event (Figure 1.1A). In the case of recombination events, the junction in the ARG is also labeled with the genomic position of the recombination (this information is not relevant for coalescence events).

Based on these labels for recombination events, one can extract a local tree for any position in the genome from the ARG. First, one identifies the lineage associated with each present-day sample. These lineages are traced backward

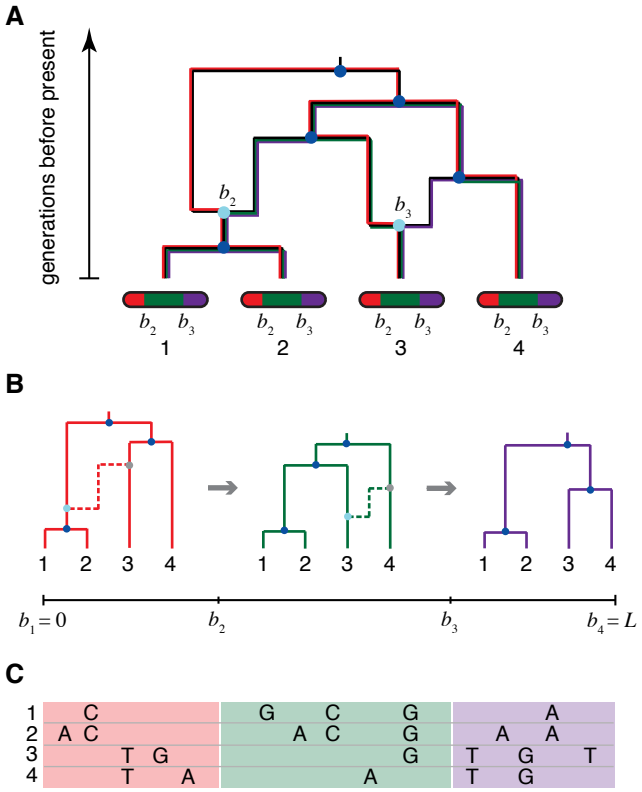

Figure 1.1: A. Schematic of an ARG with 4 lineages in the present, and 2 ancestral recombination events along a region of length L. Tracing the history backwards until all lineages have reached a common ancestor. B. An alternative view of the ARG depicted in A, showing the local tree between each pair of recombination breakpoints. The dotted lines on the tree show the recombination event which transforms the tree on the left side of the breakpoint into the tree on the right side. C. The data underlying this ARG, where only derived alleles at variant sites are shown. Figure adapted from [1].

through the ARG, and coalescences between them are noted. When a recombination event is identified, one of the two possible paths is selected based on the relationship of the position in question to the annotated recombination breakpoint. Specifically, if position is to the left of the breakpoint, then the left path is taken; and if the position is to the right of the breakpoint, then the right path is taken. (Because recombination breakpoints by definition occur between nucleotides, one of these two cases must hold.) Thus, the paths from the present-day samples to the root will coalesce only, never splitting, and therefore must define a tree. Furthermore the tree will be the same for all genomic positions between two recombination breakpoints, differing only between positions on opposite sites of a breakpoint.

Another way to think about the ARG, then, is that it defines a series of operations on trees along the length of a chromosome. As one walks along a chromosome from left to right, the local tree remains fixed until a recombination breakpoint is encountered, and then that tree is altered to form a new tree, in the specific manner defined by the change in path at the corresponding recombination node in the ARG (Figure 1.1B). The ARG, therefore, can be thought of as being interchangeable with a sequence of local trees and the associated recombination events that transform each tree to the next. In practice, this is the representation of the ARG assumed by the Sequentially Markov Coalescent (SMC') and used by ARGweaver, and in this chapter we will generally treat the ARG as a collection of trees and recombination events. Nevertheless, it should be noted that this representation does not strictly capture all of the information in the ARG. The full ARG also describes "trapped genetic material" that falls between two linked ancestral loci, but is not passed on to any present-day sample. Ignoring this trapped material substantially simplifies modeling and inference

algorithms, with what appear to be only minor costs in accuracy [1, 4, 5].

1.1.3 Why would you want to estimate an ARG?

As discussed above, if the ARG could be estimated accurately and easily, it would be useful for almost every question in population genetics. In practice, of course, there are limitations in the accuracy of inferred ARGs, and they require substantial time and effort to obtain. So, when does it make sense to take the trouble to run ARGweaver, instead of making use of simpler or more standard population genetic summary statistics and tools? Some reasons to consider sampling ARGs with ARGweaver include:

- *Trees/genealogies.* ARGweaver estimates explicit genealogies (with branch lengths) along the genome, considering both patterns of local mutation and local linkage disequilibrium. It may be particularly interesting to inspect trees at particular regions suspected to be under selection or to have experienced introgression.
- *Times/dates.* These trees allow the timings of various events to be estimated, including times to most recent common ancestry, other coalescence times, and the ages of derived alleles. If desired, posterior expected values of these times can be computed by averaging over the sampled trees.
- *Ancient introgression.* ARGweaver is a powerful method for detecting introgression and identifying specific introgressed haplotypes, particularly ancient introgression events that conventional methods may miss (e.g., [2]).

- *Bayesian treatment of uncertainty.* Unlike many simpler methods, ARGweaver attempts to fully account for the uncertainty in the ARG given the sequence data and an evolutionary model, by sampling from a posterior distribution of ARGs. This approach can mitigate biases from the inference method in addressing biological questions of interest.
- *Flexibility in addressing "custom" evolutionary questions.* By producing explicit ARGs, ARGweaver allows almost any evolutionary question to be addressed, including unusual ones not easily addressed with standard summary statistics (For example: at what fraction of sites do individuals A and B coalesce with one another before either coalesces with individual C? What is the average TMRCA for genes of functional category X? Are recombination events more likely to occur in introns or intergenic regions?)
- *Technical limitations of the data*. ARGweaver can accommodate unphased data, low-coverage sequences, archaic samples, and other unusual data types that may not be easy to analyze using other methods.

1.1.4 Practical considerations

ARGweaver is designed to run on genome sequencing data for small to moderate numbers of individuals—anywhere from two to a maximum of about 100. These individuals should be unrelated but come from the same species or from recently diverged species (such as humans and chimpanzees). Phasing of diploid genome sequences is not necessary—ARGweaver can phase "on the fly," integrating over possible phasings—but the algorithm converges faster and, in some cases, performs better on phased data (depending on the rate of phasing errors). Similarly, ARGweaver can be used on low-coverage sequenc-

ing data, making use of genotype probabilities to weight the observed bases, but high-coverage sequence data is always preferable.

In gauging the feasibility of ARG inference, it is important to recognize that the processes of mutation and recombination are opposing forces in reconstructing an ARG. The more mutations there are, the more information there is to guide the inference of tree topologies (genealogies). Recombination events, however, break up the sequences into smaller blocks, effectively limiting the information for tree inference in each block. Thus, the quality of ARG inference depends on the ratio of mutation to recombination rates per nucleotide position. In human data, this ratio is close to one, but recombination events tend to be concentrated in recombination hotspots, which makes the effective ratio greater than one for most of the genome. ARGweaver appears to work quite well in this setting. Nevertheless, the method works better when this ratio is even higher, and it will break down if this ratio falls significantly below one. Another consideration is ARGweaver's assumption of at most one recombination event per site (see below), which generally appears to have little effect but could lead to biased estimates in cases of particularly high recombination rates, large sample sizes, large evolutionary distances, or large effective population sizes. Finally, because ARGweaver depends on haplotype-scale information for inference, it is generally not useful for short sequences, deriving, for example, from RAD-seq or a de novo short-read assembly.

In terms of the number of genomes analyzed, the "sweet spot" for ARGweaver is generally between a handful of individuals and a few dozen. As the number of genomes increases, more approximate models (such as the Li and Stevens model [6]) or conventional population genetic summary statis-

tics become increasingly accurate and informative, and the relative advantage of using ARGweaver over other methods decreases. In addition, the run time and size of the ARGweaver output increases with the number of genomes, and these factors become prohibitive with more than about 100 samples. Running ARGweaver genome-wide generally requires breaking the genome into chunks of a few megabases and running ARGweaver in parallel on each chunk using a computer cluster. When running ARGweaver genome-wide is not a realistic possibility, it may still be of interest to apply ARGweaver to specific genomic regions of interest, such as candidate selective sweeps or introgressed regions. It may also be useful to run ARGweaver on subsets of the available genome sequences, for example, to shed light on genealogy structure, ancient introgression, or allele age—features ARGweaver may estimate more accurately than other methods.

Another practical consideration is that while ARGweaver's output is richly informative, it is not straightforward to interpret. The program does come with tools to compute various local summary statistics from sampled ARGs, including times to the most recent common ancestor, allele ages, and distances between samples. But many less standard analyses will require custom programs to extract the desired information from ARGs or local genealogies.

1.1.5 ARGweaver algorithm

ARGweaver uses a Markov chain Monte Carlo (MCMC) algorithm to sample ARGs at frequencies proportional to their probability, conditional on the observed DNA sequence data (X) and the model parameters (θ). The MCMC al-

gorithm starts with an initial ARG, G^0, and then repeatedly removes a subset of the ARG and resamples that subset from an appropriate conditional probability distribution. This process generates a sequence of ARGs, $G^0, G^1, \ldots, G^m$, where m is the total number of iterations of the algorithm. Although G^0 may be a poor guess with low probability, by sampling each new G^i according to the appropriate distribution, the chain will eventually converge to the desired distribution—i.e., for sufficiently large i, G^i will represent a draw from the posterior distribution over ARGs given the data and the model, $P(G^i|X, \theta)$. In practice, it is customary to plot the posterior probability as a function of the iteration number, i, observe the point at which it ceases to trend upward and becomes stable, and then to discard the ARGs sampled before this point (from what is known as the "burn-in" of the MCMC algorithm).

Even once the algorithm has converged, successive samples G^i and G^{i+1}—while they both represent samples from the posterior distribution—are not *independent* samples. Rather they are strongly correlated, since only part of the ARG is resampled on each step of the algorithm. Therefore, in order to achieve a distribution of nearly independent ARGs—both to save space and processing time, and to better assess the variance of estimates derived from the samples—it is useful to "thin" the chain, recording only every j^{th} sample (the default thinning parameter in ARGweaver is $j = 10$). After discarding the initial "burn-in" and performing thinning, the ARGs G^i that remain can be stored and treated as a collection of samples representative of the distribution of ARGs given the data and the model, $P(G|X, \theta)$.

The technical details of the ARGweaver algorithm will not be reviewed here (see [1]), but the main idea is to remove a single haploid genome from the ARG,

and then to "thread" this genome back through the ARG, by sampling both its coalescence points with the remaining sequences and the the associated recombination points. There is also another, slightly more complicated, version of this threading operation, called "subtree threading," that resamples internal branches in genealogies, and is essential for ARGweaver to efficiently explore the full space of possible ARGs. In both cases, a hidden Markov model (HMM) is used to efficiently sample new coalescent points for the new lineage across the chromosome. This HMM depends on several key modeling assumptions, which are important for users to understand, and which, therefore, will be reviewed in the next section.

ARGweaver model and assumptions

The HMM underlying ARGweaver depends on the following assumptions:

- *SMC' or SMC*: ARGweaver was originally written under the Sequentially Markov Coalescent model [4], but has been adapted to use the closely related SMC' [5]. The differences between these models are subtle and will be described in Section 1.5. These models posit that the distribution over genealogies at each nucleotide position directly depends only on the genealogy at the previous position, not on the genealogies at positions further upstream—a feature known in probability theory as the *Markov property*, after the Russian mathematician Andrey Markov. More formally, the SMC and SMC' assume that the genealogy at position $i + 1$ is independent of the genealogies at positions $1, \ldots, i - 1$, given the genealogy at position i. While the SMC' is technically more accurate, the SMC model may be considerably faster on data sets with large numbers of samples.

ARGweaver therefore allows the user to choose either model (SMC by default, `--smc-prime` for the SMC').

- *Discrete time*: All recombination and coalescent events are assumed to occur at a predefined collection of discrete time points. The total number of time points, K, can be chosen by the user (using `--ntimes <K>`) and can be arbitrarily large, with the ARGweaver model approaching a continuous-time model as K approaches infinity. However, the computational complexity of the threading algorithm is proportional to K^2, so, in practice, K must be kept modest in size. The default value of K in ARGweaver is 20. The time points are uniformly spaced on a logarithmic scale, so that they are more closely clustered at recent time points, when there are more lineages and coalescence rates are larger. The algorithm forces all lineages to coalesce by the final time point, t_K.

- *No more than one recombination event between neighboring nucleotides*. For simplicity, the algorithm permits at most one recombination event at every "step" along the sequence, meaning between two adjacent nucleotide positions. This assumption means that adjacent genomic positions must either have identical genealogies or ones that differ by a single recombination event. In practice, this assumption is minimally restrictive, because the information about genealogies comes primarily from variable sites, which tend to be sparse along the genome. If ARGweaver should need to account for multiple recombination events between variable sites, it typically can spread those events across a series of intervening invariant sites with minimal impact on accuracy. If the data are such that multiple recombinations between neighboring sites occur frequently, then it is likely that the haplotype structure is too broken down to make use of ARGweaver.

- *Population size known*: ARGweaver assumes that the effective population size N_e (which determines the coalescence rate) is provided by the user. In the simplest case, a single global value of N_e can be provided. But ARGweaver can accommodate different values of N_e for different discrete time intervals. Values of N_e can typically be obtained from the literature or estimated from the same data using one of the many available programs for inferring demographic histories (such as SMC++ [7], PSMC [8], MSMC [9], G-PhoCS [10], and diCal [11]). Note the user-provided values of N_e define a "prior" for coalescence rates in ARGweaver, so it is not necessary for them to be perfectly estimated; ARGweaver will consider the data together with this prior distribution in sampling coalescence events.
- *Mutation and recombination rates known*. The ARGweaver model also depends on pre-defined mutation and recombination rates. These rates can be assumed to be constant across the genome, or variable rates can be provided in a position-specific map along the genome. These values are also "priors" in the same sense as the population size (see above).
- *Jukes-Cantor model of base substitution*. ARGweaver makes use of a Jukes-Cantor model for nucleotide substitutions. This model assumes that all nucleotide substitutions are equally probable—an obvious oversimplification, but one that seems to have minimal costs at the close evolutionary distances typically considered by ARGweaver. The symmetries inherent in the Jukes-Cantor model can be exploited to optimize the likelihood calculations in ARGweaver.

1.2 Integrating over haplotype phase

The original ARGweaver method was designed for phased genomes, however most sequencing technologies produce unphased genomes. Many methods exist for computationally phasing genomes [12–14]; however accurate phasing requires a large reference panel, which is not always available. Even then, switch errors between 0.5-5% are observed [12], depending on the reference panel size and method used. When phasing is performed prior to running ARGweaver, the ARG will be conditioned on all the phase errors. An error in the phase usually causes ARGweaver to erroneously infer recombination or mutation events. Ideally, ARGweaver would be able to create ARGs for unphased samples, or take phase uncertainty into account. To this end, we developed an approach to integrate over possible phasings while running ARGweaver. This saves the user a data processing step, eliminates the need for a reference panel, and is more robust to specific phasing errors.

When using phase integration, all individuals are randomly phased at the start of the algorithm (or if available, initialization may be done with pre-phased haplotypes). Most of the algorithm is performed conditional on the current phase of each individual. However, the leaf threading operation is performed without regard to the phase of the individual whose lineage is being re-threaded, and is followed by re-sampling the phase for that individual. The genotype phases of other individuals are held constant during this step. Leaf threading can be performed on an unphased individual by summing over the two possible phase configurations of each heterozygous site when computing the probability of the sequence data for a particular ARG. These probabilities are used as the emissions probabilities in the hidden Markov Model,

which is used to sample the new threading. After the threading is complete, the phase for that individual is sampled at each heterozygous site according to the relative probabilities of the two possible phasings under the newly sampled ARG. Note that the phasings are held constant during subtree threading steps. Phase sampling is implemented in the ARGweaver source code available at `http://github.com/CSHLSiepelLab/argweaver`, using the `--sample-phase` option.

We performed a simulation study to assess the effectiveness of phase integration. We used `ms` [15] to simulate a 1Mb region with population size 10,000, mutation rate 2.5e-8/bp/generation, and recombination rate 1.5e-8/bp/generation for 2, 4, and 8 diploid individuals. We then ran ARGweaver on each data set, with and without phase integration, and with increasing levels of phase errors in the initial data. Figure 1.2 shows the ability of ARGweaver to recover the total branch length of an ARG, as well as the number of recombination events, as the error rate in phase increases. Overall the effects are promising; phase integration seems to help the inference, and works quite well for 2 and 4 genomes even when the phase is completely random. For larger numbers of genomes it still overestimates the statistics, however not as badly as without phase integration, and does quite well when the amount of phase error is $\leq$ 10%. We therefore recommend a hybrid solution, in which the samples are initialized with pre-phased haplotypes when possible, but phase integration is used to integrate over possible switch errors.

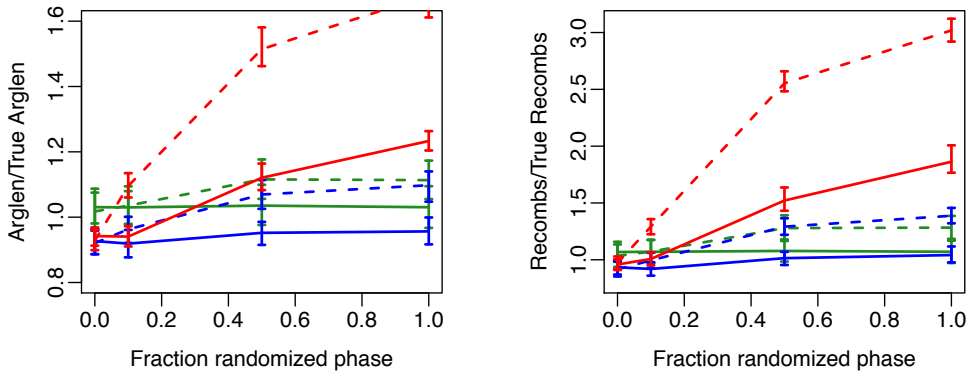

Figure 1.2: Simulation results showing effects of phase integration. ARGs were simulated with $n = 2$ (green), $n = 4$ (blue), and $n = 8$ (red) diploid genomes. The true phases were randomized with probability shown on the x-axis. ARGs were inferred with phase integration (solid lines) and without (dashed lines). The y-axis shows the inferred total length of the ARG (left) and number of recombinations (right), compared to the truth.

1.3 Accounting for ancient sampling dates

The original implementation of ARGweaver was not written with ancient samples in mind; it assumed that all genomes represented current-day individuals. However, ancient genomes are becoming increasingly common; not only the Neanderthal [16] and Denisovan [17] genomes, which will be extensively analyzed in this thesis, but there are now thousands of ancient modern human genomes available [18–22]. There are also ancient DNA data available for other species, including the woolly mammoth [23], cave bear [24], and ancient horse [25].

Although it might be possible to estimate the age of ancient samples along with the ARG, in practice the age of an ancient genome is estimated more simply by computing the fraction of missing mutations to an outgroup, compared to a modern-day sample. Given the age of the genome as an input, it is fairly

straightforward to modify ARGweaver to take this age into account. First, the age of the sample is rounded to the nearest discrete time used in the ARGweaver model, t_a. The leaf branches coming from an ancient sample start at this time, instead of at $t = 0$. This means that, when threading an ancient sample, the only valid states in the HMM are those with times $\geq t_a$. Ancient samples also affect the coalescence rates used when threading any branch. When threading non-ancient samples, the rates of coalescence depend on the number of branches in each time interval. When all the samples are present-day, this number only decreases going backwards in time, as coalescences reduce the number of branches. However, with ancient samples, this value may increase, since lineages not present at $t = 0$ do appear at t_a.

The changes described above were implemented in the ARGweaver software and can be used with the option `--age-file`, which takes a file name listing ancient samples and their ages in generations. Figure 1.3 shows that this feature effectively corrects for bias in statistics that exist when the age of ancient samples are ignored. In this example, we simulated 10 haploid lineages; 4 of which were modern-day samples, and one each with ages 50kya, 100kya, 150kya, 200kya, 250kya, and 300kya. We used a generation time of 29 years, and other demographic parameters matching ARGweaver's defaults (a mutation rate of 2.5e-8/bp/generation, and a recombination rate of 1.5e-8/bp/generation, and a constant population size of 10000). This example is fairly extreme, in that most of the samples are ancient, and many of them are older than most ancient samples from which DNA has been successfully extracted and sequenced. Overall, it is impressive that many of the statistics of the ARG are estimated quite well even when all of the samples are incorrectly treated as present-day samples. This is true for the time to the most recent com-

mon ancestor (TMRCA) as well as the total branch length. However, there is a noticeable bias in the branch lengths of the ancient leaves, as well as in estimates of allele age. When ARGweaver is given the sample ages, it successfully incorporates these into the algorithm and the biases disappear.

1.4 Correcting for low-quality genomes and application to *Sporophila*

ARGweaver has also been extended to work on low-quality genomes. This was inspired by a collaboration with researchers at the Cornell's Lab of Ornithology. They had sequenced the genomes of 72 individuals from nine species of capuchino seedeaters from the genus *Sporophila*, collected across South America [26]. The sequencing coverage of each individual ranged from 1.9x to 9.8x. When running ARGweaver on higher-coverage genomes, standard practice is to mask out uncertain genotypes. In this case, however, most of the genotypes have some uncertainty and it is necessary to account for this in a more rigorous way, as masking would sacrifice too much of the data.

The implementation of this feature was fairly staightforward. It only required a modification to the emissions probabilities used in ARGweaver's threading HMM. These are computed at each site using Felsenstein's algorithm [27]. The algorithm is initialized by assigning a vector (p_A, p_C, p_G, p_T) at each leaf so that $p_i = I[a = i]$, where a is the observed allele for the sample corresponding to the leaf node. But, this vector represents probabilities of each allele, so they can simply be used instead, allowing ARGweaver to account for genotype uncertainty.

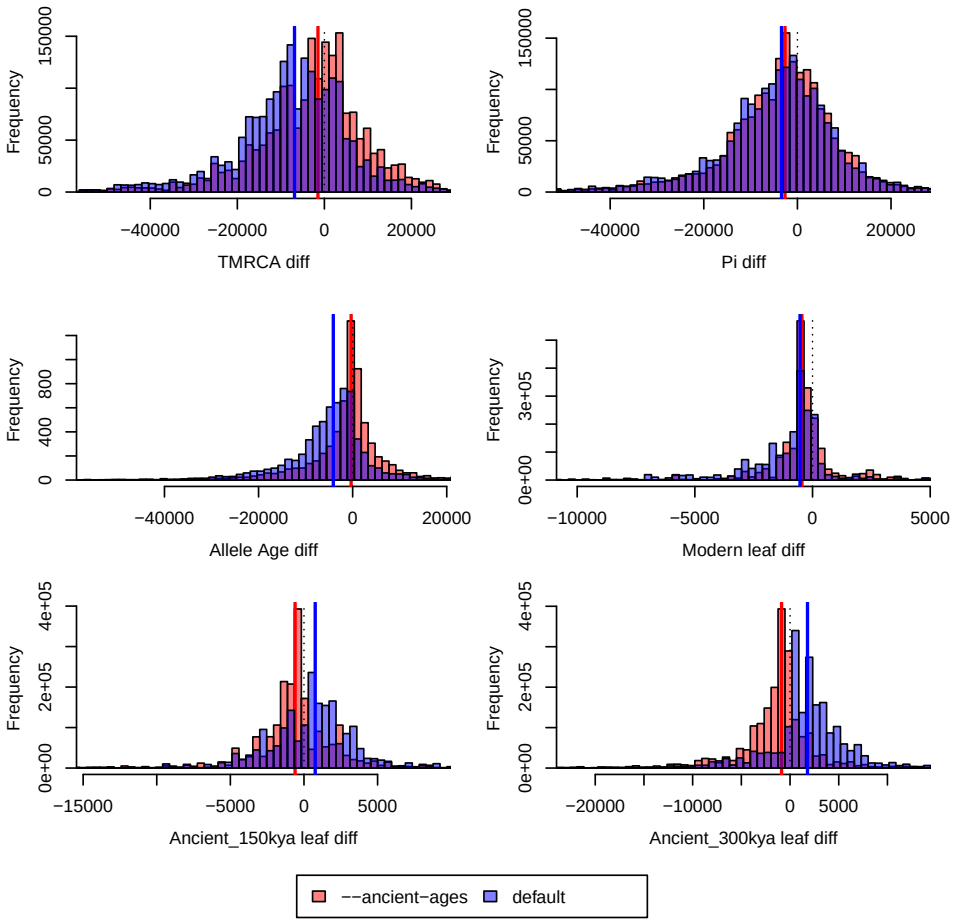

Figure 1.3: **Effect of accounting for ancient sampling ages.** Looking at the ARG across a 2Mb region, at every base we compute the difference between the true statistic and the median from ARGs sampled across 2000 MCMC iterations. The pink distribution shows the ARGs inferred while accounting for ancient sampling dates; the blue uses the default parameter. (Purple is the overlap between the two). The dotted black line is at $x = 0$, and the red and blue lines are at the medians of the pink and blue distributions. The statistic for each plot is named in the x-axis, and the names are as follows: TMRCA (time to most recent common ancestor, in generations); Pi (average distance between two leaf nodes, in generations); Allele age (age of derived alleles); Modern leaf (coalescence time of a leaf node for a present-day sample); Ancient_150kya leaf (coalescence time of a leaf sampled 150kya), and Ancient_300kya leaf (coalescence time of a leaf sampled 300kya).

I demonstrate the effectiveness of this approach with a simple simulation study. I used `msprime` [28] to simulate 10 500kb regions with 6 haploid genomes, population size 10000, mutation rate 2.5e-8/bp/generation, recombination rate 1.25e-8/bp/generation. I then used a custom script to introduce errors, so that with probability $r \in \{0.0001, 0.0005, 0.001, 0.005\}$, the true allele was changed to a randomly chosen allele. I then ran ARGweaver on each data set twice; an "uncorrected" run which did not use allele probabilities, and a "corrected" run in which this simple error model was taken account (so that the probability of the observed, sometimes incorrect allele, is set to $(1 - r) + r/4$, and all other alleles have probability $r/4$). The difference in performance is illustrated in Figure 1.4 and shows a striking improvement. Each plot shows the difference in estimated TMRCA from the true TMRCA. When $r = 0.0001$, the distributions are both centered at 0 and mostly overlapping. For all other values of r, the "uncorrected" TMRCA is overestimated, getting worse as the error rate increases. But in the "corrected" ARGs, the distribution of TMRCAs does not appear to change as the error rate increases.

In real data, the application is more complicated, both because models of genotyping error are different for heterozygous and homozyous sites, and because most variant calling algorithms do not output well-calibrated probabilities of error that properly corrects for reference bias [29]. Nevertheless, popular genotype callers, including GATK [30], compute genotype probabilties. I modified ARGweaver to read in VCF files and genotype probabilities encoded in them using any of the PL (phred-scaled genotype likelihoods), GL (genotype likelihood), or PP (genotype posterior probability) formats [31].

Applying this to the *Sporophila* set, we seem to have at least some success in

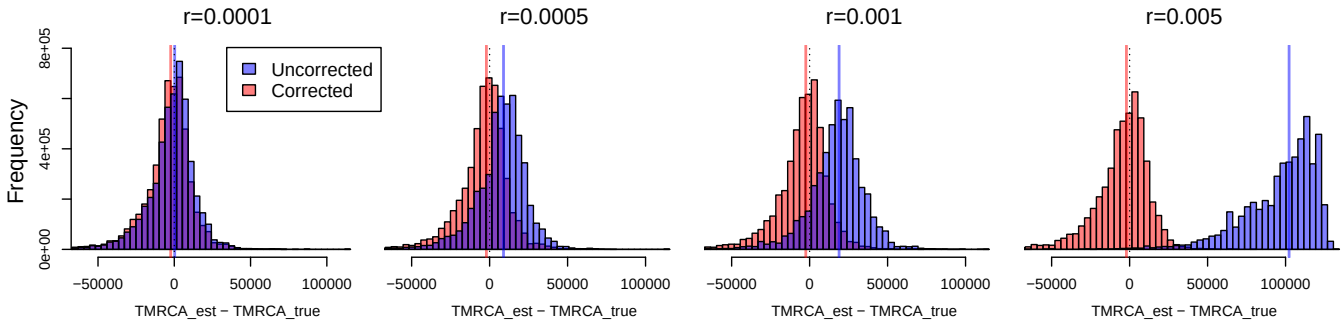

Figure 1.4: **Effectiveness of incorportating allele probabilities**. Each histogram shows a distribution showing the difference in ARGweaver's estimate of the TMRCA from the true TMRCA, calculated at every base across a 5Mb simulation. For ARGweaver's estimate I used the 50% quantile over 50 ARGs sampled across 500 iterations, after 500 iterations of burn-in. Each plot shows two overlayed historgrams calculated on the same data set. The alleles in each data set are randomized at every individual/base at the rate indicated in the heading. The blue plots show the distribution obtained when the data is used without allele probabilities; the pink plots show the distribution when ARGweaver takes the error model into account. The vertical blue/pink lines show the median of each distribution, and a dotted black line is at $x = 0$ for reference.

generating reasonable ARGs despite the low coverage of many of the samples. While the data set consisted of samples from nine different species, they have diverged in the past 100,000 years and have a very high population size [32], so that almost no genetic variation is observed between individuals from different species [26] A previous analysis of the same data set [26] identified several local peaks in F_{ST} [33] between different species, that may underlie the phenotypic differences in color pattern and mate selection observed in the field. The ARGs produced by ARGweaver tell the same story; the local trees do not seem to be organized by species along most of the genome. However, we see regions where species are clustered that coincide with the previously detected F_{ST} peaks. Fig-

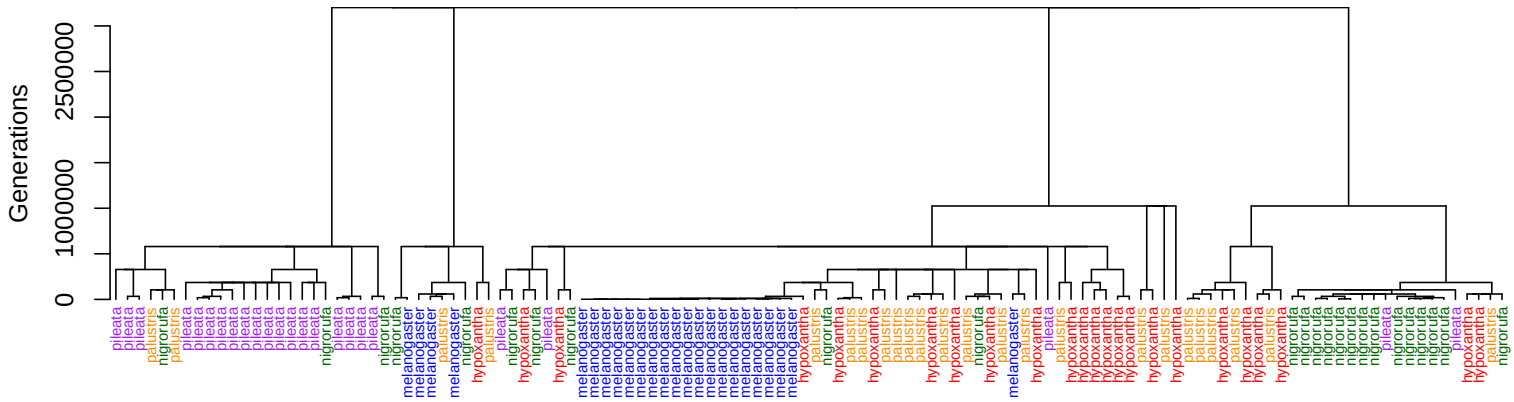

Figure 1.5: Tree sampled by ARGweaver for the *Sporophila* data set at Contig252 position 470000. This position overlaps peak in F_{ST} at a gene (ASIP) associated with plumage color [26]. This tree represents individuals from the five species with the most individuals sampled. The species name is written underneath each leaf and is color-coded: purple=*S. pileata*; red=*S. hypoxantha*; orange=*S. palustris*; green=*S. nigrorufa*; blue=*S. melanogaster*.

ure 1.5 shows one example tree produced by ARGweaver in the ASIP gene on Contig252, which is one of the highest F_{ST} peaks. In this tree, there is a high degree of clustering by species. There is also rapid coalescence of most of the *S. melanogaster* individuals (blue), and also of the *S. nigrorufa* (green) individuals to a lesser extent, which is suggestive of selective sweeps in these populations at this gene. Other members of our group are developing methods to more formally classify potential selective sweeps from the ARG output.

1.5 Estimating ARGs under the SMC′

1.5.1 Introduction to the SMC and SMC′

ARGweaver was originally designed under a discretized approximation of the Sequentially Markov Coalescent (SMC) model. However, it has been shown that the SMC′ more accurately captures the properties of the full coalescent-with-recombination (CwR) [5,34]. The SMC and SMC′ are both Markov models which describe the distribution of a local tree at position $i+1$ on a chromosome, given the local tree at position i and demographic parameters (rates of coalescence and recombination). The only difference between the SMC and the SMC′ is a subtle change in how branches broken by recombination are treated by each model. To describe the difference, it is necessary to first overview the models. A tree at position i with n leaf nodes is denoted as T_i^n. Each branch of the tree represents a path, tracing the genetic ancestors who passed along this locus through the generations. When two branches merge (coalescence), it means that the branches have found a common genetic ancestor.

If recombination did not occur between sites i and $i+1$ in any of the ancestors represented in T_i^n's branches, then T_{i+1}^n will be identical to T_i^n. Otherwise, we assume that recombination is rare enough, and the branches short enough, that no more than one recombination event occurs between any two sites. This recombination is assigned a branch (b) adn time (t) according to which ancestor was the source of the recombined chromosome. This ancestor passed along, as a single chromosome, a fusion of their mother's and father's chromosomes, with the split point being between positions i and $i+1$. The genetic history at site i might follow this ancestor's mother, while the history at site $i+1$ follows their

father. So, while this chromosome has the same history from present day until time t, the history before time t is split into two separate branches, until farther back in time when these branches recoalesce.

Thinking about this as a spatial process across a chromosome, when a recombination event is encountered on a local tree at position i, T^n_{i+1} will have all the same branches as T^n_i, except that the recombination branch b "breaks" at time t and has a different history at position $i + 1$. In the SMC, this branch would choose a new common ancestor among all the other branches remaining in the tree at or before time t (including the root branch, which extends to infinity). The SMC' improved this model by recognizing that the broken branch also represents an existing lineage in the population, and that it may so happen to be the most recent common ancestor of the new branch. In other words, the recombining ancestor's parents may happen to be more closely related to each other at site $i + 1$ than to any other lineages in the tree. In this case, while T^n_i and T^n_{i+1} technically trace through different ancestors, the topology and branch lengths of the two trees are identical. This type of recombination is referred to as "invisible", "circular", or "bubble" recombination, and it is allowed under the SMC' but not the SMC.

Note that invisible recombinations are already possible in ARGweaver under the SMC due to the discretization of time, because the broken branch may re-coalesce on the sister or parent branch at the same time as in the previous tree (an event that would be virtually impossible using a continuous-time model). However, the SMC' models these invisible recombinations in a way that is more consistent with coalescent theory. The fewer lineages that exist in the local tree at the time of recombination, and the longer the branch, the more likely that

the broken branch will choose itself as its recoalescence point. Conditional on the coalescence of a recombining branch occuring at a given time, the coalescence branch is *a priori* chosen uniformly among all branches that exist at that time. So, the probability of choosing the recombining branch is inversely proportional to the number of branches in the tree at the time of coalescence. In general, when performing inference, the SMC may under-estimate the recombination rate because it does not infer enough invisible recombination events, and this problem will be worse in time intervals when there are fewer lineages and invisible recombinations are more likely.

1.5.2 SMC′ implementation

Internally, ARGweaver stores the ARG as a series of "blocks" across a chromosome. Each block has a start and an end coordinate, a local tree that is constant across the entire block, and a recombination/recoalescence event which occurs after the final site in the block, producing the local tree in the next block. The transition probabilities of the HMM (i.e., the probability of the new branch coalescing at a particular point in the tree given its coalescence point at the previous site) are the same for all pairs of sites within a block, and need only be computed once per block (assuming that the recombination rate is also constant across the block). Furthermore, because of the assumption of only one recombination event per local tree, new recombination events can only be sampled between sites within the same block, since a recombination event already exists between neighboring blocks. Also recall that ARGweaver samples recombinations and coalescence events separately. The threading HMM samples coalescence points for the new branch, while integrating over possible recombination events. The

recombination events are chosen in a subsequent step, conditional on the chosen coalescence points.

When implementing the SMC′, we chose to treat self-recombinations as a separate case. Self-recombinations are not stored in the ARG structure at any point during the algorithm, but can optionally be sampled prior to storing the ARG (with the option `--invisible-recombs`). Most of the challenge in implementing the SMC′ is in adjusting the transition probabilities of the HMM to account for the possibility of invisible recombinations. The fact that a lineage which is broken by recombination may re-coalesce back onto itself means that this extra lineage must be accounted for in the coalescence probabilities.

The possibility of a lineage re-coalescing back on itself means that calculating the transition probabilities under the SMC′ will be more computationally complex than it was under the SMC. Intuitively, the reason for this is that under the SMC, when the new branch is broken by recombination, it must recoalesce onto an already-existing lineage of the tree, and these branches are all known, so that the recoalescence probabilties are a function of the tree with the broken lineage removed, T_i^{n-1}. Therefore, the recoalescence probabilities are independent of the state chosen at site i. Under the SMC′, we need to consider the coalescence time of the new branch at the previous site in order to calculate the probability of recoalescence at any point in the tree. Fortunately, we can simplify this dependence by breaking the calculations into three cases: time intervals when the new branch at the previous site has not yet coalesced, the time interval when the new branch coalesces, and time intervals when the new branch does not exist because it coalesced more recently. In the first case, we add one to the lineage counts which go into the coalescence probabilities; in the second case, we add

one in the half-time interval immediately before the coalescence but not after, and also need to consider the existence of an additional node when computing the probabilities of choosing a specific branch; in the third case, there is no extra branch to consider and the calculations are the same as in the SMC.

As an example, consider the case of computing the transition probability of the threading branch v coalescing at state (y, b) at site $i + 1$ given that it has coalesced at state (x, a) at site i. The state notation (x, a) indicates that v coalesces onto branch x at time t_a. For now, assume that sites i and $i + 1$ are in the same block, so that there is no recombination currently sampled between them, and the local trees T_i^{n-1} and T_{i+1}^{n-1} are the same. We will further assume that $x \neq y$, so that a recombination must have occurred on branch v for this transition to be possible. This situation is illustrated in Figure 1.6, and shows the "broken lineage" that must be considered under the SMC′. The broken lineage exists in the time range $[t_k, t_a]$.

When $b < a$, the broken lineage is present throughout the duration of the recombination and recoalescence of v. Therefore, it is simple to adjust the transition probability from the original SMC in this case: we simply add 1 to the lineage counts used for computing coalescence probabilities. This includes both the number of lineages used for computing coalescence rates, as well as the factor describing the probability of choosing a specific branch for coalesence.

When $a < b$, it is somewhat more complicated. In this case, the extra lineage exists in the range $[t_k, t_a]$, and 1 is added to the coalesence rates used in this time interval. But in the range $[t_a, t_b]$, the extra lineage no longer exists, and the coalescence rates used by the original SMC model are used.

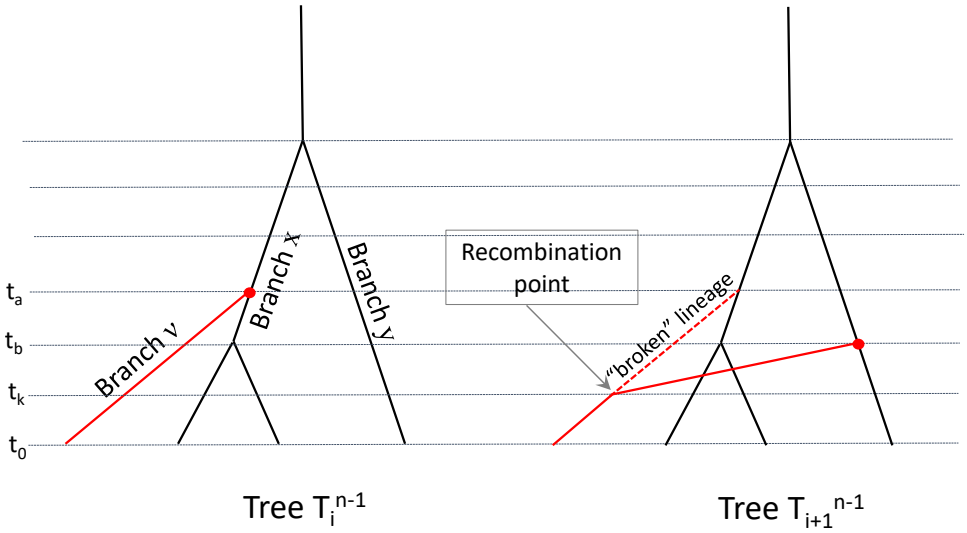

Figure 1.6: Example of a recombination on the threading branch v between two sites with no previous recombination events. Branch v coalesces at state (x, a) at site i, and at state (y, b) at state $i + 1$, with a recombination on v at time k. Under the SMC′, the broken lineage is a candidate for re-coalescence of v and needs to be taken into account when computing coalescence probabilities.

When $a = b$, the coalescence rate in the half-time interval above a (which is rounded to a) does not contain an extra lineage, whereas the intervals below a do. Furthermore, because the broken branch forms an additional node at time b, the number of possible branches existing at this time increases by 2 (this does not affect the coalescence rate, but the probability of choosing each branch when coalescence occurs).

Similar logic is used for other cases (such as $x = y$, or when there is already a recombination between the adjacent trees). The derivations are much the same as in the SMC model, with additional lineage counts added when the broken lineage exists, and an extra node added where it coalesced at the previous site. One other addition is in the case where there is no previous recombination, and v coalesces at the same point at sites i and $i + 1$. As in the SMC, this could

happen because there was no recombination, or because of recombination and re-coalescence back to the same discrete point. However, in the SMC′, we also need to account for the possibility of invisible recombinations. Because invisible recombinations are not represented in the tree, there could potentially be an invisible recombination on any of the branches. The probability of this is not trivial to compute; it requires summing across all branches and computing the probability of a recombination at any point along the branch which coalesces at any point farther up the same branch. This probability is dependent on where the new branch coalesces, so has to be computed for every possible state.

With the exception of the "invisible recombination" addition above, the transition probabilities are all identical to those in the SMC model, except with adjustments to the input lineage/node counts that are input. However, the fact that the lineage counts depend on the previous state make optimization of these calculations more complicated. For example, in the original SMC algorithm, the cumulative coalescence rate was stored in a variable $C_m = \sum_{j=0}^{m} \frac{l_j \Delta t_j}{2N_j}$. Then, the coalescence rates between times k and b would be given by $C_b - C_{k-1}$. Under the SMC′, we need to store a second version of the cumulative coalescence rates, in which an extra lineage exists: $C'_m = \sum_{j=0}^{m} \frac{(l_j+1)\Delta t_j}{2N_j}$. Then, in this example where the extra lineage exists until time a, the coalescence rate between time k and b would be given by $C_b - C_{a-1} + C'_{a-1} - C'_{k-1}$. In the same way, the probability of coalescence at a particular time interval was previously stored in a single vector; under the SMC′, three versions of this vector are required: one in which no broken lineage exists, one where it coalesces and forms a node, and one where it exists but does not coalesce. The derivations and optimized formulas are described in Appendix A.

1.5.3 SMC′ demonstration

In this section, I present a few simulation results demonstrating that the SMC′ implementation seems to work at least as well, and sometimes better, than the original SMC model.

First I present a simple demonstration that key parameters are estimated as accurately with the SMC′ as with the SMC. In the following figure, I generated 100 data sets of length 2Mb and 8 haploid samples with the program `ms` [15]. I used a constant diploid population size of 10000, mutation rate of 2.5e-8/gen/bp, and recombination rate of 1.5e-8/gen/bp.

I then ran ARGweaver on each data set 6 times, varying the number of haploid samples (2, 4, or all 8) and the model used (SMC or SMC′). All other ARGweaver parameters were set to the default. Then, I extract various statistics from the ARG., such as the time to most recent common ancestor (TMRCA), total branch length (branchlen), or average distance between leaf nodes (pi). In Figure 1.7, I show the distribution of the difference between the median of the estimated statistics, and the true statistics. There is no discernable difference between this distribution computed under the SMC′, vs under the SMC.

To look a bit deeper into the accuracy of statistics, I then looked at the distribution of statistics predicted by ARGweaver as a function of the true statistic. Figure 1.8 shows that there is no difference in the accuracy for any quantile of the true TMRCA between the SMC and SMC′. For higher quantiles, there is a tendency for ARGweaver to underestimate the TMRCA. This is likely due to the discretization, and the fact that there is a maximum time set in ARGweaver, so that any true TMRCA past this time will be necessarily underestimated.

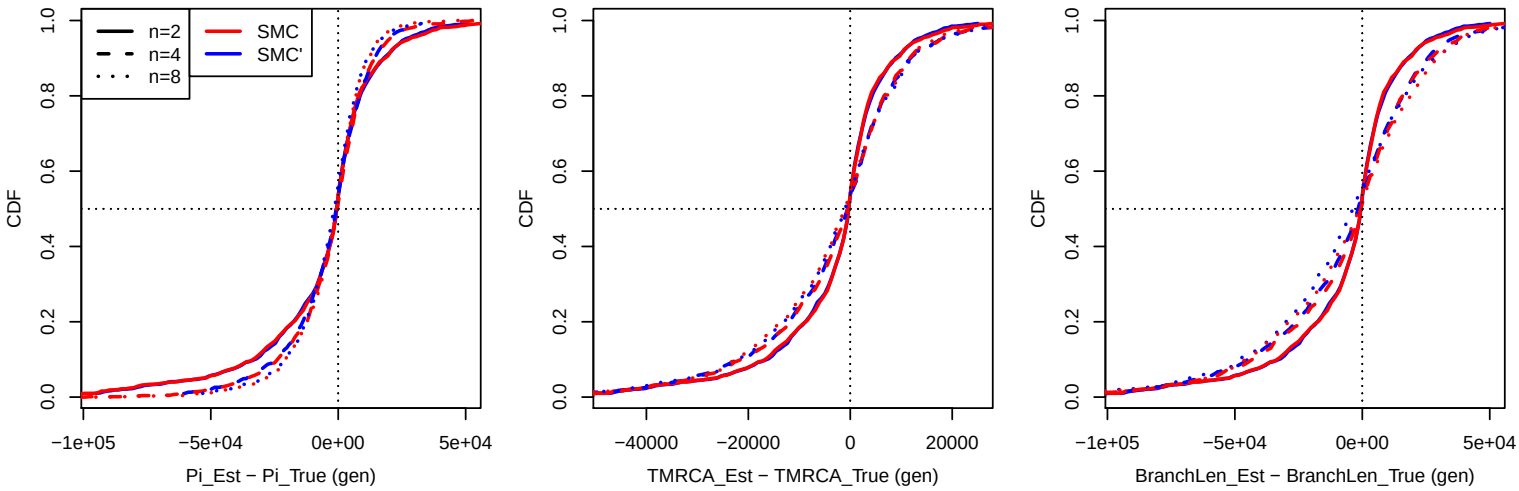

Figure 1.7: **Tree statistic accuracy SMC vs SMC′**. These plots show the difference between the median estimated statistic (Pi, TMRCA, and total branch length) from the true statistics. The distributions are taken from sampling the difference in these statistics at every base across each simulated data set (total 2e8 bases).

Next, I looked at estimated recombination rates across these regions. Figure 1.9 shows the estimated rate in each set of inferred ARGs, compared to the true value of 1.5e-8 events/bp/generation. It also shows the rate of invisible recombinations inferred, which is higher for smaller sample sizes and old branches, as expected. It does appear that the SMC′ is more accurate in estimating recombination rate. It is not clear why there is such an over-estimation in the first time interval using the SMC. This may not have to do with the SMC model, but with improvements in how the rounding of recombination times is handled in the SMC′.

Finally, I looked at the joint distribution of coalescence times between sites separated by various distances, under $n = 2$. This distribution captures the main difference in the two models, and a previously published analysis showed that

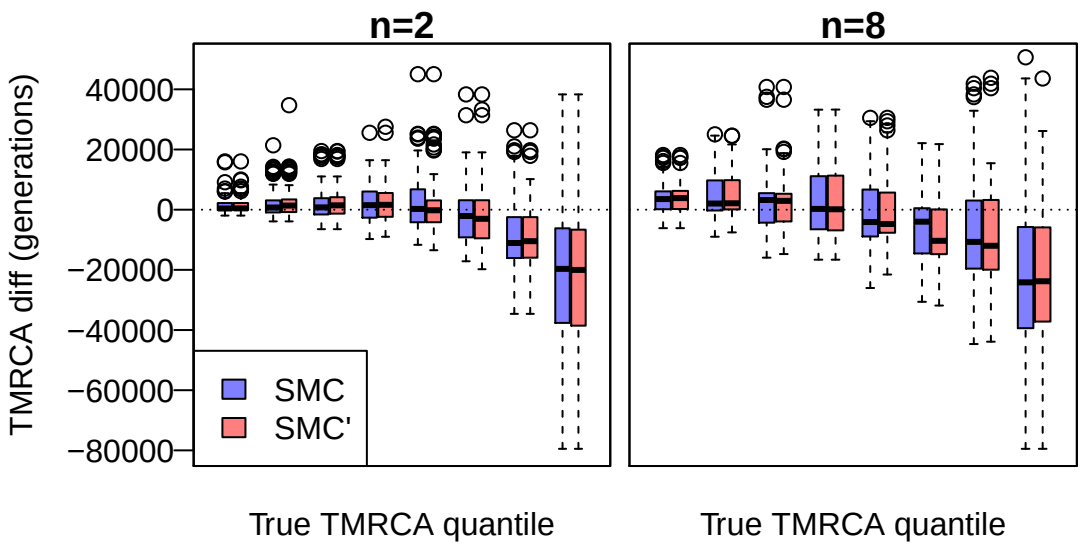

Figure 1.8: **Parameter accuracy by value**. The true TMRCA distribution, taken as the TMRCA at every base along the 2e8 bases in the statistic, was divided into 8 quantiles. This plot shows the difference in the estimated vs true TMRCA for each quantile, for SMC and SMC'

this distribution matches the full coalescent-with-recombination (CwR) much better under the SMC' than the SMC [34]. That was just done for ARGs simulated with continuous-time models. Here, we compare whether this is true for ARGs generated by ARGweaver under the two models. The results are shown in Figure 1.10. This plot is generated with a different data set than the previous ones. Here, I have run ARGweaver with a masked data set, so that ARGs are generated from the prior distribution. With $n = 2$, each threading operation removes one lineage and replaces it, so that the ARGs generated at each iteration can be thought of as independent samples; from each sample I record the TMRCA for two sites at the distances of interest (100bp, 1kb, 10kb, 100kb).

Figure 1.10 does show that there seems to be a better match between the SMC' and CwR than between the SMC an CwR. However, the difference be-

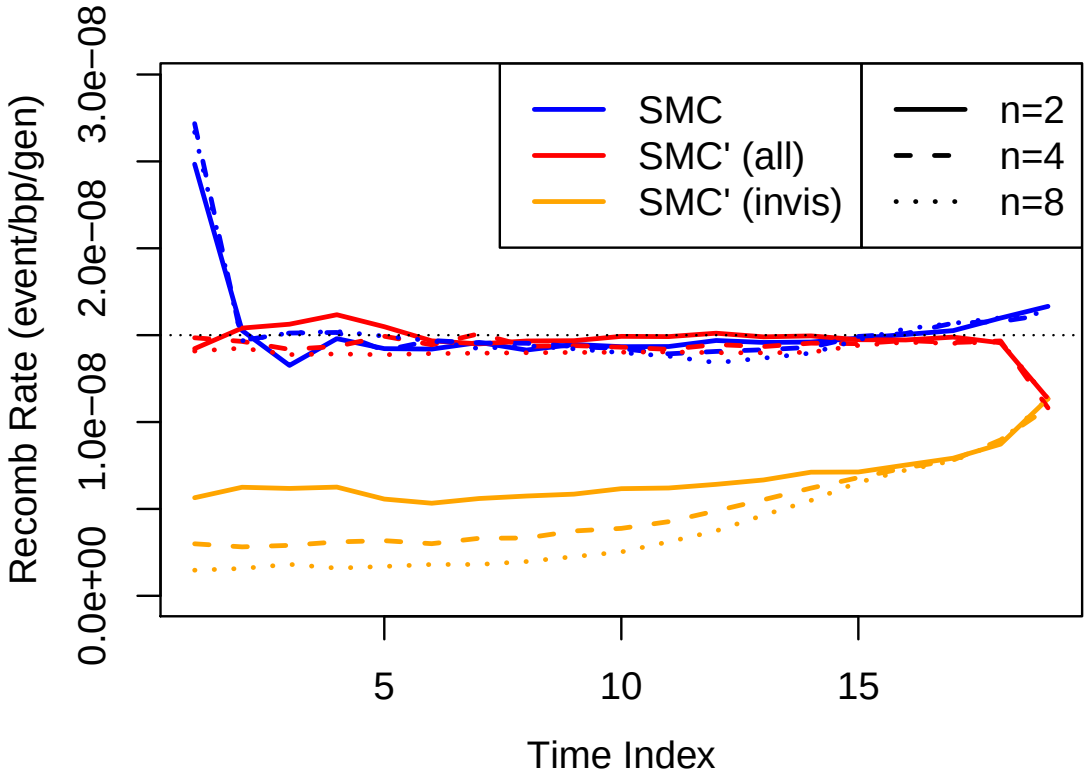

Figure 1.9: **Estimated recombination rate in the SMC vs SMC′**. The true value (black dotted line) is 1.5e-8. The x-axis (Time index) refers to the ARGWeaver discretized time index, with the lowest values being the most recent time. The recombination rate was estimated as number of inferred events in each discrete time interval, divided by the total branch length in each time interval. The orange line shows only invisible recombination events, inferred under the SMC′, whereas the red line shows the total recombination rate.

tween the two plots is fairly subtle. Overall, I conclude that many other approximations of the ARGweaver model also contribute to differences with the CwR, especially discretization of time, and ways for spreading probability mass for coalescence and recombination onto zero-length nodes.

Lastly, I examine at the run-time of the SMC vs SMC′. For this, I performed additional simulations with $n = 16$, $n = 32$, and $n = 64$, and compare the runtime

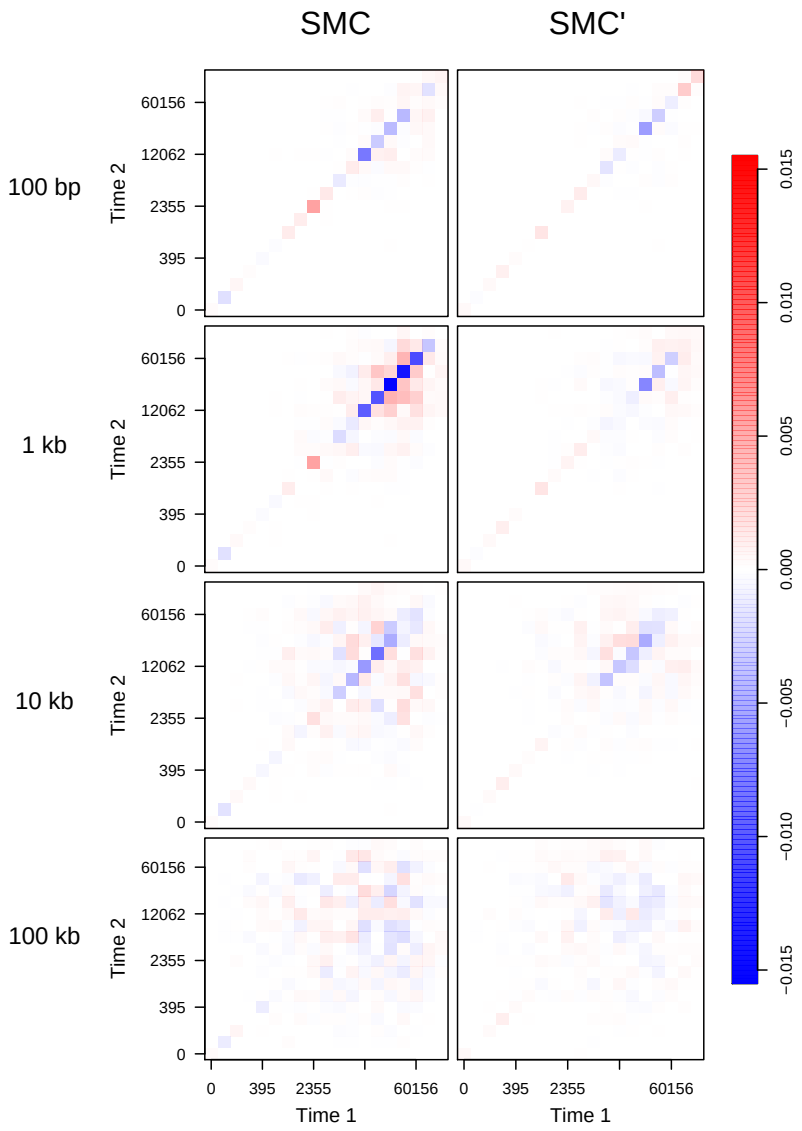

Figure 1.10: **Difference in joint distribution of TMRCA between the SMC/SMC′ and the CwR.** The coalescence time of one site is shown on the x-axis, the coalescence of another site spaced some distance apart (shown in the left column) is shown on the right axis. The color represents the difference in the fraction of sites falling in this bin between one model (SMC on left and SMC′ on right) and the CwR.

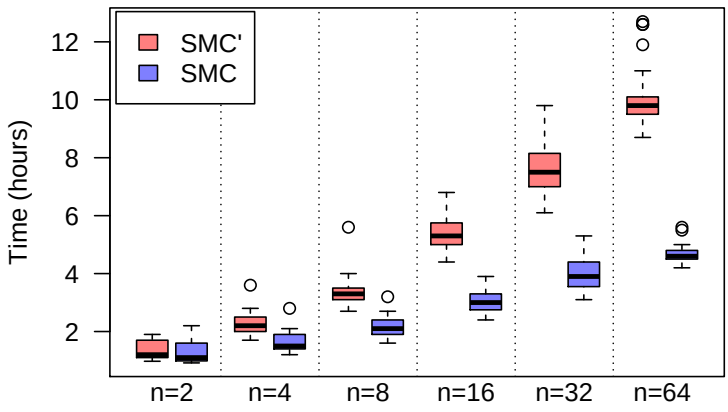

Figure 1.11: **Run-time of SMC vs SMC′. The boxplot shows distributions of run-time over 100 simulation replicates, where arg-weaver was run on each replicate for 1000 iterations.**

in hours using the two models. The results are shown in Figure 1.11. Despite my efforts to optimize the SMC′ equations, I was not able to simplify the equation for the probability of no invisible recombinations, creating a bottle-neck in the run-time and increasingly worse performance as n increases. If I were to re-do this implementation, I would choose to keep invisible recombinations in the ARG structure, so that I would not have to integrate over their presence. However, given that overall the SMC′ does not significantly improve inference, it does not seem like a worthwhile endeavour.

To summarize, I implemented the SMC′ in ARGweaver, but it is not clear that it makes a significant difference to the quality of inference, and for large n it is much slower. Therefore, ARGweaver continues to use the SMC by default; the SMC′ can be used with the option `--smc-prime`.

However, the effort spent implementing the SMC′ is not entirely a loss. First, the SMC′ was a requested feature of ARGweaver, and is now available to anyone who wants to try it out and compare results with those from the SMC. Ad-

ditionally, there turn out to be some parallels between the SMC' and the multi-population model, which will be the focus of Chapter 3. As will be described there, the multi-population model also requires knowledge of which state was chosen at the previous site to compute lineage counts, as different states will add a lineage to different populations. It also turns out that the multi-population model makes more sense under the SMC' than the SMC, because some recombination events may change populations but not change the tree topology. Therefore, the SMC' model will get extensive use in Chapter 3.

1.6 Conclusion

The first half of this chapter was an introduction to ARGweaver, written by M. Hubisz and A. Siepel as the first half of a book chapter on using ARGweaver, which will be published in the book "Statistical Population Genomics", edited by J. Dutheil and published open-source by Springer. After that, I described and demonstrated several important features that have been added to ARGweaver over the course of my dissertation and used in various studies. These include: integration over phase, ancient sample ages, low-quality genomes, reading VCF files, and implementation of the SMC'. An additional feature, sampling from a demographic model with population divergences and migrations, will be presented as part of Chapter 3.

CHAPTER 2

AN EARLY ADMIXTURE EVENT BETWEEN ANCIENT HUMANS AND NEANDERTHALS

Note: With the exception of the introduction and conclusion (which are not published elsewhere), this chapter was adapted from text published in [2]. While I was not a first author on that paper, all the text and figures used here are taken from my contributions to that paper and its Supplementary Material.

2.1 Introduction

When I first started studying human evolution, the chimpanzee was regarded as the closest relative to human that could potentially be used for genetic comparison. I was even involved in some of the early studies comparing the newly sequenced human and chimp genomes, sometimes with an outgroup, and looking for signals of adaptation [35, 36]. But this approach gives us insight into adaption along a lineage which is estimated anywhere between 6-12Mya in length [37]. The sequencing of the Neanderthal genome [16, 38, 39], and later the Denisovan genome [17], has enabled us to zoom in on the past half million years of hominin history, revolutionizing the study of human evolution and leading to many new insights about ancient hominin history.

The most notable discovery was that humans and Neanderthals interbred, and that the hybrids were healthy enough that thir genes persist in the genomes of humans today, so that all non-African humans contain $\sim$ 2% Neanderthal DNA [16,39,40]. This was followed by the discovery of the Denisovans, and the

realization that they too left an even higher amount of DNA (4-6%) in modern Oceanian humans [41]. There has been much speculation about the effect of these introgressed genome segments on humans. Because Neanderthals adapted to the cold Eurasian climate several hundred thousand years earlier than humans, it seems possible that the interbreeding may have had adaptive benefits for humans, allowing us to inherit gene variants that helped us survive the harsh climate. However, it has also been argued [42, 43] that the smaller population size of Neanderthals means that they carried a higher genetic load, and that interbreeding therefore contributed unhealthy variants to the human gene pool. Furthermore, it is also possible that some hybrid incompatibilities had arisen between humans and Neanderthals, so that gene variants that are healthy on a Neanderthal background may be deleterious in a human, or vice versa. Likely, all of these possibilities are true to some extent. For example, a recent study [44] showed evidence of adaptive introgression in genes that interact with viruses, and there is also the striking example of Tibetans inheriting a variant of EPAS1 from Denisovans which helps them survive high altitudes. On the other hand, there are also signals that some archaic DNA was deleterious, including an almost complete lack of introgression on the X chromosome, as well as several autosomal "deserts of introgression" [40, 45]. While it has been claimed that introgressed Neanderthal and Denisovan DNA is depleted near genes [40, 45], this has lately been called into question [46, 47], but there does remain some signal of depletion near regulatory elements [47]. Understanding the effects of this interbreeding is made difficult by the lack of archaic hominin data, and the many biases in our ability to detect introgression, as well as the complexity of ancient hominin interbreeding events.

It turns out that ARGweaver is very well-suited to studying the history of

archaic hominins. First, there are very limited numbers of archaic samples available (at the time of this study, there was only 1 each of high-quality Neanderthal and Denisovan genomes). While ARGweaver is a computationally expensive tool to use, it makes full use of this very limited data set, taking full haplotype patterns into account with no reduction to summary statistics. There are high levels of incomplete lineage sorting (ILS) between Neanderthals, Denisovans, and humans, so that the local trees produced by ARGweaver are a natural way to model their relationships, and to examine coalescence times to try to tease apart introgression from ILS.

The study in this chapter was part of a collaboration with the other authors in [2]. Before I was involved, they had already devised a hypothesis that there may have been some introgression of human DNA in the Neanderthal genome (due to evidence I describe below). This was a surprising idea, since the only available Neanderthal genome at the time (the Altai Neanderthal) is estimated to be far older ($\sim$ 115 kya) than human's out-of-Africa migration, which occurred roughly 50kya. The idea was that an ancient group of humans left Africa earlier than 115kya and encountered Neanderthals, possibly in the Middle East, and interbred with them, leaving some of their genes in the Neanderthal gene pool. This group of humans then died out, or perhaps was subsumed into the Neanderthal population, so that they are not direct ancestors of any current-day human.

It was already known that Neanderthals share more alleles with African humans than Denisovans, but this was attributed to a hypothesized introgression from an unknown super-archaic hominin (possibly *Homo erectus*) into the Denisovan (referred to as Sup$\rightarrow$Den). This event is supported by especially high

D statistics (which measure the increase in Neanderthal/African allele sharing compared to Denisovan/African sharing) at variant sites that have reached a high-frequency, or have fixed, in the African population [16]. The presence of this event makes it more difficult to make a case for Hum→Nea introgression, as both cause a skew in allele sharing in the same direction.

The initial evidence for a Hum→Nea event came observations about divergence and heterozygosity measured in 100kb genomic windows. It turns out that the windows with the lowest divergence between Neanderthals and African humans have higher Neanderthal-Denisovan divergence than expected, as well as higher Neanderthal heterozygosity levels [2]. This cannot be explained by Sup→Den, but would be expected under a Hum→Nea event. By contrast, the regions with the highest African-Denisovan divergence also have high Neanderthal-Denisovan divergence, and high Denisovan heterozygosity, which can be explained by Sup→Den, but not Hum→Nea. This analysis suggest that both events may be true, and was backed up by a simulation study.

The group also undertook an analysis using G-PhOCS [10], to try to infer a demographic model, including population size changes and migration events. G-PhoCS is a Bayesian method which examines many short, neutral, unlinked regions of aligned genomes, and samples local trees for each region, fitting these trees to a demographic model. The analysis did find support for both Sup→Den and Hum→Nea migrations.

However, it would be desirable to have another line of evidence which looked at all the available data, and modeled the relationship between humans, Neanderthals, and Denisovans. This was the motivation for the study described below. The main idea was to run ARGweaver on Neanderthal, Densiovan, and

African human data, and identify regions where Neanderthal and Africans coalesce more recently than the Neanderthal/African split: these are candidate Hum→Nea regions. Regions where the inferred local trees have unusually high coalesence times for the Denisovan are candidate Sup→Den regions. The local trees produced by ARGweaver make the difference between Sup→Den and Hum→Nea very clear, whereas other allele sharing statistics do not. In the analysis described below, I show how this approach helped make a convincing case for the Hum→Nea event.

2.2 Results and Discussion

2.2.1 Excess of young 'African' haplotypes in Neanderthal genome

We ran ARGweaver on all the autosomal chromosomes, with a data set that included the Neanderthal, Denisovan, 4 Africans, and chimpanzee (haploid) genomes. We then scanned the resulting ARGs for contiguous regions (≥ 50kb) where one archaic individual coalesces with Africans before it coalesces with the other archaic individual. We refer to these segments as 'African' haplotypes, and we recorded the archaic/African coalescence time for these segments. More details about the ARGweaver runs and the 'African' haplotype criteria are in Section 2.3.2. Figure 2.1a shows the distribution of these segments identified in both Neanderthal and Denisovan, for each of the discrete times in the ARGweaver model. There is clearly an excess of young 'African' haplotypes in the Neanderthal.

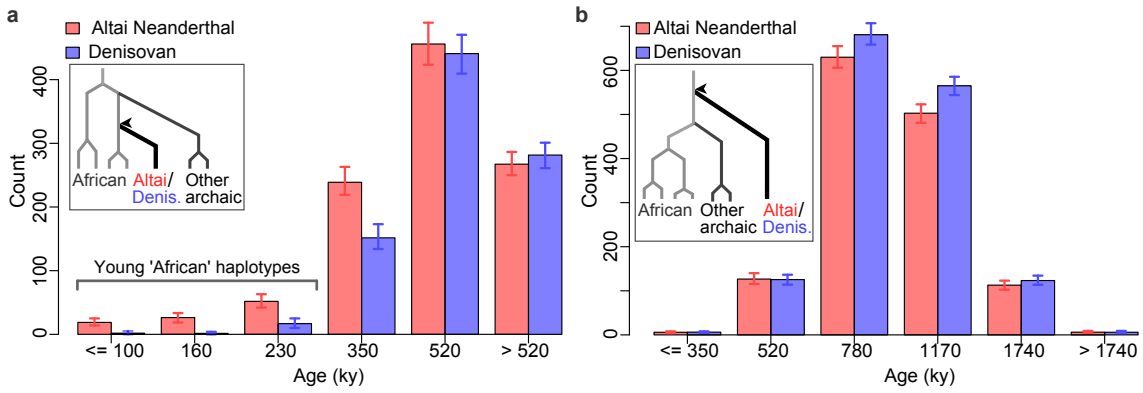

Figure 2.1: **Distinguishing between two scenarios of introgression into archaic humans (From [2])** a, The age distribution of 'African' haplotypes (≥ 50 kb) in the Altai Neanderthal and the Denisovan genomes as inferred by ARGweaver. Error bars represent the 95% credible intervals from 302 Markov chain Monte Carlo (MCMC) replicates. An 'African' haplotype coalesces within the African subtree before coalescing with the other archaic individual (inset), and its age is inferred as that coalescent time (arrowhead). The majority of the young 'African' haplotypes in the Altai Neanderthal genome are estimated to coalesce 100,000-230,000 years ago, with just a few estimated to coalesce less than 100,000 years ago. b, The age distribution of 'deep ancestral' haplotypes (≥ 50 kb) in the Altai Neanderthal and Denisovan genomes. A 'deep ancestral' haplotype coalesces above the African subtree and the other archaic lineage (inset), and its age is inferred as that coalescent time (arrowhead). ky, thousand years.

We also identified regions where the local tree indicates one of the archaic individuals is an outgroup to all other samples: we refer to these as 'deep ancestral' haplotypes. The distribution of these and their ages are shown in Figure 2.1b. We observe a slight excess of these regions in Denisovan compared to Neanderthal for the time intervals t = 780ky and t = 1170ky.

Potentially introgressed segments in the Altai Neanderthal We expect most of the inferred 'African' segments to be a result of incomplete lineage sorting and

not necessarily the result of introgression. We thus wanted to choose a set of 'African' segments in the Altai Neanderthal that are strong candidates for being introgressed from modern humans. We chose a length cutoff of 50Kb because we expect introgressed segments to be long, relative to older haplotypes resulting from incomplete lineage sorting. In addition, long haplotypes harbor more mutations, giving ARGweaver more power to accurately date coalescence events, so this length cutoff also filters out less informative regions. However, none of our results changed substantially when varying the length cutoff from 20kb up to 100kb (affecting the overall but not relative counts between the Altai Neanderthal and Denisovan). Looking at Figure 2.1a, which uses a length cutoff of $\geq$ 50kb and shows the distribution of haplotype ages for 'African' haplotypes, the age cutoff $\leq$ 230ky was chosen to classify potentially introgressed segments in Altai Neanderthal, as there are few 'African' segments in the Denisovan genome meeting this criteria. There are an average of 97 (95% CI: 86-108) 'African' segments in the Altai Neanderthal genome meeting this length and age criteria, covering 7.2Mb (95% CI: 6.5-7.9 Mb). Conversely, in the Denisovan genome, there are an average of 20 'African' segments (95% CI: 13-28) covering 1.3Mb (95% CI: 0.9-1.9Mb). Some of the 'African' segments in the Altai Neanderthal that are older than 230ky may also be due to the proposed introgression event, however, given the high levels of 'African' haplotypes in the Denisovan older than 230ky, it seems likely that most of these segments are better explained by incomplete lineage sorting. This observation is also confirmed by simulations (see Simulation study section below).

African source population of potentially introgressed segments Looking at the set of potentially introgressed segments in the Altai Neanderthal genome, we examined whether these segments primarily coalesce within a particular African

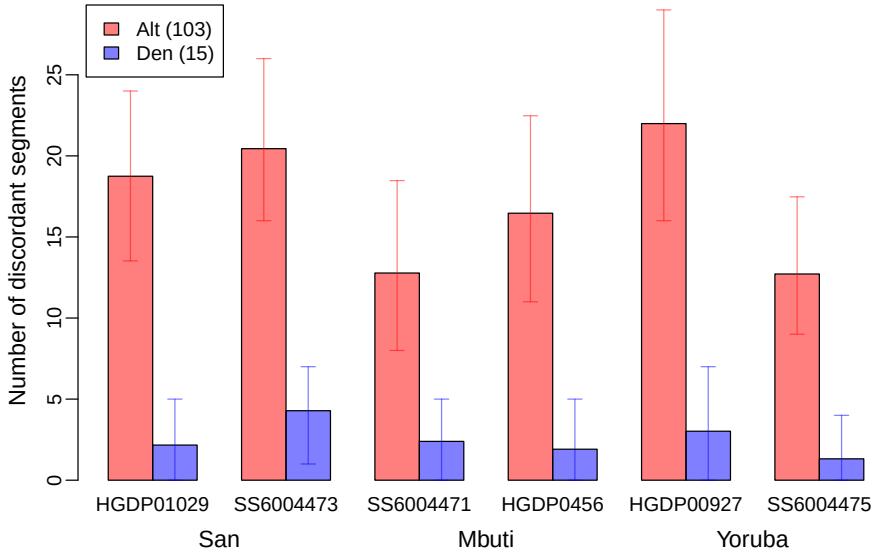

Figure 2.2: Number of potentially introgressed segments (≥ 50kb with ages ≤ 230ky) in the Altai Neanderthal and Denisovan genomes based on their coalescence times with each African individual. Red=Altai, blue=Denisova, bar height represents means, and error bars give 95% confidence interval.

population. In Figure 2.2, we show the number of segments in the Altai Neanderthal defined by their coalescence time with each African individual. There appear to be somewhat fewer 'African' segments from the individuals from the Mbuti population; however this difference is not statistically significant, and the data support a model in which the three African populations contribute equally to the introgression event.

Overlap of segments with 'African' haplotypes and ancestral segments As shown in Figure 2.1b, there is an excess of ancestral segments in the Denisovan, which is presumably due to introgression of an unknown archaic population into the Denisovan genome [16]. This archaic introgression complicates interpretation

of the ARGs. Both scenarios of introgression – super-archaic introgression into Denisovan lineage and modern human introgression into Altai Neanderthal lineage – are likely to lead to an excess of Denisovan ancestral lineages, as well as an excess of 'African' Altai Neanderthal lineages. However, the observed excess of 'African' haplotypes in the Altai Neanderthal at $\leq$ 230ky is not expected from super-archaic introgression into the Denisovan lineage alone. Nevertheless, we wanted to check that this signal is not an artifact due to the excess ancestral segments in the Denisovan genome.

To this end, we created a version of Figure 2.1a, which shows the excess of young segments in the Altai Neanderthal genome with 'African' haplotypes. In this version, we removed all 'African' segments which overlap any ancestral segment in either lineage of the other archaic individual. The ancestral segments used for this purpose were not filtered for informativeness or length, in order to use the most complete (rather than confident) set of ancestral segments. The results are shown in Figure 2.3. 67% of the Altai Neanderthal segments with 'African' haplotypes were removed, and 62% of the Denisovan segments were removed; however the excess of young Altai Neanderthal segments with 'African' haplotypes is still statistically significant.

Argweaver analysis of the two European Neanderthals on chromosome 21
ARGweaver was also run on chromosome 21 with the addition of data from two additional Neanderthals (El Sidrón and Vindija), for which targeted sequencing was used on chromosome 21. We then called 'African' haplotypes as described above. Having three Neanderthals in the analysis makes it less likely that a Neanderthal lineage will coalesce into the human subtree before coalescing with another Neanderthal, under a model with no modern human intro-

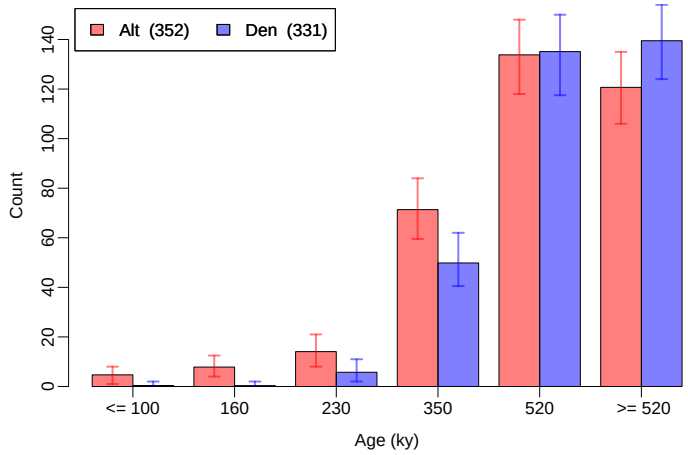

Figure 2.3: Distribution of ages in the Altai Neanderthal and Denisovan genomes for segments (≥ 50kb) with 'African' haplotypes, after removing any 'African' segment which overlaps segments where the other archaic individual is an outgroup to all other archaic and present-day humans. Numbers in the legend give total numbers of segments. Bar height represents means across all ARGweaver runs, and error bars give 95% confidence intervals.

gression. Therefore, we expect fewer 'African' haplotypes in any of the three Neanderthals compared to the Denisovan. Still, the analysis yielded one region on chromosome 21 which was called as a long, young homozygous 'African' haplotype in Altai Neanderthal (chr21:30,368,000-30,458,000; hg19 coordinates). No such region was found for the El Sidrón, Vindija or Denisovan chromosome 21 (Figure 2.4).

Simulation study to address the ages of the archaic individuals One limitation of ARGweaver is that, at the time of this study, it did not have an option to handle the age of archaic individuals. For this reason, all individuals in each data set were treated as if they were present-day individuals, but we were concerned

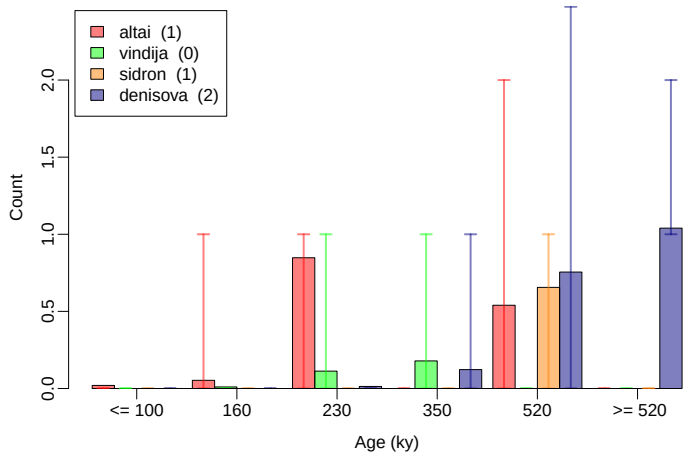

Figure 2.4: Number of 'African' haplotypes found in each archaic individual in the ARGweaver analysis of their chromosome 21.

that this approach could potentially lead to a bias in the coalescence times inferred by ARGweaver. If the archaic individuals were all of the same age, we would expect this bias to be the same for the Altai Neanderthal and Denisovan. However, since the Altai Neanderthal is likely older than the Denisovan, there is an additional concern that this could lead to a larger bias in the Altai Neanderthal compared to the Denisovan, and this differential bias could be falsely interpreted as a signal of introgression.

In order to explore the effects of this model misspecification, we conducted a simulation study. We generated data sets using realistic demographic parameters, including ancient sampling ages for the archaic individuals, but without any migration between populations (see Section 2.3.3 for details).

ARGweaver was run on each simulated data set as in the real data analysis: sample ages were ignored, haplotype phases were randomized for each individ-

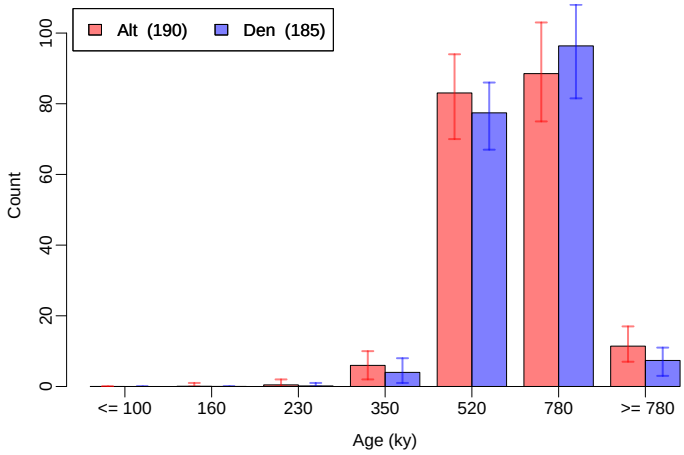

Figure 2.5: Distribution of the ages of segments with 'African' haplotypes in the simulated data set, which was generated with no recent modern human admixture, but with sampling times of 70ky for the Altai Neanderthal and 50ky for the Denisovan. Bar height represents mean, and error bars give 95% confidence intervals.

ual, and the phase integration feature was used. Then, 'African' segments in the Altai Neanderthal and Denisovan were identified, and the distribution of their ages compared. Unlike in the real data analysis, there was no significant excess of young segments with 'African' haplotypes in the Altai Neanderthal genome (Figure 2.5). Therefore, the excess observed in the real data analysis appears not to be an artifact due to the difference in the ages of the archaic individuals.

Effect of phase integration We analyzed the simulated data sets in two ways: once with the true haplotype phase treated as known, and once with randomized phase and phase integration. All results presented here are from the analysis with phase integration, as this is how the real data was analyzed. Similar figures produced from the runs with known phase look extremely similar in shape (not shown). However, the absolute numbers of long 'African' haplotypes were

40% lower in the runs with phase integration. This appears to be largely a result of long haplotypes being broken up by phase errors. On a basewise level, the performance of the two runs was more similar: 'African' segments were identified with a true positive rate of 77.7% and a false positive rate of 3.9% when the true phase was used, compared to a true positive of 74.2% and false positive of 4.7% with phase integration. Overall, we expect that, if phase were known in the real data, our analysis would have yielded more long 'African' haplotypes, but phase integration does not seem to have impacted the ages or relative counts of these segments.

Simulations with archaic introgression into the Denisovan We conducted an additional simulation study to explore the effects of introgression into the Denisovan from an unknown archaic hominin. The simulation parameters were similar to the above section, but with an additional population simulated with 1Mya divergence from the other hominins, and 10% migration rate into the Denisovan 300kya. Further details are in Section 2.3.4.

Applying the same ARGweaver analysis to this data set, we obtain an age distribution of 'African' haplotypes shown in Figure 2.6; the distribution for deep ancestral haplotypes is in Figure 2.7. In this case, the excess of ancestral segments in Denisovan compared to Altai Neanderthal is much higher than observed in the real data, suggesting that the simulations contain an exaggerated amount of super-archaic introgression into the Denisovan. As expected, this introgression does cause the Altai Neanderthal to have more 'African' haplotypes than the Denisovan. However, it does not cause any skew in the ages of these haplotypes compared to those shown in the simulations without archaic introgression. Notably, there are no 'African' haplotypes in the Altai Neanderthal

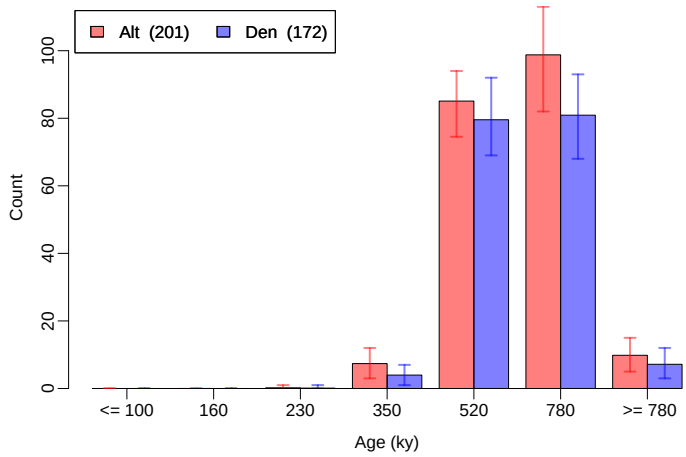

Figure 2.6: Distribution of 'African' haplotype ages in data simulated with introgression into the Denisovan from an unknown archaic hominin.

with ages ≤ 350ky, and only a small excess in the 350kya category compared to the Denisovan. Therefore, it is unlikely that the youngest 'African' haplotypes identified for in the Altai Neanderthal genome could be explained by super-archaic introgression into the Denisovan.

Simulations with modern human introgression into the Altai Neanderthal We performed a final simulation study to test the power of the ARGweaver approach to detect the type of introgression event proposed in this manuscript. These simulations included migration from a modern human population into the Altai Neanderthal lineage 100kya at a rate of 3.55% for a single generation, and also included Sup→Den introgression at a more modest rate than the previous section. Full details are given in Section 2.3.5. We ran ARGweaver on this data set, and produced 'African' haplotypes ≥ 50kb; these are shown in Figure 2.8. In the top panel, the distribution of African haplotype ages is shown for the Altai Neanderthal and Denisovan. These simulations show an excess of young

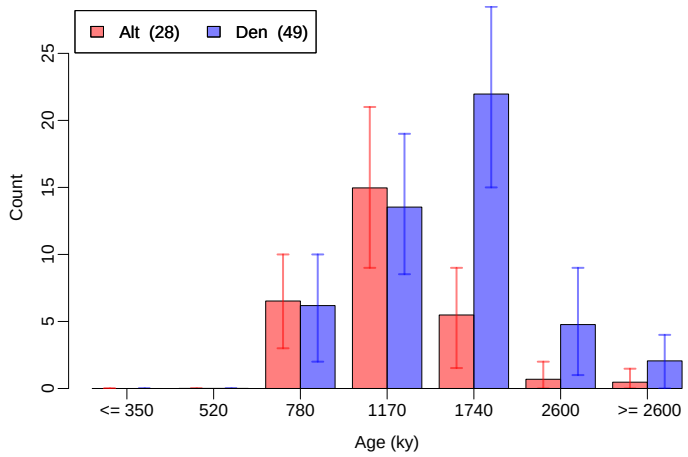

Figure 2.7: Distribution of ancestral segment ages in data simulated with introgression into the Denisovan from an unknown archaic hominin (super-archaic introgression).

African haplotypes dated ≤ 234kya, as well as an absence of such haplotypes in the Denisovan, which is quite similar to the results from the real data. Therefore, a migration event similar to the one simulated can produce a signal much like the one we observe, and our ARGweaver analysis has the power to detect this signal.

These simulations also allowed us to compare the haplotype ages computed by ARGweaver to the true haplotype ages available in the trees produced by `ms`. For each 'African' haplotype predicted by ARGweaver, we computed its true age as the average time (from present) to the first coalescence between an African lineage and a target lineage (either the Altai Neanderthal or Denisovan) across the predicted region. In Figure 2.8b, we show the distribution of true ages for each set of 'African' haplotypes sharing a particular estimated age from ARGweaver. The figure is divided into true positives, false positives, true neg-

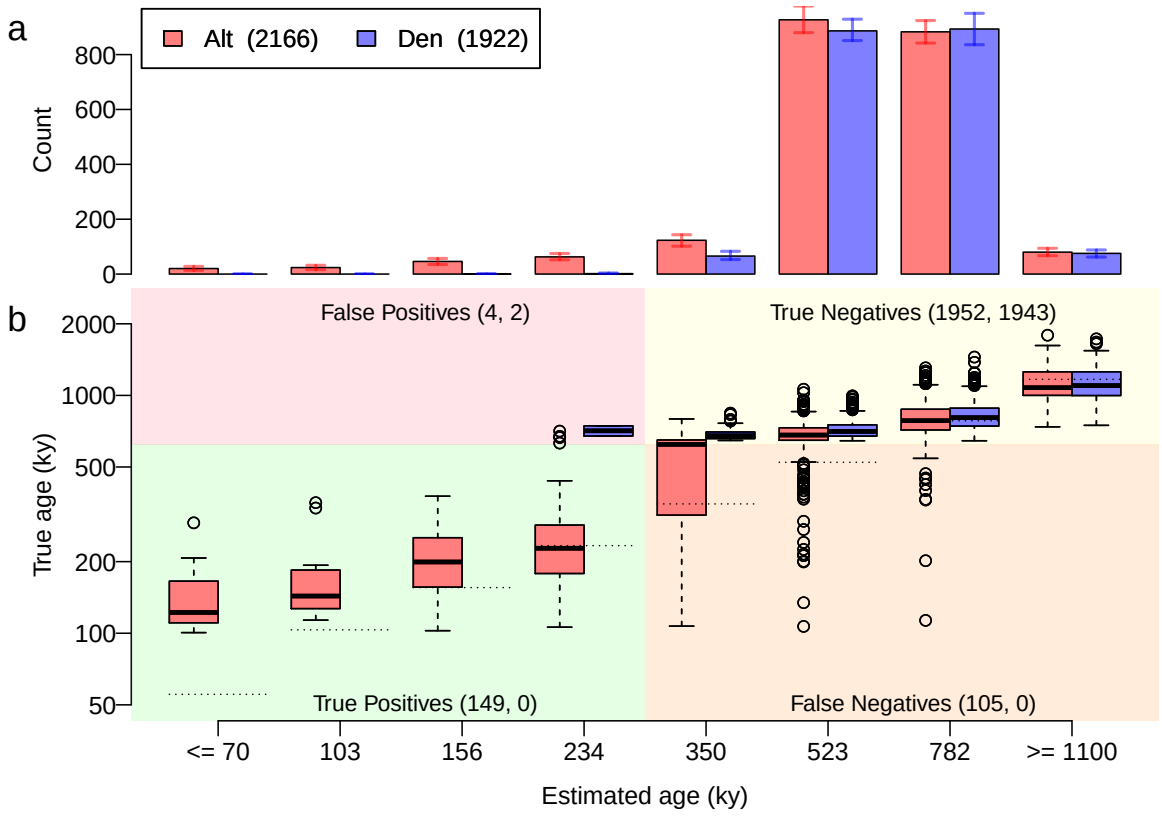

Figure 2.8: a. Distribution of African haplotype ages in sequences simulated with introgression into the Altai Neanderthal lineage from modern humans 100,000 years ago. 'African' haplotypes are identified as in Figure 2.1. Error bars represent the 95% Bayesian credible intervals from 302 MCMC replicates. b. Distribution of true haplotype ages for each of the estimated ages. The horizontal dotted lines show the estimated age. The plot is divided into four quadrants; the lower half represents 'African' haplotypes having true ages between 100,000 and 620,000 years ago (the divergence time between archaic and present-day humans), which are necessarily due to post-divergence gene flow from modern humans. The left side of the plot represents regions that would be identified as introgressed based on a threshold of≤234,000 years. The counts in each quadrant are for Altai Neanderthal (red) and Denisovan (blue), respectively. The counts for the Denisovan in the lower two quadrants are zero because there was no simulated migration from modern humans into the Denisovan lineage. Note that this is a somewhat nonstandard plot of true age versus estimated age; a more standard, reversed view is given in Figure 2.9 and demonstrates that the estimated ages are largely unbiased. Error bars as in the standard Tukey box plot (R boxplot function).

atives, and false negatives based upon ARGweavers ability to identify these introgressed haplotypes resulting from the modern human gene flow into the ancestors of the Altai Neanderthal, using an identification threshold of $\leq$ 234kya. Note that this figure considers only 'African' haplotypes which were identified by ARGweaver and which pass the length threshold of 50 kb; there are many more false and true negatives which are not considered here. Overall, these simulations reinforce our choice of 234kya as a threshold for choosing a confident set of potential introgressed regions in the Altai Neanderthal genome with a very low false positive rate. This choice is supported both by the distribution of true ages in each age bin for the Altai Neanderthal, as well as by the contrast between the Altai Neanderthal and Denisovan. A higher threshold would identify more truly introgressed regions, but would disproportionately increase the false positive rate.

It is apparent from Figure 2.8b that the times produced by ARGweaver behave reasonably, with the expected linear relationship between estimated and true times. However, the estimated times are quite noisy, and cannot be used to precisely date a particular haplotype. It also appears from this figure that the ARGweaver age estimates may be biased downward, but this is actually an artifact of our simulation settings and the distribution of true haplotype ages. For example, because the simulated migration into Altai Neanderthal occurred at 100kya, all haplotypes with age estimates less than 100kya are necessarily underestimated in this scenario. The same effect is seen for the Denisovan, which has no haplotypes with true ages younger than the archaic/modern human divergence time of 620kya. To confirm that the bias is an artifact, we show in Figure 2.9 a more standard view of the accuracy of the ARGweaver age estimates, with the estimated ages shown as a function of (binned) true ages. Figure 2.9

confirms that the estimates are largely unbiased, especially for the younger ages where ARGweavers discrete times are sampled densely. However, Figure 2.8b demonstrates that the youngest haplotypes predicted by ARGweaver are almost certainly underestimates, due to their noise; therefore the ages of the youngest haplotypes do not provide a lower bound on the date of introgression. In the simulation, 13.2% of haplotypes with ages estimated $\leq$ 234kya have dates earlier than the true introgression event. In the real data, only 8.6% of haplotypes are dated more recently than 100kya, suggesting that the true introgression was likely older than 100kya.

2.3 Methods

2.3.1 ARGweaver settings

Data used ARGweaver was run genome-wide using a data set consisting of the Altai Neanderthal and Denisovan, six present-day humans, and chimpanzee (panTro4). The six present-day African humans included two each of Yoruban (HGDP00927, SS6004475), Mbuti (SS6004471, HGDP0456), and San (HGDP01029, SS6004473) individuals, all sequenced to high coverage [16]. For this study, we only examined autosomal chromosomes. We had additional sequence data for two Neanderthals (El Sidrón and Vindija) on chromosome 21 only; these were included in a separate analysis of chr21.

Genomic Filters Genomic filters Various genomic filters were applied to minimize the influence of sequencing and alignment errors. The following regions were masked in the analysis: 1) simple repeats identified by Tandem Repeats

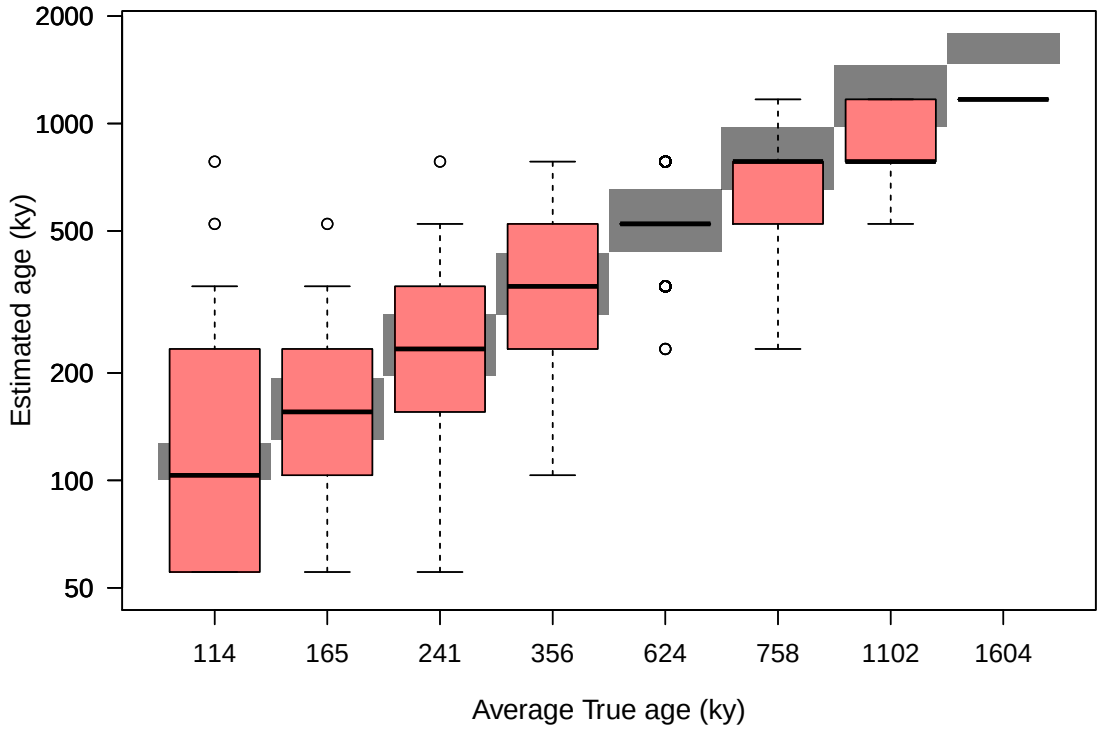

Figure 2.9: **Haplotype age accuracy.** Distribution of estimated 'African' haplotype ages in Altai Neanderthal genome as a function of true haplotype ages, in a simulation scenario with migration from modern humans into the Altai Neanderthal lineage 100kya. The gray boxes in the background show the range of true haplotype ages in each bin; the boxplot shows the distribution of estimated ages for each bin. Note that the boxplots represent distributions over a somewhat coarse collection of discretized times used by ARGweaver.

Finder (TRF) [48] (Simple Repeats track for GRCh37/hg19 downloaded from the UCSC Genome Browser); 2) recent segmental duplications in the human genome [49] (Segmental Dups tracks for GRCh37/hg19), 3) transposable elements identified by Repeat Masker (`http://www.repeatmasker.org`) with $\leq 20\%$ divergence from their consensus sequences, 4) regions with a mappability score in the Duke 20mer uniqueness score different from 1; 5) sites flagged as systematic errors [50]; and 6) regions not showing conserved synteny between human and chimpanzee (according to the UCSC syntenic net of the alignment between GRCh37/hg19 and panTro2).

Demographic Parameters ARGweaver requires prior distributions for the coalescence, mutation, and recombination rates, and these were chosen as in [1]. Specifically, a population size of 11,534 was used, and the recombination rate was based on the HapMap Phase II recombination map [51]. The average per-generation mutation rate was 1.26x10-8, with the rate in every 100kb segment scaled to reflect the observed substitution rate in that region between chimpanzee (panTro2), orangutan (ponAbe2), and macaque (rheMac2). Other ARGweaver parameters were a maximum time of 1,000,000 generations, 20 discrete time steps (distributed on a logarithmic scale, such that recent time intervals are shorter than more ancient ones, using the ARGweaver parameter $\delta = 0.01$) and 5,000 MCMC iterations. We used a site compression rate of 10 (`-c 10`), which decreases compute time 10-fold by combining groups of 10 sites into a single compressed site, and increasing mutation and recombination rates correspondingly. Site compression is implemented in a dynamic way, ensuring that variant sites are never compressed together, so that information is not lost. The genome was divided into 5Mb chunks with 1Mb overlap, in order to run the analysis in parallel across many processors. The ARGs for each chunk of a chro-

mosome were then pasted together at the midpoints of the overlaps between them before they were analyzed for signs of introgression. ARGs were sampled every 20th iteration starting with iteration 2000. ARGweaver did not take the ages of the archaic individuals into account, however we expected that the six present-day samples would dominate in determining the coalescence times. We explored the effect of this model misspecification using simulations (see the Simulation study section below).

Integration over phase Phase integration was used for all samples in the data (except the chimpanzee sequence, which was treated as a single haploid sample). The effects of using phase integration for this analysis were explored by simulation (described below).

2.3.2 Identifying 'African' and 'deep ancestral' haplotyes

'African' haplotypes Segments coalescing within the African subtree (2.1a) were identified for each haploid chromosome of each archaic individual, based on an ARG output by ARGweaver (representing a single MCMC sample from the posterior distribution of ARGs). The ARG defines a local tree at every position along the genome, with each tree having two leaf nodes per individual representing its two haploid chromosomes or lineages. The 'African' haplotypes for a particular archaic lineage (the target lineage) are determined by looking at the times to the most recent common ancestor (TMRCA) between the target lineage and other lineages in each local tree. Let T_{Afr} be the set of TMRCAs between the target lineage and all African lineages in the tree, and let T_{Anc} be the set of TMRCAs between the target lineage and all lineages coming from other archaic

individuals. Then, if $\min(T_{Afr}) < \min(T_{Anc})$ and $\min(T_{Afr}) < \max(T_{Afr})$, the local tree is considered discordant, and thus, may contain 'African' haplotypes. These two conditions guarantee that the archaic lineage is more closely related to at least one African lineage than to other archaic individuals, and that it falls within the range of African variation for the segment in question. Note that the other lineage from the target individual is not used to define TAnc, so that both heterozygous and homozygous 'African' haplotypes will be detected.

Once the sequence segments with 'African' haplotypes have been identified, the age for each segment is set to $\min(T_{Afr})$. Adjacent segments with 'African' haplotypes with the same age are combined into a single segment. Finally, a filter was applied which removed any segment in which the overall polymorphism level (across all individuals in the data set excluding the chimpanzee outgroup) was less than 1 polymorphic site per 1000 bases. This was done in order to remove segments with long stretches of masked sites.

Sequence segments coalescing beyond the African and archaic tree The ancestral segments shown in Figure 2.1b were defined using the same quantities defined in the previous section; in this case choosing regions for which $\min(T_{Afr})$, $\max(T_{Afr})$, $\min(T_{Anc})$, and $\max(T_{Anc})$ are all equal. This indicates that the target lineage is an outgroup to all African lineages as well as all lineages from other archaic individuals. The age of the ancestral segments was set to the TMRCA of this lineage with all other lineages, and adjacent ancestral segments with the same age were combined. The same filter for the polymorphism level in each segment was applied to these ancestral segments.

Averaging over MCMC replicates and the effect of homozygosity The ARGweaver analysis produced 151 sampled ARGs, as samples were taken every 20th

MCMC iteration from iteration 2000 to 5000. Each of these 151 ARGs produced a set of 'African' and ancestral haplotypes for each haploid lineage of each archaic individual (Altai Neanderthal and Denisovan). For a given archaic individual, then, there are 302 sets of segments with 'African' haplotypes as well as 302 sets of ancestral segments. Statistics presented here (such as counts or genomic coverage of these segments, or a selected subset of them) are calculated on each of the 302 replicates separately, and means and 95% confidence intervals across these values are reported.

Note that all 'African' and ancestral segments are defined for a single haploid lineage of an archaic individual, while ignoring the other lineage from that individual, so that segments are identified without regard to whether they are homozygous or heterozygous. This was done in order to fairly compare numbers between the Altai Neanderthal and the Denisovan, despite the higher level of homozygosity in the Altai Neanderthal. Homozygous segments are expected to be identified in both lineages, so that the effective number of replicates in homozygous regions will be closer to 151 rather than 302. This may result in somewhat more noise in our estimates for the Altai Neanderthal compared to the Denisovan, but it should be a minor effect, and importantly, there should be no impact on the expected values of our statistics due to an individuals level of homozygosity.

2.3.3 Simulations to assess the effects of ancient sample ages

We used ms [15], to simulate one hundred 2 Mb regions consisting of four Africans, one Altai Neanderthal, and one Denisovan individual. We used de-

mographic parameters consistent with estimates from G-PhoCS on our data [2]: an African population size of 24,000, an Altai Neanderthal population size of 750 from 70kya to 140kya, changing to 3,200 from 140kya to 450kya, a Denisovan population size of 2,500 from 50kya to 450 kya, the Altai Neanderthal and Denisovan populations coalesce at 450kya and have a population size of 8,000, and this population coalesces with Africa at 620kya, with an ancestral population size of 17,800. The Altai Neanderthal age was modeled at 70kya, and the Denisovan at 50kya. In order to do the simulations with ms, we set the Altai Neanderthal and Denisovan population sizes to a very high number (10,000 4N0 generations) from the present until the sampling time, so that the two lineages from each individual would not coalesce with each other. A recombination rate of 1.25x10-8 recombs/generation/base pair was chosen (by trial and error), with the aim of producing a similar distribution of lengths of 'African' haplotypes as observed in the real data. The following is the `ms` command used, obtained by converting the above sizes into units of 4xN0, and converting times to units of 4xN0 generations by dividing by 29x4xN0. N0 can be chosen arbitrarily and was set to 1000. The first eight samples correspond to the African population, the next two to the Altai Neanderthal, and the final two to the Denisovan. The `ms` command is as follows:

```
ms 12 1 -T -seeds <seed1> <seed2> <seed3> -r 10 2000000 -I
3 8 2 2 -n 1 24 -n 2 10000 -n 3 10000 -en 0.431 3 2.5 -en 0.603
2 0.75 -en 1.21 2 3.2 -ej 3.88 3 2 -en 3.88 2 8 -ej 5.34 2 1
-en 5.34 1 17.8
```

The trees output by `ms` were then modified (using a custom perl script) to shorten the branches of each ancient individual, subtracting the sample age.

These modified trees were then given to the program Seq-Gen v1.3.3 [52] to simulate the sequences. The `Seq-Gen` call was the following:

```
seq-gen -q -z <seed4> -p <nump> -mHKY -t3.0 -f0.3,0.2,0.2,0.3
-l2000000 -s 0.00005 < trees.txt
```

where `<nump>` is the number of trees output by `ms`, and trees.txt contains the modified output from `ms`. The mutation rate corresponds to $4N_0 1.25e-8$. The ms and seq-gen commands above were run 100 times with different values of seed1, seed2, seed3, seed4 to produce 100 sets of 2Mb sequences.

2.3.4 Simulations to assess the effects of Sup→Den migration

The simulation parameters were the same as before, except that we now include an unsampled archaic hominin population with a divergence time of one million years from the ancestral human population. Admixture with the Denisovan was simulated to occur 300kya, such that 1% of the Denisovan genome came from this archaic hominin every generation for 10 generations. The `ms` command was:

```
ms 12 1 -T -seeds <seed1> <seed2> <seed3> -r 10 2000000 -I
4 8 2 2 0 -n 1 24 -n 2 100000 -n 3 100000 -n 4 5 -en 0.431 3
2.5 -en 0.603 2 0.75 -en 1.21 2 3.2 -ej 3.88 3 2 -en 3.88 2
8 -ej 5.34 2 1 -en 5.34 1 17.8 -em 2.586 3 4 40 -em 2.589 3
4 0 -ej 12.93 4 1
```

Sequences were generated from the trees created by this command as before, followed by the same ARGweaver analysis.

2.3.5 Positive simulations with Hum→Nea migration

This simulation was done similarly to the previous ones, with the addition of migration from a modern human population into the Altai Neanderthal lineage 100kya at a rate of 3.55% for a single generation. While we did include Sup→Den migration, we used a more modest rate (1% migration for a single generation 300kya). Ancient sample ages of 70kya for the Altai Neanderthal and 50kya for the Denisovan were implemented as before by post-processing the `ms` output. The `ms` command was:

```
ms 12 1 -T -seeds <seed1> <seed2> <seed3> -r 10 2000000 -I
4 8 2 2 0 -n 1 24 -n 2 100000 -n 3 100000 -n 4 5 -en 0.431 3
2.5 -en 0.603 2 0.75 -en 1.21 2 3.2 -ej 3.88 3 2 -en 3.88 2
8 -ej 5.34 2 1 -en 5.34 1 17.8 -em 0.862069 2 1 142 -em 0.862319
2 1 0 -em 2.586207 3 4 40 -em 2.586457 3 4 0 -ej 12.93 4 1
```

1000 replicates of this simulation were created and analyzed by ARGweaver in the same manner as the real data analysis, including randomizing the initial haplotype phasings and use of ARGweavers phase integration feature.

2.4 Conclusion and Future Directions

The ARGweaver analysis conclusively shows that there is an excess of young, long 'African' haplotypes in the Altai Neanderthal genome. The simulations show that this excess cannot be explained by the ages of the archaic individuals, or by introgression into the Denisovan from an unknown archaic hominin. Our simulations do find that this excess is consistent with the signal produced by

an early modern human population into the Altai Neanderthal lineage 100kya. However, the noise in ARGweaver's haplotype ages, as well as the large number of possible migration scenarios to consider, makes it difficult to get a precise estimate of the time of this gene flow event. It is also difficult to precisely identify the introgressed regions; in particular, there is a large excess of 'African' haplotypes that are 350ky old, but they cannot be confidently distinguished from ILS (Figure 2.1a). We also cannot confidently identify potentially introgressed segments of the genome shorter than 50kb without risking a very high false positive rate.

Thus, while this approach was powerful to establish strong evidence for the Hum→Nea migration event, it falls short in its ability to identify and characterize the introgressed regions. We also saw that it had no power at all to confidently detect regions in the Denisovan genome introgressed from a super-archaic hominin (Figure 2.1b). These shortcomings motivate the next chapter of this dissertation. There, we will build ARGs under a demographic model and get much finer resolution and higher power to examine genomic introgression.

CHAPTER 3

MAPPING GENE FLOW BETWEEN ANCIENT HOMININS THROUGH DEMOGRAPHY-AWARE INFERENCE OF THE ANCESTRAL RECOMBINATION GRAPH

3.1 Introduction

It is well established that gene flow occurred among various ancient hominin species over the past several hundred thousand years. The most well-studied example is the interbreeding that occurred when humans migrated out of Africa and came into contact with Neanderthals in Eurasia roughly 50,000 years ago [16, 39]. This left a genetic legacy in modern humans which persists today: between 1-3% of the DNA of non-African humans can be traced to Neanderthals [40]. We also now know that an extinct sister group to the Neanderthals, the Denisovans, intermixed with humans in Asia, leaving behind genomic fragments in 2-4% of the DNA of modern Oceanian humans [17, 45].

Many other admixture events have been hypothesized, creating a complex web of ancient hominin interactions across time and space. These include: between Neanderthals and Denisovans (Nea↔Den) [16, 53]; between Neanderthals and ancient humans who left Africa over 100 thousand years (Hum→Nea) [2]; between an unknown diverged or "super-archaic" hominin (possibly *Homo erectus*) and Denisovans (Sup→Den) [16, 54]; and between other unknown archaic hominins and various human populations in Africa (Sup→Afr) [55, 56]. (In the above notation, the arrows indicate the direction of gene flow hypothesized; in many cases it may have gone both ways, but we lack samples to test the other direction).

As the network of interactions gets more complex, it becomes more difficult to apply standard methods to test for gene flow or identify introgressed regions [47]. For example, a positive value has been observed for the statistic $D(Neaderthal, Denisovan, African, Chimp)$ [16], indicating that there is excess allele sharing between Neanderthals and African humans, as compared to Denisovans and Africans. But, this could be explained by gene flow between Neanderthals and Africans, or from super-archaic hominins into Denisovans, or some combination. The main strategy for teasing apart these scenarios is to examine the age of shared alleles. In this case, the D statistic is highest at sites where the derived allele is fixed or high-frequency in Africa, implying that many of the excess shared alleles are older than the Neanderthal/human divergence, so cannot be explained by Hum↔Nea gene flow. This forms the basis for the hypothesis of super-archaic introgression into Denisovans [16], which predicts a deficit of African-Denisovan shared alleles, as opposed to a surplus of African-Neanderthal sharing. However, it has also been noted that many genomic windows with the lowest Neanderthal-Africa divergence nevertheless have high Neanderthal-Denisovan divergence, which is best explained by Hum→Nea gene flow [2]. Currently, both events have support from multiple studies, including: model-based demography estimation by GPhoCS [2, 10], using ARGweaver [1] to examine coalescence times for gene trees that do not match the species tree [2], and comparing the frequency-stratified D-statistics with those from extensive simulations under various models of gene flow [54].

While both Sup→Den and Hum→Nea events have substantial support, it remains challenging to identify introgressed genomic regions that result from them. This problem is more difficult than identifying regions introgressed into modern non-African humans from Neanderthals and Denisovans, both because

we do not have a sequence from the super-archaic hominin, and because these events are likely older, and therefore the haplotypes more broken up by recombination. We are further limited by the very small numbers of sequenced Neanderthal and Denisovan genomes. Current approaches, including the conditional random field (CRF) [40,45] and the S* statistic [57,58] (and recent variant Sprime [59]), have been tuned to the problem of finding recent introgression into humans. Furthermore, they only use a small number of summary statistics, such as locations of specific patterns of allele sharing. When the genomic signal is more subtle, it may be necessary to incorporate all the data with careful methodology in order to have sufficient power to confidently detect these regions.

In this paper, we present ARGweaver-D, which infers ancestral recombination graphs (ARGs) [60–62] conditional on a generic demographic model that includes population splits, size changes, and migration events. The ARG consists of local trees across a chromosome, representing the ancestral relationships among a set of sequenced individuals at every genomic position. In this extension to ARGweaver, the ARGs also contain information about the population membership of each lineage at every time point, so that introgressed regions are encoded in the ARG as lineages that follow a migrant path. Unlike most other methods, this approach allows multiple types of introgression to be inferred simultaneously, and takes into account the full haplotype structure of the input seqeunces. It works on unphased genomes and can accommodate changing migration and recombination rates. ARGweaver-D is a Bayesian method, using Markov chain Monte Carlo (MCMC) iterations to remove and "rethread" branches into the local trees; as a result, the output of ARGweaver-D is a series of ARGs that are sampled from the posterior distribution of ARGs conditional

on the input data and demographic model. From these, we can extract posterior probabilities of introgression for any lineage at any genomic position.

Another recent method, `dical-admix` [46], is similar to ARGweaver-D in that it is designed to accommodate generic demographic models, and takes the full haplotype structure of the input sequences into account. However, there are some important differences that make our approach more applicable to the complex history of ancient hominins. `dical-admix` assumes that there are only a few admixed individuals, and that other genomes are "trunk" lineages that help define the haplotype structure of their respective populations. It therefore cannot infer admixture from an unsampled population, nor is it designed to work when all individuals have some degree of admixed ancestry. Additionally, ARGweaver-D can handle unphased genomes, which is important since there are not enough Neanderthal or Denisovan samples to reliably phase these archaic genomes.

After introducing ARGweaver-D, we present simulation studies showing it can successfully detect Nea→Hum introgression, even when using a limited number of genomes. We then use simulations to show that it can also detect older migration events, including Hum→Nea, Sup→Den, and Sup→Afr, depending on the underlying demographic parameters. We then apply this method to African humans and ancient hominins, classifying 3% of the Neanderthal genome as introgressed from ancient humans, and 1% of the Denisovan genome as introgressed from a super-archaic hominin. In contrast to Nea→Hum introgression, we do not see any evidence of selection against Hum→Nea introgression.

3.2 Results

3.2.1 ARGweaver-D can estimate genealogies conditional on arbitrary demographic model

ARGweaver-D is an extension of ARGweaver [1]that can infer ARGs conditional on a user-defined population model. This model can consist of an arbitrary number of present-day populations that share ancestry in the past, coalescing to a single panmictic population by the most ancestral discrete time point. Population sizes can be specified separately for each time interval in each population. Migration events between populations can also be added; they are assumed to occur instantaneously, with the time and probability defined by the user.

Recall that ARGweaver is a MCMC sampler, in which each iteration consists of removing a branch from every local tree in the ARG ("unthreading"), followed by the "threading" step, which resamples the coalescence points for the removed branches. This threading step is the main engine behind ARGweaver, and is accomplished with a hidden Markov model (HMM), in which the set of states at a particular site consists of all possible coalescence points in the local tree. In the original version of ARGweaver (with a single panmictic population), each of these states is defined by a branch and time. In ARGweaver-D, each state has a third property, which we call the "population path", representing the population(s) assigned to the new branch throughout its time span. The modified threading algorithm is illustrated and further described in Fig 3.1.

Without migration events, and assuming that present-day population as-

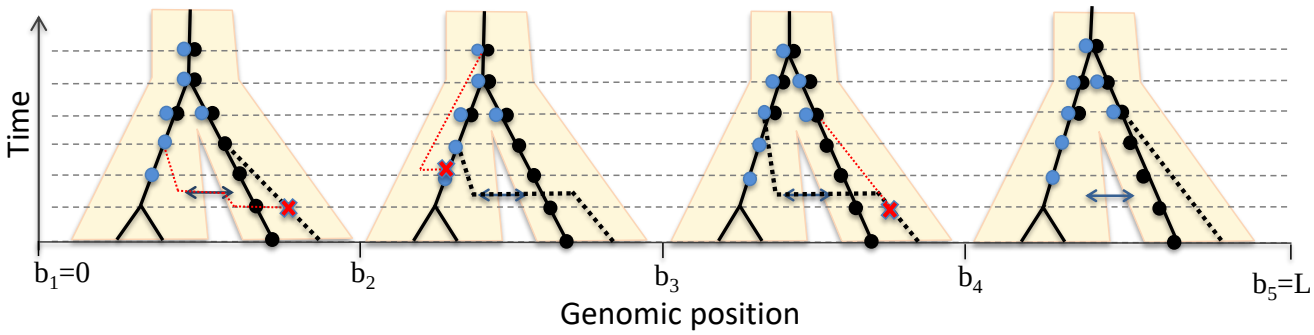

Figure 3.1: **Illustration of the "threading" operation under a model with two populations and a single migration band.** The gray horizontal dashed lines represent the discrete time points in the ARGweaver model, when coalescence and recombination events occur; migration and population divergence times are pre-specified by the user and rounded to the nearest "half time-point" (midway between the dashed lines). Migration is assumed to occur instantaneously at a rate p_M specified by the uesr. This ARG currently has three haploid samples, indicated by the solid black lines. A fourth sample from the right-hand population is being threaded into the ARG, with the dotted black line representing one possible threading outcome. Each dot on the tree is a potential coalescence point for the new branch, representing a state in the threading HMM. The black dots are states from a population path with no migration, whereas the blue dots are from the migrant population path. Recombination events occur immediately before positions b_2, b_3, and b_4, as indicated by the red X on the trees preceding those positions. The dotted red line shows the recoalescence of the broken branch, which defines the tree at the next site. The recombinations before b_2 and b_4 would be sampled after the threading algorithm, as they occur on the branch being threaded, whereas the recombination before b_3 is part of the ARG before the threading, and therefore not modified at this stage. Here, we only show a single tree in each interval between recombination events; the local tree is identical within each of these intervals. The lineage being threaded enters an introgressed state at position b_2, and leaves it at b_4. The transition probabilities of the HMM are calculated between each pair of adjacent states; the probability of migration p_M is a factor in the transition observed at b_2. It is not a factor at b_3 because the new branch is already in a migrant state. The transition probability at b_4 includes a factor of $1 - p_M$.

signment for each branch is known, the ARGweaver-D is more efficient than the original panmictic version. This is because coalescence is not possible unless two branches are in the same population at the same time, and so the state space of potential coalescence points will be a subset of the original state space. However, as migration events are added, coalescence points in other populations become possible, and some coalescence points may be reachable by multiple population paths (see Fig 3.1). Therefore, the complexity of the algorithm can quickly increase. Whereas the original threading algorithm had an asymptotic running time of $O(Lnk^2)$ (where L is the number of sites, n the number of samples, and k the number of time points), ARGweaver-D is $O(Lnk^2P^2)$, where P is the maximum number of population paths available to any single lineage.

One way to improve the efficiency is to allow at most one migration event at any genomic location. Note that this assumption still allows multiple lineages to be introgressed at the same genomic position, if they are descended from a common migrant ancestor. This assumption is reasonable when the number of samples is small and the migration rate is low, and is set as a default in ARGweaver-D that we use throughout this paper. It has two advantageous side-effects: it avoids strange parts of the state space that could cause MCMC mixing problems (such as back-migrations, or population label switching issues). It also means that if we are modelling introgression from a "ghost" population such as a super-archaic hominin (from which we have no samples), there will be at most one (migrant) lineage in the population at any location. Therefore, the population size of ghost populations does not matter as coalescence will not occur within them.

After running ARGweaver-D, it is straightforward to identify predicted in-

trogressed regions; they are encoded in each ARG as lineages that follow a migration band. By examining the set of ARGs produced by the MCMC sampler, ARGweaver-D can compute posterior probabilities of introgression across the genome; this can be done in any way that the user would like: as overall probabilities of migration anywhere in the tree, or probabilities of a specific sampled genome having an ancestral lineage that passes through a particular migration band. For a diploid individual, we can look at probabilities of being heterozygously or homozygously introgressed. Throughout this paper we use the cutoff of $p \geq 0.5$ to define predicted introgressed regions, and compute total rates of called introgression for a diploid individual as the average amount called across each haploid lineage.

More details of the ARGweaver-D algorithm are given in Fig 3.1 and the Supplementary Text. ARGweaver-D is built into the ARGweaver source code, which is available at: `http://github.com/CshlSiepelLab/argweaver`.

3.2.2 ARGweaver-D can accurately identify archaic introgression in modern humans

We performed a set of simulations to assess the performance of ARGweaver-D for identifying Neanderthal introgression into modern humans. These simulations realistically mimic human and archaic demography, as well as variation in mutation and recombination rates (see Methods). We compared the performance with the CRF algorithm [40]; Fig 3.2 summarizes the results. Overall, ARGweaver-D has improved performance over the CRF, which is subtle for long segments but becomes more pronounced for shorter segments. This gain in

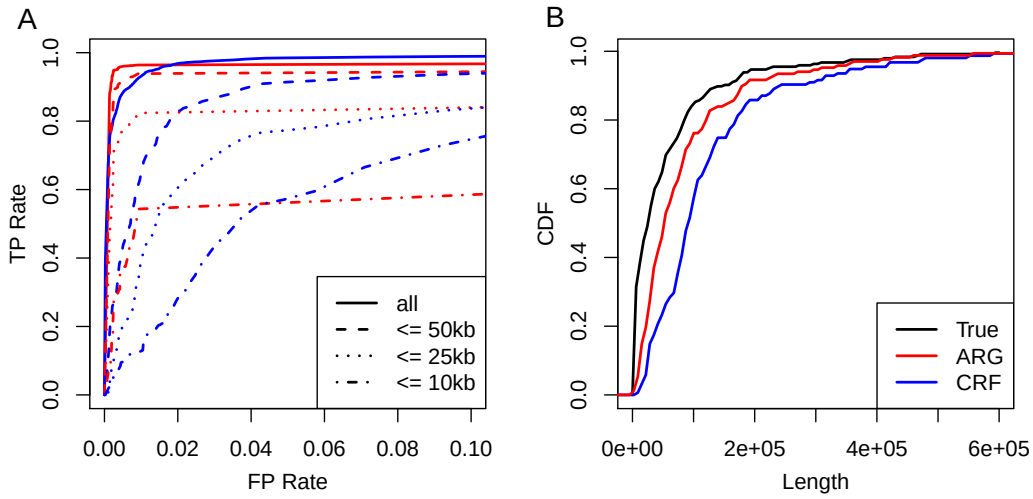

Figure 3.2: **Performance on Nea→Hum simulations.** **A:** ROC curves showing basewise performance of ARGweaver-D (red) and the CRF (blue) for predicting introgressed regions in simulated data. The two methods predicted introgression in the same simulated European individuals, however the CRF made use of the full reference panel (43 diploid Africans), whereas ARGweaver-D only used a small subset of the reference panel (2 diploid Africans). Different line patterns correspond to different maximum segment lengths. **B:** The length distribution of real and predicted introgressed regions for the same simulations and predictions shown in panel A.

power is despite the fact that the CRF used a much larger panel of African samples than was used by ARGweaver-D. (CRF used 43 Africans, ARGweaver-D used only 2 to save computational cost).

Next, we predicted introgressed regions in two non-African human samples from the Simons Genome Diversity Panel (SGDP), one European (Basque) and a Papuan. The ARGweaver-D model used is illustrated in Fig 3.3; in this case only the "Recent migration" bands were included. We compared to calls on the same individuals from the CRF. Again, ARGweaver-D used two Africans, whereas the CRF used 43. And while the CRF uses Africans as a control group,

ARGweaver-D allows for introgression into any of the human samples. The results are summarized in Fig 3.4. Overall, the two methods call a large fraction of overlapping regions, but each method also produces a substantial fraction not called by the other method (between 15-40%), and ARGweaver-D generally calls more regions. While ARGweaver-D seems to have greater power in simulations, another factor in the discrepancy may be that the ARGweaver-D segments were called with the inclusion of both the Altai and Vindija Neanderthals, whereas the CRF calls were produced with the Altai Neanderthal only. Both methods show a strong depletion of introgression on the X chromosome, especially in the Basque individual.

Notably, ARGweaver-D calls close to 0.5% introgression from Neanderthal into each of the African individuals. These calls may be explained by a combination of false positives and back-migration into Africa from Europe. Another possibility is that regions introgressed into Neanderthals from ancient humans [2] may be identified in the wrong direction under this model. With few samples, it is likely difficult to determine the direction of migration between two sister populations. Indeed, when we simulate migration in both directions, but still have only a Nea→Hum migration band in the ARGweaver-D model, 8% of Hum→Nea bases are identified as Nea→Hum. (See Supplementary Text). This is our motivation for excluding non-African samples when looking for introgressed regions from older migration events in the next section.

Finally, we compared the rate of calls in the Basque individual to predictions of Neanderthal ancestry based on the F4-Ratio statistic F4(Altai, chimp; Basque, African)/F4(Altai, chimp; Vindija, African) [47]. Both ARGweaver-D and CRF predicted fewer elements (1.95% and 1.56%, respectively), compared to the F4

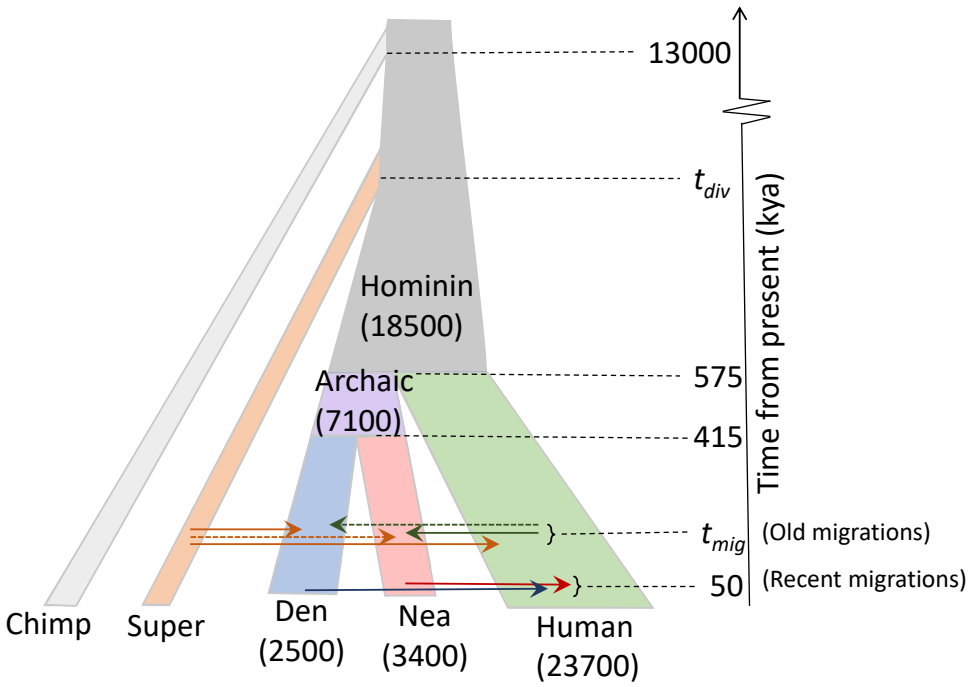

Figure 3.3: **Population model used for ARGweaver-D analysis**. Population sizes are given in parentheses. The model is invariant to the population sizes of the chimpanzee and super-archaic hominin, as no more than a single lineage ever exists in each of these populations. Migration events are shown by arrows between populations; solid arrows are used for events which have been proposed by previous studies. All parameters except t_{mig} and t_{div} are held constant at the values given.

ratio statistic (2.31%). Looking across the chromosomes, there is a higher correlation between coverage predicted by ARGweaver-D and the expectation from the F4 ratio (Spearman's $\rho = 0.75$), than between CRF and the F4 ratio ($\rho = 0.51$) (Fig 3.5).

3.2.3 ARGweaver-D can detect older introgression events

We next did a series of simulations to assess ARGweaver-D's power to detect other ancient introgression events that have been previously proposed. To fo-

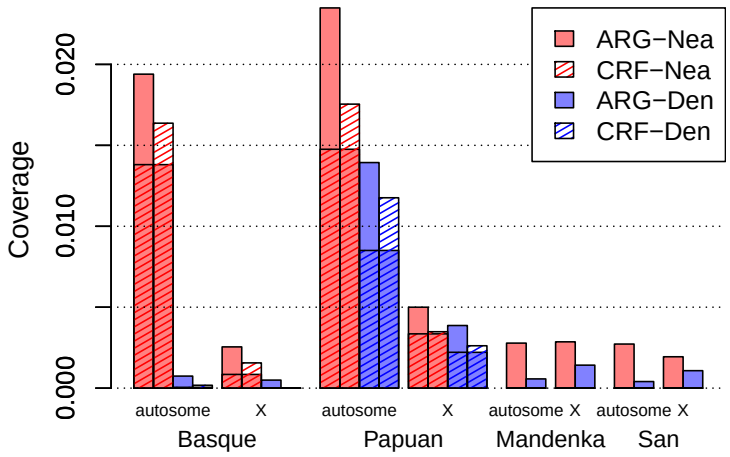

Figure 3.4: **Average coverage of predicted introgressed regions into four SGDP individuals.**s The regions that are both colored and striped represent regions called by both CRF and ARGweaver-D. The CRF calls were only produced for non-African individuals, so only ARGweaver-D results are shown for Mandenka and San.

cus on these older events, we simulated the modern human samples using on a model of African human population history, and as such did not include the migration from Neanderthals or Denisovans into non-African humans. The simulations included three migration events: from modern humans into Neanderthals (Hum→Nea), from a "super-archaic" unsampled hominin into Denisovans (Sup→Den), and also from super-archaic into Africans (Sup→Afr). (Note that although both Sup→Afr and Sup→Den involve introgression from the same super-archaic population, it is only meant to represent introgression from any unsampled, diverged hominin species, and does not necessarily imply that the same population admixed with both Africans and Denisovans.) The simulations included many realistic features: ancient sampling dates for the archaic hominins, variation in mutation and recombination rates, randomized phase, and levels of missing data modeled after the SGDP and ancient genomes that

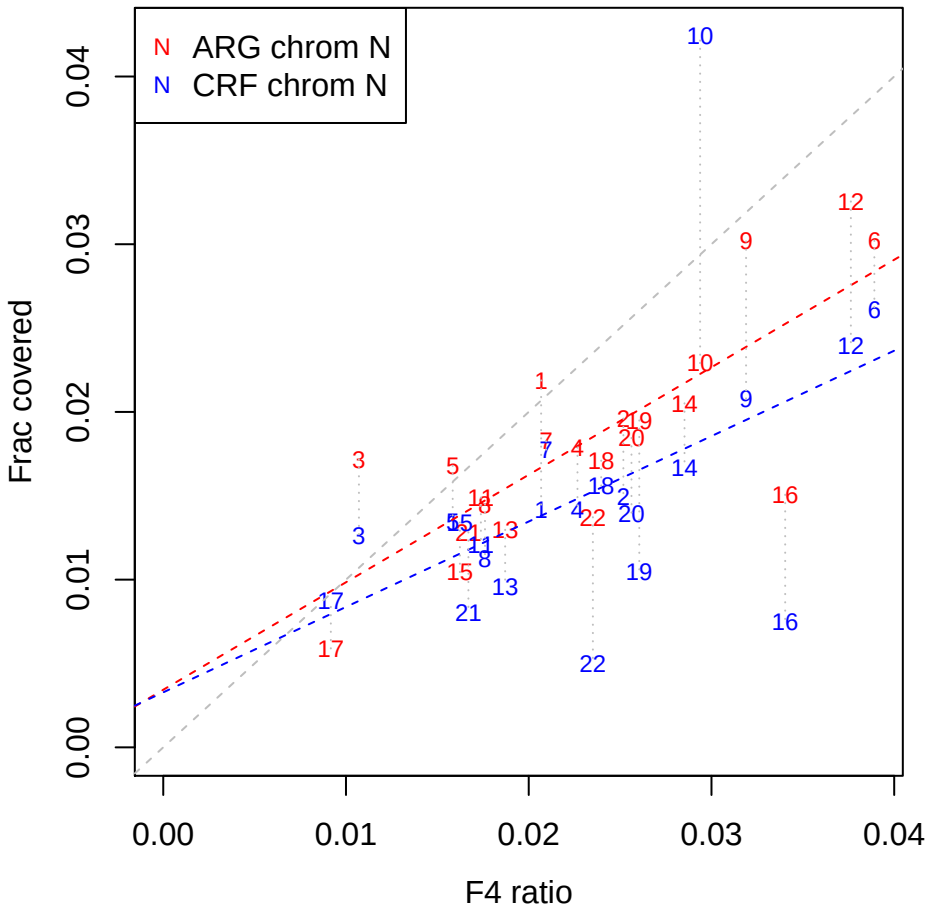

Figure 3.5: **Coverage of introgression predictions vs. F4 Ratio.** For each autosomal chromosome, we plot the expected fraction predicted introgressed by ARGweaver-D (red) and CRF (blue) into the Basque individual, vs the F4Ratio statistic F4(Altai, chimp, Basque, Afr)/F4(Altai, chimp, Vindija, Africa) computed for variants on each chromosome. The dashed red and blue lines show the best linear fit for each method, whereas the gray dashed line shows $x = y$ for reference. Dotted gray lines connect results from the same chromosome.

we use for analysis (see Methods). Each set of simulations contained all three types of migration and ARGweaver-D detected all migration events in a single run with multiple migration bands in the model.

We analyzed these data sets with ARGweaver-D using the model depicted in Fig 3.3, with only the "old migration" bands. As we do not have good prior estimates for the migration time (t_{mig}) or super-archaic divergence time (t_{div}), we tried four values of t_{mig} (50kya, 150kya, 250kya, 350kya) and two values of t_{div} (1Mya, 1.5Mya). We generated data sets under all 8 combinations of t_{mig} and t_{div}, and then analyzed each data set with ARGweaver-D under all 8 models, in order to assess the effects of model misspecification on the inference.

The power of ARGweaver-D to detect introgression is summarized in Fig 3.6. The left side of the plot represents simulations generated with t_{div} = 1Mya, whereas the right side used t_{div} = 1.5Mya. Power to detect super-archaic introgression is clearly much higher when the divergence is higher, but (as expected) does not affect power to detect Hum→Nea introgression. Looking from top to bottom, the plots show the effect of increasing the true time of migration. In the top plot with t_{mig} = 50kya, only results for Sup→Afr are shown because the archaic hominin fossil ages pre-date the migration time. For all events, we see power decrease as the true migration time decreases.

For a given simulation set, the effect of the parameters used by ARGweaver-D are generally more subtle. We note that power tends to be better when older migration times are used in the model, even when the true migration time is recent; in particular, the power when t_{mig} = 150kya (red bars) is often much worse than when later times are used, especially for the Hum→Nea event. Similarly, power is often better when t_{div} is set to 1Mya in the ARGweaver-D model, as

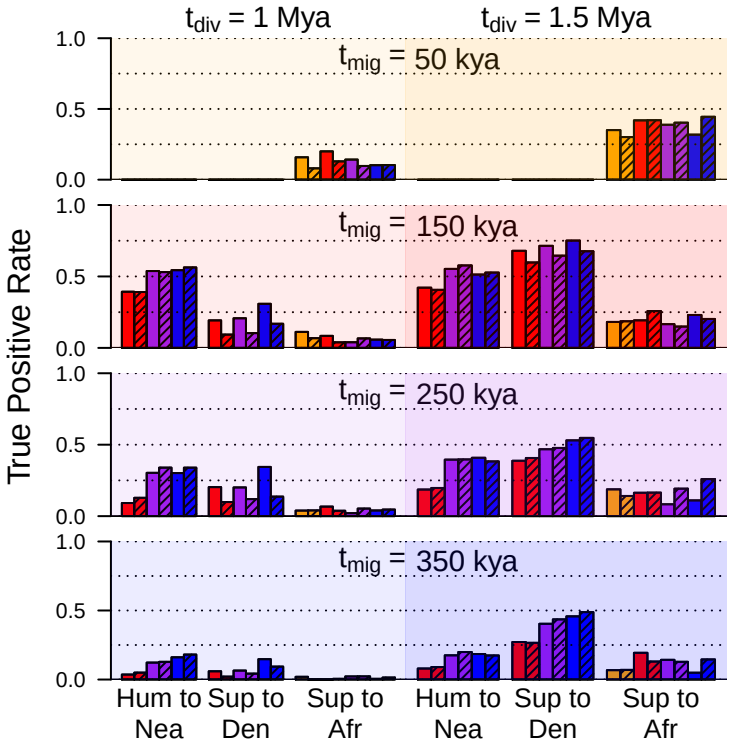

Figure 3.6: **Simulation results.** Each shaded box represents a set of simulations generated with a different value of t_{mig} and t_{div}. Each bar gives the basewise true positive rate for a particular migration event and ARGweaver-D model, using a posterior probability threshold of 0.5. The color of each bar represents the value of t_{mig} used in the inference model (orange=50kya, red=150kya, purple=250kya, blue=350kya); shaded bars have t_div=1.5Mya in the inference model, whereas solid bars use t_{div} = 1.0Mya. Because the archaic hominin fossil ages are older than 50kya, results for t_{mig} = 50kya (top) are only applicable for introgression into humans.

opposed to 1.5Mya.

In summary, ARGweaver-D has reasonably good power to detect super-archaic introgression when the divergence time is old, but power is more limited as the divergence decreases. The power to detect Sup→Afr is always lower than the power to detect Sup→Den, as the African population size is much larger, making introgression more difficult to distinguish from incomplete lineage sort-

ing. For the Hum→Nea event, we have around 50% power if the migration time is 150kya, and around 30% power when it is 250kya.

False positive rates are less than 1% when a posterior probability threshold of 0.5 is used (Fig 3.7). When analyzing the simulated data sets, we included two additional migration bands in the ARGweaver-D model as controls: one from the super-archaic population to Neanderthal (Sup→Nea), and another from humans into Denisova (Hum→Den). The rates of calling these events were also less than 1% for all models. Importantly, the rate of mis-classification is very low for all categories (Fig 3.8); in particular, the model can easily tell the difference between Hum→Nea and Sup→Den events, despite both resulting in similar D statistics [2, 54].

More details about the simulation results are available in the Supplementary Text. One issue to note is that, although the simulated data sets were generated with a human recombination map, the ARGweaver-D model used a simple constant recombination rate. Performance is somewhat better when ARGweaver-D uses the true recombination map, but in practice there are not enough Neanderthal or Denisovan samples to generate a reliable recombination map, and there is no data to infer the recombination map for the super-archaic population. The Supplementary Text also shows results when we simulate more African individuals. We find that performance does not improve as samples are added, so in the main text we focus on analysis with two African samples (four haploid genomes).

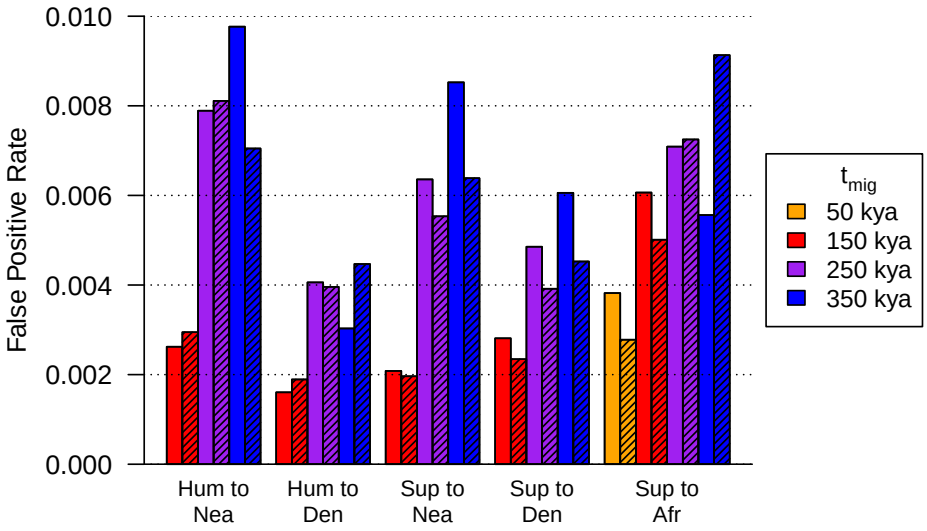

Figure 3.7: **False positive rates calculated from simulations, using several ARGweaver-D models.** Color indicates the value of t_{mig}. Shaded bars have t_{div} = 1.5Mya and solid bars have t_{div} = 1.0Mya. False positive rates are calculated base-wise using a posterior probability cut-off of 0.5. The same set of underlying data was used for all the calculations in this plot; it was simulated as in Fig 3.3, but with no true migration events.

3.2.4 Deep introgression results

We next applied the models from the previous section to real modern and archaic human genomes. Our goals were to identify and characterize introgressed regions from previously proposed migration events, as well as to see if we find evidence for other migrations which may not be detectable using other methods. Our data set consisted of two Africans from the SGDP [63], two Neanderthals [16,54], the Denisovan [17], and a chimpanzee outgroup. We again use the demography illustrated in Fig 3.3, with old migration events only. We focus on the model with t_{mig}=250kya and t_{div}=1Mya, because this model seemed to have high power in all our simulation scenarios, and because our results sug-

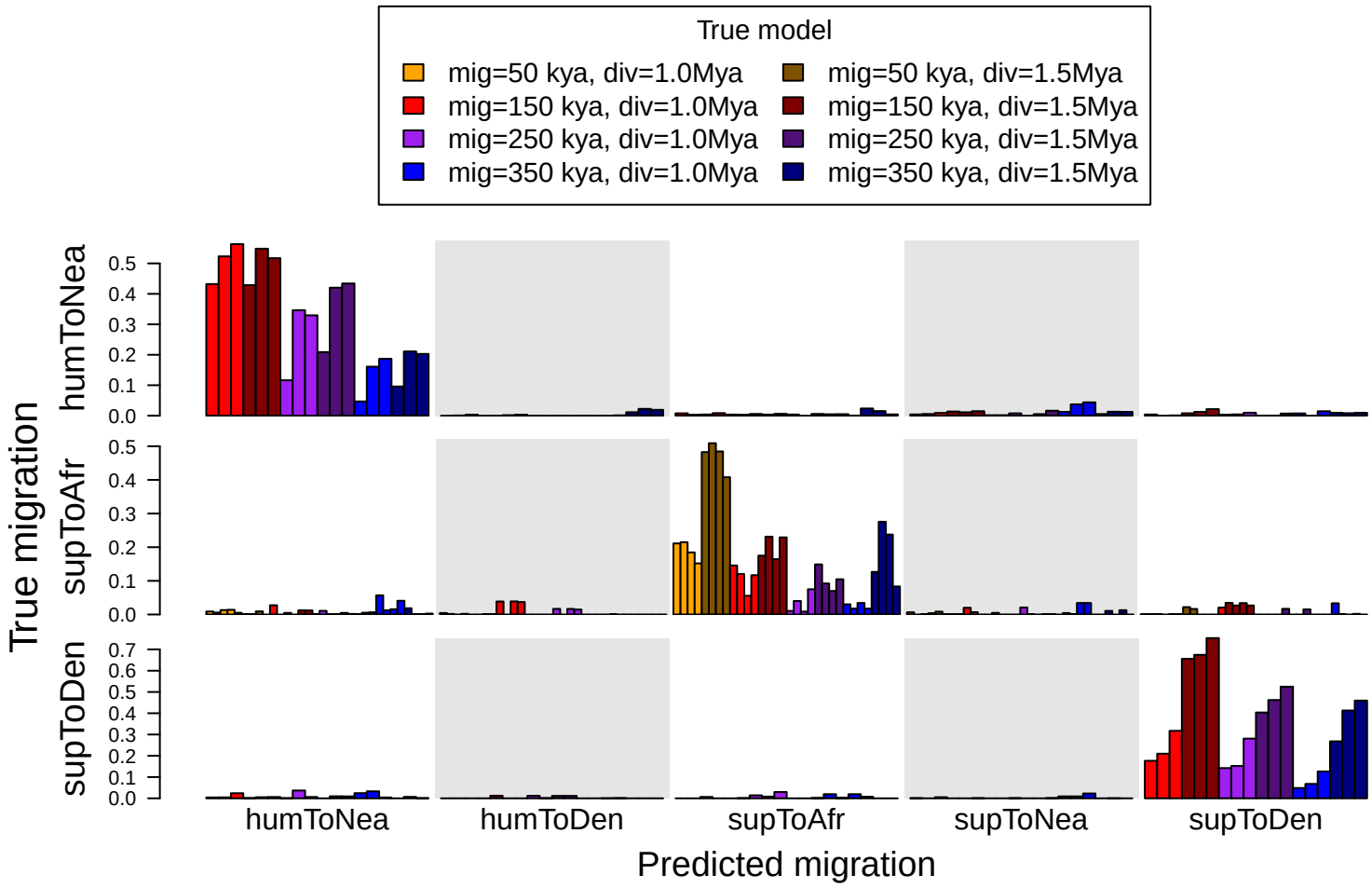

Figure 3.8: **Detailed simulation results.** Each row shows a true migration category, and each column shows the fraction of bases predicted in the category indicated at the foot of the column. The color of each bar represents the true parameters used in simulation, as indicated in the legend, with darker colors used for the older super-archaic divergence time. Multiple bars of the same color show results on the same data set, using an ARGweaver-D model with a different t_{mig}. The value of t_{mig} used by ARGweaver-D is not indicated in the plot, but increases from left-to-right: t_{mig} = 50, 150, 250, 350kya, with 50kya only shown for Sup→Afr. All the models used t_{div} = 1Mya; the plot with t_{div} = 1.5Mya is nearly identical.

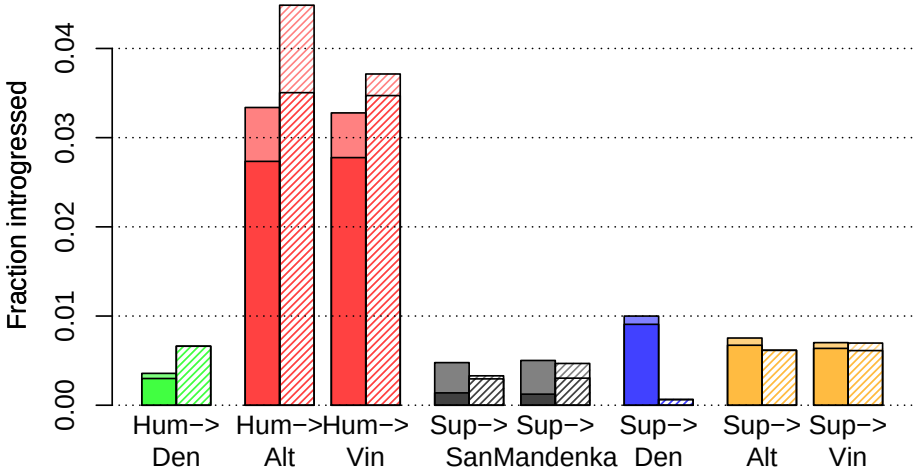

Figure 3.9: **Genome-wide coverage of predicted ancient introgression**. These results use a posterior probability cutoff of 0.5; solid bars are for autosomes, and striped bars for chromosome X. Each bar shows total average coverage for a haploid genome; the darker bottom portions of each bar represent homozygous calls.

gest that it may be the most realistic (as discussed below). The results using other models are consistent with those presented here, and are described in the Supplementary Text.

An overview of the coverage of predicted introgressed regions is depicted in Fig 3.9; a more detailed summary is given in Fig 3.10. The most immediate observation is that Hum→Nea regions are called most frequently, at a rate of ~ 3% in both the Altai and Vindija Neanderthal. This number is almost certainly an underestimate, given that the true positive rate for this model was measured between 30-55%. By contrast, only ~ 0.37% of regions are classified as Hum→Den. As no previous study has found evidence for Hum→Den migration, this serves as a control, verifying that our false positive rate estimated in simulations is likely fairly accurate, as we estimated a FP rate of 0.41%.

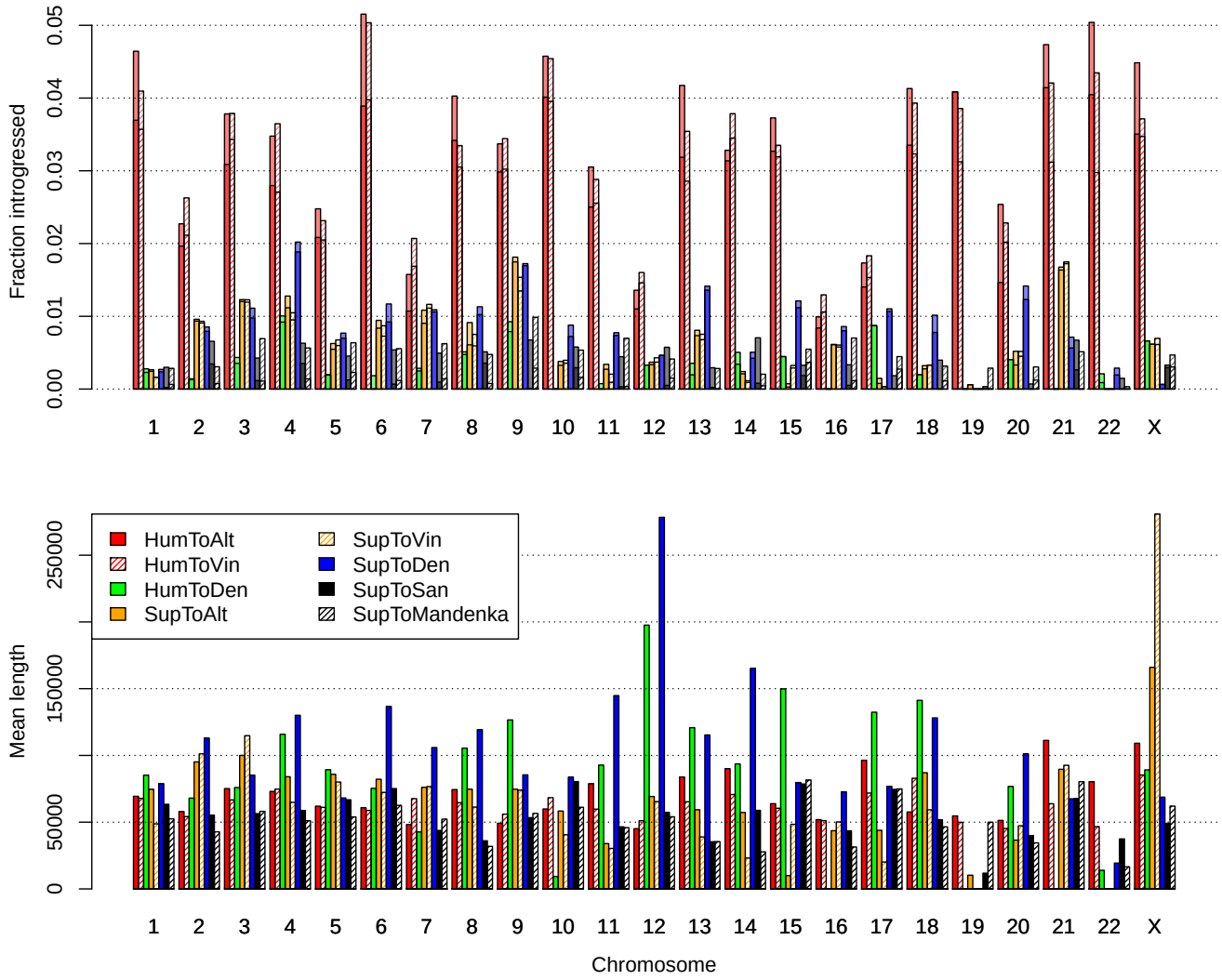

Figure 3.10: **Properties of introgressed regions by chromosome** The top plot shows average coverage of predicted introgressed regions per haploid genome, with darker portions representing homozygous regions. The bottom shows average length of introgressed regions by chromosome.

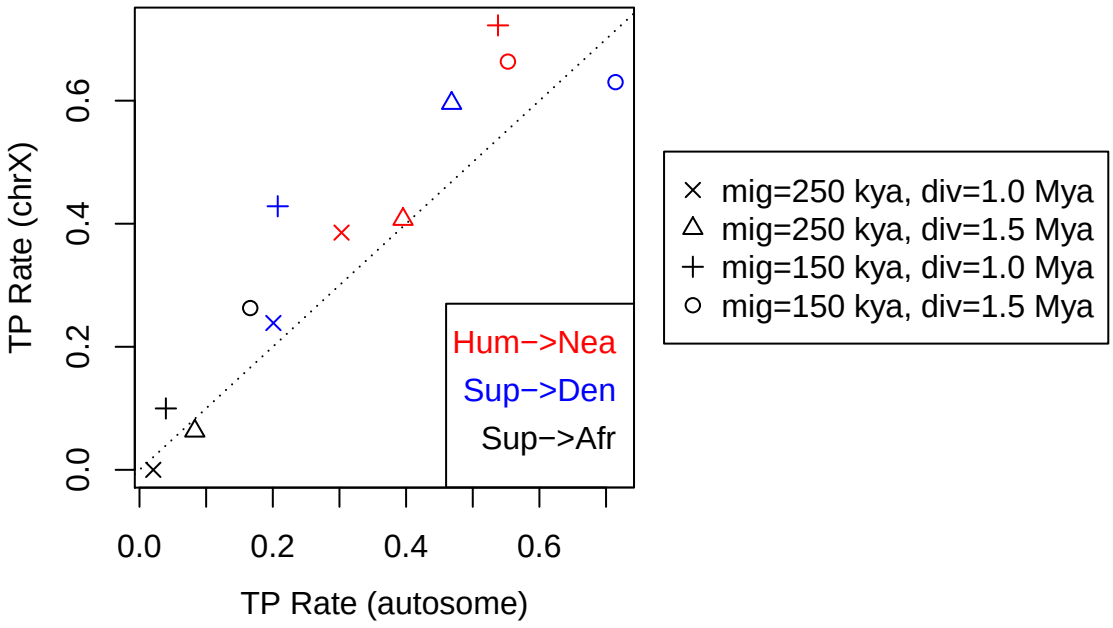

Figure 3.11: **True positive rate for simulations on X chromosome vs autosomes**. The y-axis shows true positive rates from simulations where population sizes were multiplied by 0.75 to roughly approximate X chromosome demography. Different plotting characters are used for different simulation models, as indicated in the legend. All ARGweaver-D analysis was done with t_{mig}=250kya and t_{div}=1.0Mya.

Whereas there is a depletion on the X chromosome of archaic introgression into humans, we see high coverage of Hum→Nea on the X for both Altai and Vindija. The fact that it is somewhat higher on the X than the autosomes might be partially explained by increased power on the X; simulations suggest that power will be ~ 20% higher for this event when population sizes are multiplied by 0.75 (Fig 3.11). Overall, there is a lot of variation in detected introgression across the chromosomes, and several autosomal chromosomes have higher predicted coverage than the X, including 1, 6, 21, and 22 (Fig 3.10).

Although the Vindija sample is 70ky younger than the Altai sample [54],

there is no apparent depletion of human ancestry on Vindija compared to Altai on the autosomes, suggesting that negative selection did not cause a significant loss of human introgressed regions in the Neanderthal during that time. Several chromosomes do show drops in coverage from Altai to Neanderthal, with the largest drop on the X chromosome (Fig 3.10).

Other migrations are detected at lower levels. We identify 1% of the Denisovan genome as introgressed from a super-archaic hominin, which is double our estimated false positive rate for this event. The fact that we found much less than the ~ 6% estimated by previous methods [54] might suggest that the super-archaic divergence time is closer to 1Mya, since we would expect to have more power with a higher divergence time. Still, this analysis resulted in 27Mb of sequence that may represent a partial genome sequence from a new archaic hominin. ARGweaver-D also predicted a small fraction of the Neanderthal genomes as introgressed from a super-archaic hominin (0.75% for Altai and 0.70% for Vindija). These amounts are only slightly above the estimated false positive rates (0.65%), and the Sup→Nea event has not been previously hypothesized.

One interesting aspect of Sup→Den and Sup→Nea regions is that, to the extent that these predictions are accurate, there is the potential that this super-archaic sequence was passed to modern humans through subsequent Den→Hum and Nea→Hum migrations. We explored these regions further by intersecting them with introgression predictions across the full SGDP data set. This analysis is detailed in the Supplementary Text. It first confirms that most Sup→Den and Sup→Nea regions have higher-than expected divergence to the Denisovans and Neanderthals (respectively) across all humans, and not

just the two African humans used by ARGweaver-D. 15% of the Sup→Den regions overlap with sequence introgressed into Asian and Oceanian individuals from Denisovans, and many of these regions also contain a high number of variants consistent with super-archaic introgression. We also see that 35% of the Sup→Nea regions are introgressed in at least one modern-day non-African human. We also identified one region of hg19 (chr6:8450001-8563749) which appears to be Neanderthal-introgressed and overlaps a Sup→Nea region. We compiled a list of Sup→Den and Sup→Nea regions that overlap human introgressed regions, and the genes that fall in these regions. These are given in Tables 3.1 and 3.2.

Location (hg19)	**count**	**overlapping genes**
chr15:56880301-56943860	10	RP11-1129I3.1, ZNF280D
chr5:35268551-35472820	9	U3
chr15:79936231-80045380	7	
chr2:183978771-184038340	7	NUP35
chr1:40622111-40751801	6	RLF, RNU6-1237P, TMCO2, RP1-39G22.7, ZMPSTE24
chr4:143486431-143606100	6	INPP4B, RP11-223C24.1
chr15:63493991-63599658	5	RAB8B, APH1B
chr17:30992260-31232970	3	MYO1D, RP11-220C2.1, Y_RNA, AC084809.2, AC084809.3
chr5:74577091-74897550	3	CTD-2235C13.2, HMGCR, COL4A3BP, CTD-2235C13.3, POLK, RNU7-175P, CTC-366B18.2
chr20:18369011-18456230	3	DZANK1, RNA5SP476, POLR3F, MIR3192
chr2:104441221-104575299	2	AC013727.1, AC013727.2, RP11-76I14.1
chr3:156394341-156515810	2	TIPARP, RP11-392A22.2
chr8:97918711-98192640	2	CPQ, KB-1958F4.2, KB-1958F4.1
chr8:56673021-56798570	2	TMEM68, TGS1, LYN
chr13:77606499-77899730	1	MYCBP2, MYCBP2-AS1, RP11-226E21.2
chr4:85723291-85798820	1	WDFY3, RP11-147K21.1
chr6:131062559-131237440	1	SMLR1, EPB41L2
chr7:83335141-83452959	1	
chr10:52582401-52700350	1	A1CF, RP11-449O16.2
chr3:129951121-130100515	1	COL6A5, AC093004.1

Table 3.1: **Sup→Den regions overlapping Den→Hum regions predicted by the CRF.** The "count" column shows the number of non-African SGDP individuals who have Denisovan introgression at this locus. We restricted this list to Sup→Den regions for which at least 90% of SGDP individuals without Denisovan introgression have a higher divergence to the Denisovan than to Neanderthals.

Location (hg19)	count	overlapping genes
chr6:8450001-8563749	71	HULC
chr7:44396121-44543978	37	RP5-844F9.1, NUDCD3, RNU6-1097P
chr4:121531631-121587672	32	RP11-501E14.1
chr9:30239959-30438940	21	LINC01242
chr7:50432351-50497990	17	IKZF1, CTC-736O2.1
chr9:94891421-95445440	17	snoU13, RP11-62C3.6, IARS, SNORA84, NOL8, CENPP, OGN, OMD, ASPN, ECM2, MIR4670, IPPK
chr3:16970431-17045770	15	PLCL2, MIR3714
chr6:120748701-120851630	14	RNU6-214P
chr7:85753641-85880460	14	
chr9:73603471-73725370	13	TRPM3
chr11:42691811-42766780	13	
chr4:106603601-106693390	9	INTS12, GSTCD, RP11-45L9.1
chr15:67472779-67650950	8	SMAD3, AAGAB, IQCH
chr6:41014213-41153400	7	APOBEC2, OARD1, NFYA, TREML1, TREM2
chr2:84214131-84279410	7	
chr4:42929991-43023170	5	GRXCR1
chr4:161759472-162023170	5	AC106860.1
chr5:342721-451430	5	AHRR, C5orf55, EXOC3
chr9:88250001-88377170	4	AGTPBP1, RP11-202I11.2
chr4:117545891-117601428	4	
chr4:18307191-18446628	3	
chr12:92114981-92187270	3	
chr4:18307191-18368540	3	
chr6:46356781-46434892	3	RCAN2
chr18:47574231-47700390	2	MYO5B
chr4:81260001-81631350	2	C4orf22
chr6:44893490-45311660	2	SUPT3H, MIR586, RUNX2
chr13:84262861-84384570	1	

chr13:87243951-87386382	1	
chr3:100416521-100565100	1	TFG, ABI3BP
chr4:18084421-18196990	1	
chr4:98345401-98553210	1	RP11-18N21.2, RP11-681L8.1, AC034154.1, STPG2
chr13:65910681-66015839	1	

Table 3.2: **Sup→Nea regions overlapping Nea→Hum regions predicted by the CRF.** The "count" column shows the number of non-African SGDP individuals who have Neanderthal introgression at this locus. We restricted this list to Sup→Nea regions for which at least 90% of SGDP individuals without Neanderthal introgression have a higher divergence to the Neanderthal than to Denisovans.

We examined lengths of all our sets of predicted regions, as they might be informative about the time of migration. However, we find that there is strong ascertainment bias towards finding longer regions, so that the length distributions are highly overlapping for different migration times. (See Supplementary Text).

Instead, we looked at the frequency spectrum of introgressed regions to gain insight into the times of migration events. The older the migration, the more likely that an introgressed region has drifted to high frequency and is shared across the sampled individuals. For the Hum→Nea event, we observed 37% of our regions are inferred as "doubly homozygous" (that is, introgressed across all four Neanderthal lineages). This is very close to what we observe in regions predicted from our simulations with migration at 250kya (38%), whereas simulations with migration at 150kya and 350kya had doubly-homozygous rates of 10% and 55%, respectively. To further narrow down the range of times, we did additional simulations with t_{mig} =200, 225, 275, and 300kya, and compared the frequency spectrum of introgressed regions after ascertainment with

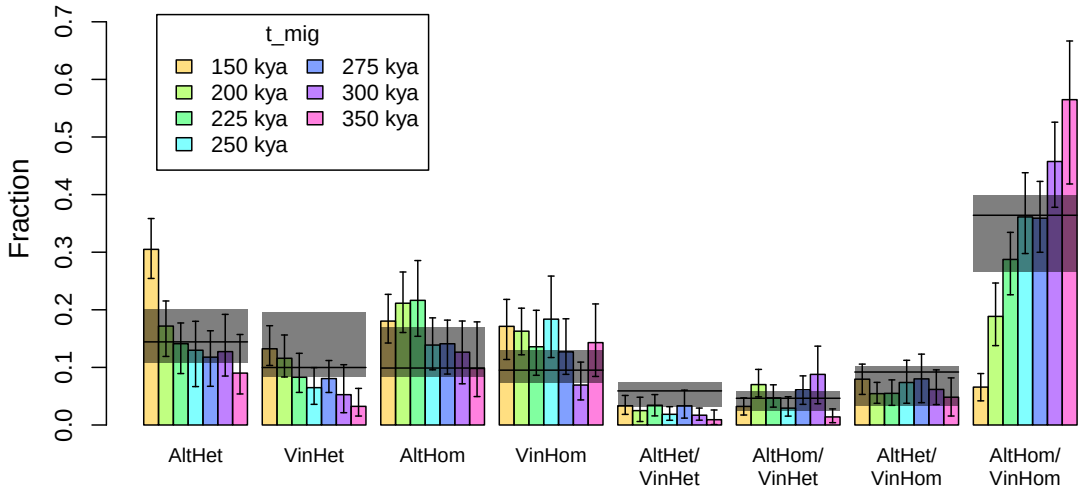

Figure 3.12: **Frequencies of Hum→Nea introgression categories**. For both the real and simulated data, Hum→Nea regions were ascertained with ARGweaver-D using a model with t_{mig} = 250kya and t_{div} = 1Mya. These regions were classified as heterozygous/homozygous in the Altai Neanderthal (AltHet/AltHom), and in the Vindija Neanderthal (VinHet/VinHom), depending on which branches are in the migrant state in the majority of sampled ARGs. Here, the colored bars represent the fraction of Hum→Nea bases in each category for simulated data sets generated with different values of t_{mig}; the error bars show 95% confidence intervals (CIs) computed using 100 bootstrap replicates across the introgressed elements. The horizontal black lines represent the amount observed in the real data, with the gray boxes showing the CIs, also obtained by the same bootstrap process.

ARGweaver-D. We find that the observed frequency spectrum is consistent with 200kya < t_{mig} < 300kya (Fig 3.12). The same approach suggests that t_{mig} > 225kya for the for the Sup→Den event (Fig 3.13).

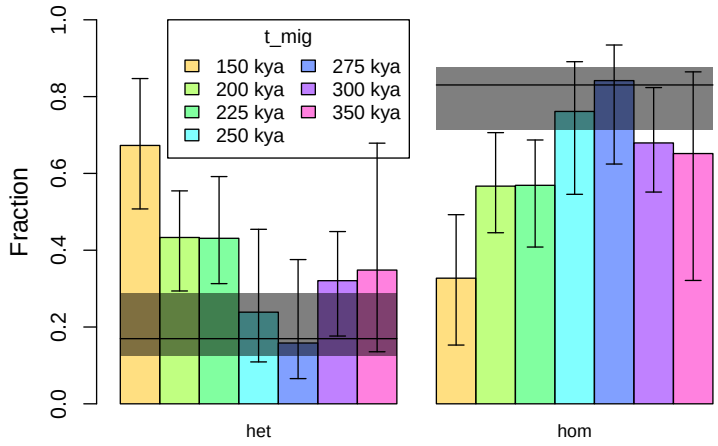

Figure 3.13: **Frequencies of Sup→Den introgression categories**. This figure is analygous to Fig 3.12; here we look at putative Sup→Den regions. Because there is only one Denisovan individual, there are only two categories: heterozygous or homozygous. Note that while we expect rates of heterozygosity to decrease with migration time, the confidence intervals here are wide, and there may be conflicting ascertainment effects that cause the apparent increase in heterozygous segments for the simulated data sets with oldest t_{mig} values.

Data release and browser tracks

Our predictions and posterior probabilities can be viewed as a track hub on the UCSC Genome Browser [64], using the URL: `http://compgen.cshl.edu/ARGweaver/introgressionHub/hub.txt`. The raw results can be found in the sub-directory: `http://compgen.cshl.edu/ARGweaver/introgressionHub/files`. Fig 3.14 shows a large region of chromosome X as viewed on the browser, with a set of tracks showing called regions, and another showing posterior probabilities. Fig 3.15 shows a zoomed-in region with a Sup→Den prediction, and Fig 3.16 shows an example Hum→Nea region. When zoomed in, there is a track showing the patterns of variation in all the individu-

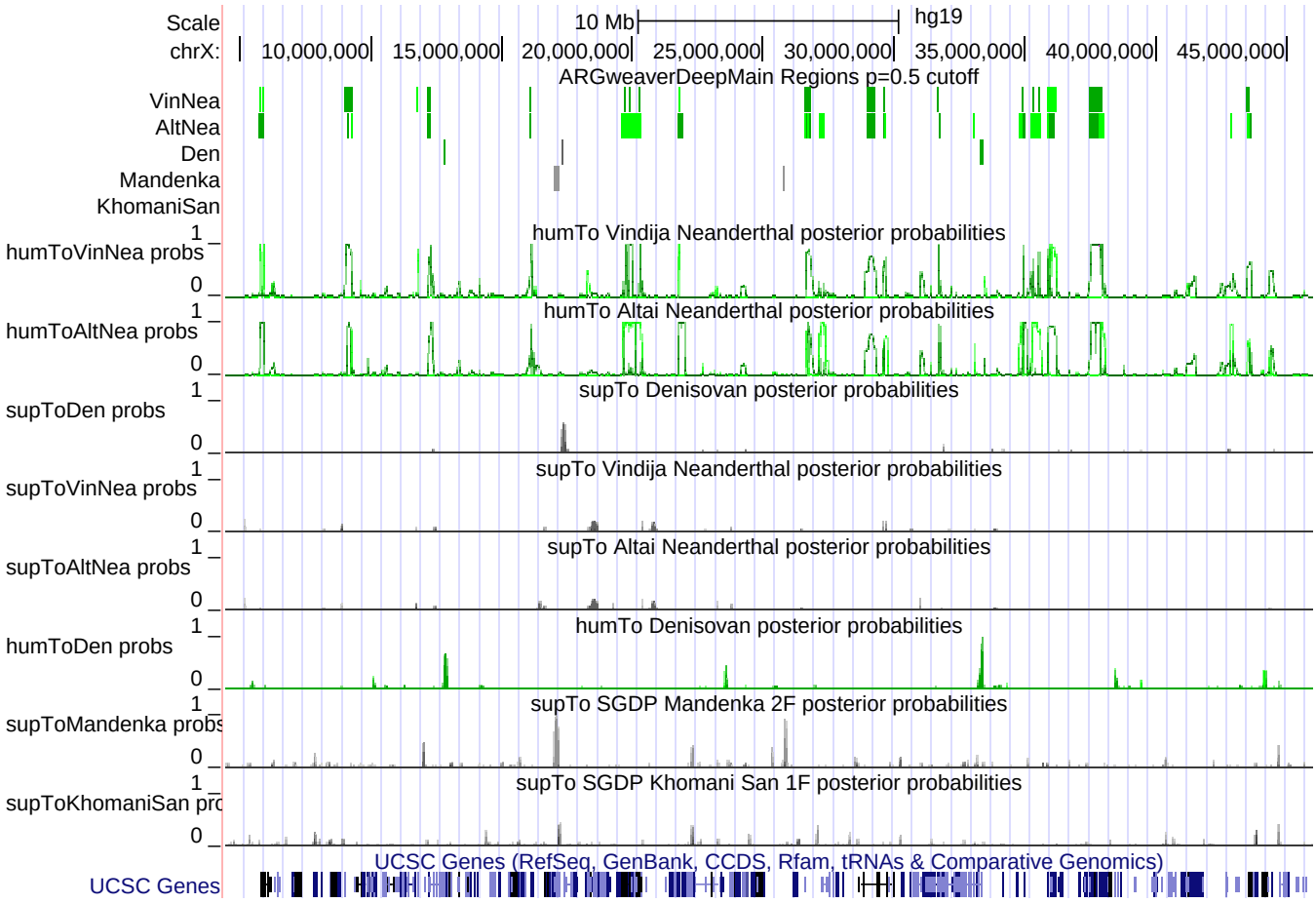

Figure 3.14: **Introgression results for a large section of chromosome X displayed on UCSC Genome browser**. The top set of tracks show predicted introgressed regions, with green indicating introgression from humans, and gray indicating introgression from a super-archaic hominin. Darker colors are used for homozygous introgression. Below that can be seen the posterior probabilities for each type of introgression into each individual.

als used for analysis, with haplotype phasing sampled from ARGweaver-D.

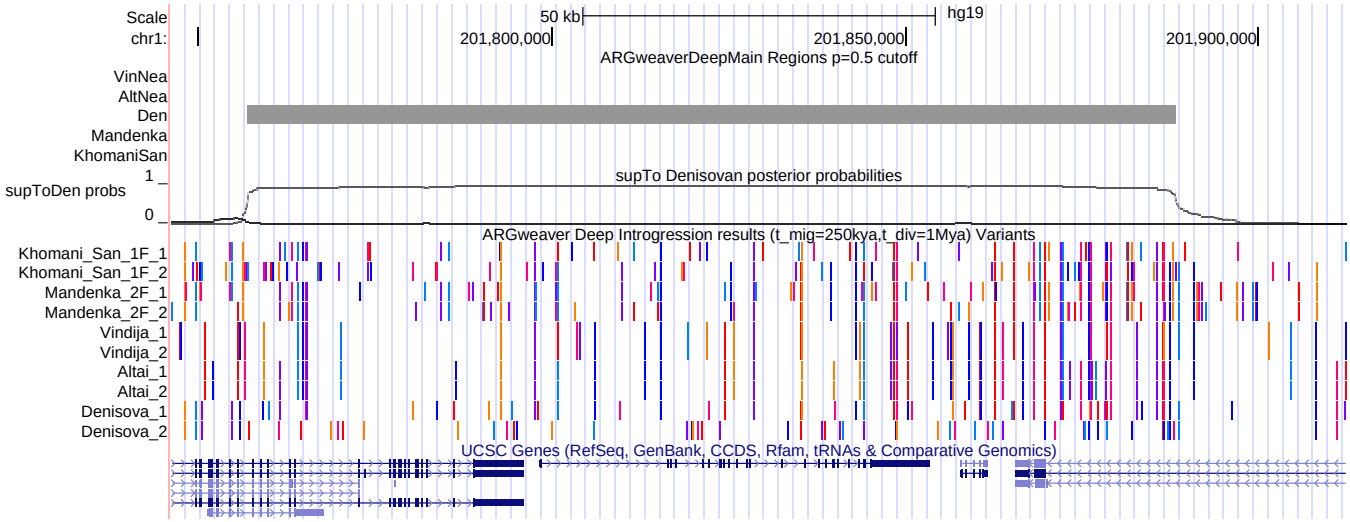

Figure 3.15: **UCSC Genome Browser shot of a region with predicted heterozygous Sup→Den introgression**. This browser shot is zoomed in on a ~ 170kb region. The first two tracks show the predicted regions and posterior probabilties as in Fig 3.14, except only the supToDen probabilities are shown. The track just below shows the variants observed in this region that are used in the ARGweaver-D analysis. Alternating colors are used for each variant site. When chimpanzee alignments are available, the non-chimp allele is colored; otherwise the minor allele is colored. Lack of a color may mean that the haplotype has the chimpanzee or major allele, or that it has missing data. The phasing of the variants represents the final phase sampled by the ARGweaver-D algorithm. Here, the Denisovan is usually homozygous and shares variants with Africans and Neanderthals outside of the introgressed region; but within it, the Denisova_2 haplotype has many singleton variants, whereas Denisova_1 continues to share many variants with Neanderthals and Africans.

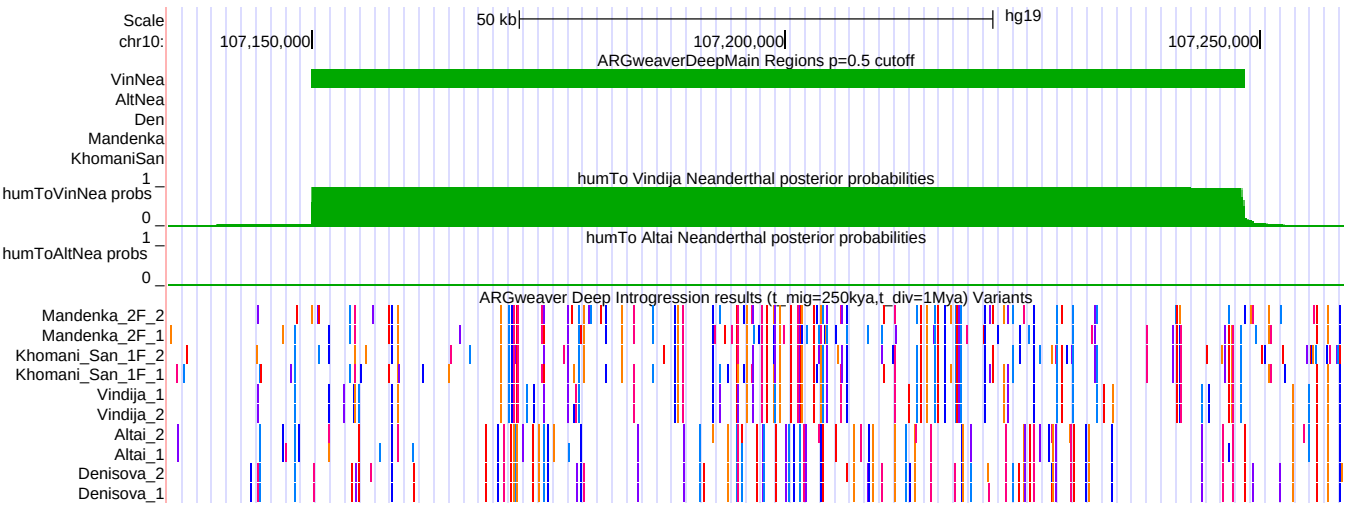

Figure 3.16: **UCSC Genome Browser shot of a region with predicted homozygous Hum→Nea introgression in Vindija**. This region on chromosome 10 has a high-probability introgressed region in both Vindija (but neither Altai) haplotypes. The top green bar indicates a predicted Hum→Nea region in Vindija, and below this is the posterior probability of introgression across the region in both Neanderthals. The variant track is similar to Fig 3.15. Here, we see almost identical haplotypes between Vindija and the Africans, whereas Altai shares haplotypes with the Denisovan.

Functional analysis of introgressed regions

Some observations in the previous section suggest that there was not strong selection against the Hum→Nea regions. We sought to look for other signals that might hint at possible functional consequences of this event.

We first looked at deserts of introgression that were detected in [45]. They noted four 10Mb deserts in which the rate of both Nea→Hum and Den→Hum introgression is $< 1/1000$. The coverage of Hum→Nea introgression within these deserts is shown in Table 3.3; the fairly high coverages suggest that these deserts are unidirectional. For two of the deserts, the Hum→Nea coverage is

Table 3.3: **Amount of Hum→Nea introgression in deserts of Nea→Hum and Den→Hum introgression.**

Location (hg19)	cov Altai	cov Vindija	num Altai	num Vindija
chr1:99-112 Mb	0.097	0.024	4	3
chr3:78-90 Mb	0.101	0.078	6	5
chr7:108-128 Mb	0.023	0.031	4	5
chr13:49-61 Mb	0.029	0.048	6	5

The columns "cov Altai" and "cov Vindija" show the coverage of Hum→Nea within the given region on each Neanderthal; "num Altai" and "num Vindija" show the number of introgressed regions > 50kb. The genome-wide average coverage for Altai and Vindija is 0.034 and 0.033, respectively.

very high, especially in the Altai Neanderthal. The third region is interesting as it overlaps the FOXP2 gene, which contains two human-chimp substitutions that have been implicated in human speech [65, 66], although the Hum→Nea introgressed region is upstream of these substitutions (Fig 3.17).

We next looked at all deserts of Nea→Hum ancestry, to see if this larger set of regions are are depleted for introgression in the other direction. Based on the CRF regions, we identified 30 regions of at least 10Mb which qualify as deserts. We looked at several statistics, including coverage of Hum→Nea in these regions, number of elements, and change in coverage between the Altai and Vindija Neanderthals; but we do not see any difference in the distribution of these statistics within deserts, as compared to randomly chosen genomic regions matched for size (Fig 3.18).

Finally, we checked for enrichments or depletions of various functional elements in our introgressed segments, relative to what would be expected if

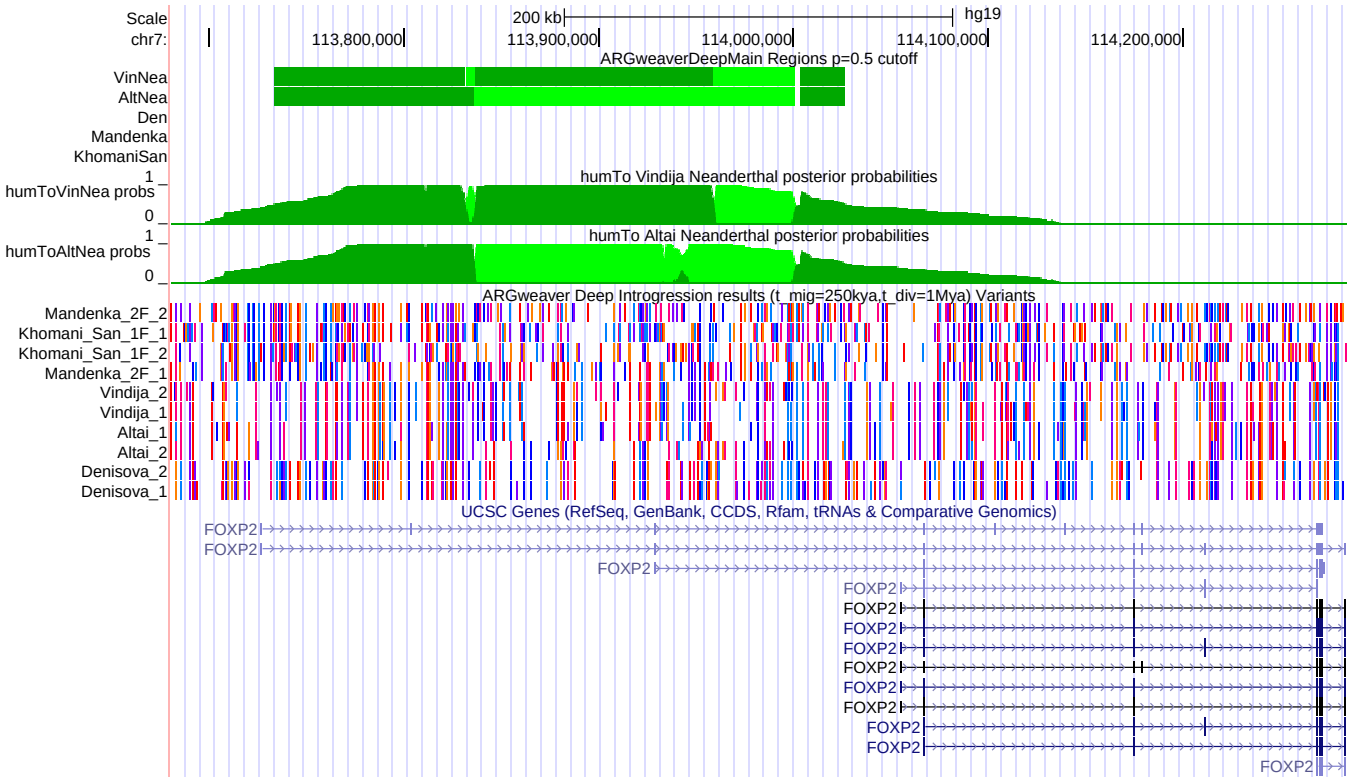

Figure 3.17: **UCSC Genome Browser shot of a predicted Hum→Nea region overlapping FOXP2.** Exon 7, which contains human-chimp substitutions shared by Neanderthals that may be involved with human speech, is located at the very right of this plot, and is not predicted introgressed. As in Fig 3.14, the light green implies heterozygous Hum→Nea introgression, whereas dark green is homozygous.

the introgressed segments were randomly distributed throughout the genome. However, the interpretation of these numbers is difficult, as local genomic factors (such as effective population size, mutation and recombination rates) affect the power to detect regions. While the overall levels of enrichment are therefore difficult to interpret, it is interesting to note that the enrichment of functional regions (such as CDS, promoters, and UTRs) tends to be higher in the Altai than the Vindija Neanderthal, which is the opposite pattern we might expect from negative selection (since the Vindija Neanderthal's fossil is much more recent).

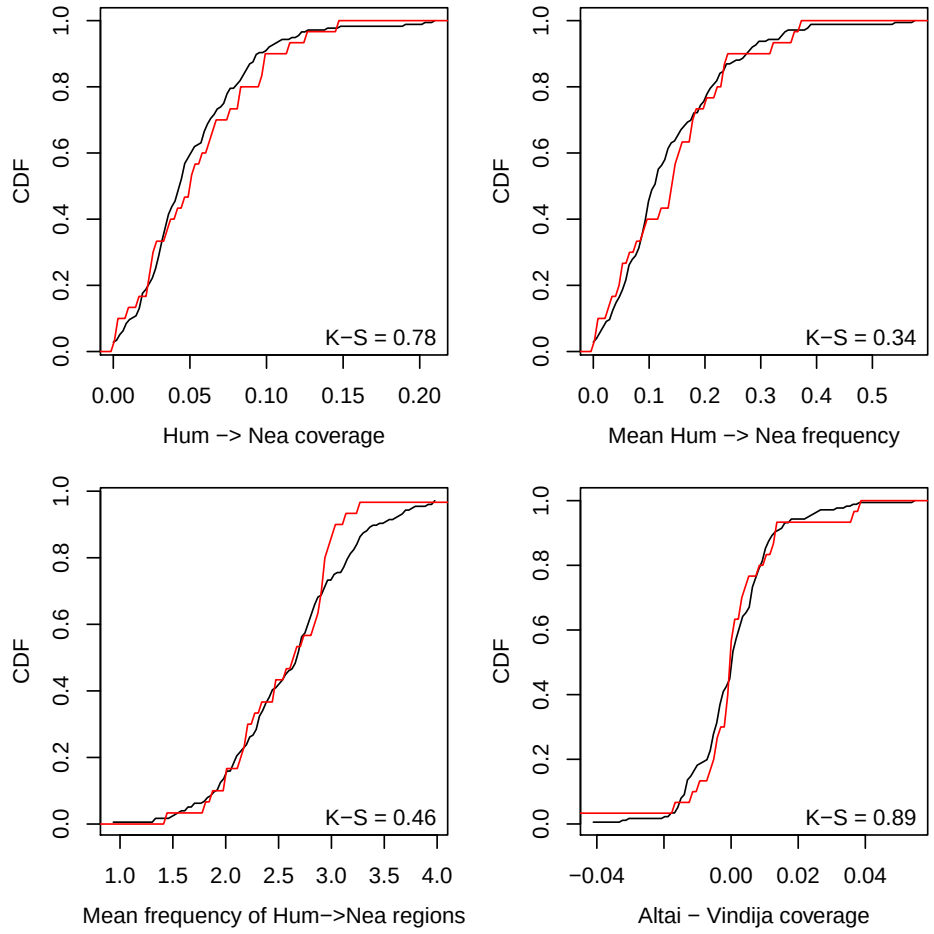

Figure 3.18: **Properties of Hum→Nea regions within Nea→Hum deserts.** We compared the distribution of various statistics across all non-overlapping 15Mb windows in the genome (black), to the distribution within deserts of Neanderthal introgression in humans of at least 10Mb (red). We excluded any window that crosses a telomere or centromere, or where ≥ 50% of the window does not pass our filters. In the bottom-right corner of each plot is shown the Kolmogorov-Smirnov statistic p-value, indicating that there is no significant difference between the black and red distributions. The statistics shown are indicated on the x-axis label. "Hum→Nea coverage" is average fraction of the window that contains any Hum→Nea region. "Mean Hum→Nea frequency" is the average number of introgressed haploid lineages of Hum→Nea across the window (where a frequency of zero indicates no introgression, and a frequency of 4 indicates homozygous introgression in Altai and Vindija). "Mean frequency of Hum→Nea regions" is the mean frequency, among regions with Hum→Nea calls. "Altai - Vindija coverage" is difference in mean coverage between the Altai and Vindija within each window.

Further enrichment results are detailed in the Supplementary Text.

3.3 Discussion

We present a new method for building ARGs under an arbitrary demographic model, and use this method as a powerful new way to identify introgressed regions. While it can detect introgressed Neanderthal and Densiovan sequences in human genomes, ARGweaver-D is too computationally complex to be used on a large scale over many human samples. However, it is very powerful even on small samples and older migration events, and has several other benefits over other methods. It does not require a reference panel of non-introgressed individuals, and can simultaneously identify introgression stemming from multiple migration events, as well as from both sampled or unsampled populations. ARGweaver-D does not rely on summary statistics, but uses a model of coalescence and recombination to generate local gene trees that are most consistent with the observed patterns of variation, even for unphased genomes. By incorporating all this information, it can successfully distinguish migration from incomplete lineage sorting, and tease apart different migration events that produce similar *D* statistics (such as Sup→Den and Hum→Nea). The code is freely available and can be applied to any number of species or demographic scenarios.

Applying this method to modern and archaic hominins, we confirm that a significant proportion of the Neanderthal genome consists of regions introgressed from ancient humans. While we identified 3% of the Neanderthal genome as introgressed, a rough extrapolation based on our estimated rates of

true and false positives suggests that the true amount is around 6%. Thus, the Neanderthal genome was likely more influenced by introgression from ancient humans, than non-African human genomes are by Neanderthal introgression. Our follow-up analysis suggests that the Hum→Nea gene flow occurred between 200-300kya. This time estimate is largely based on the frequency of introgressed elements among the two diploid Neanderthal genomes, and thus will be sensitive to the accuracy of the demographic model we used for simulation, as well as other factors such as mutation rate and generation time.

Making conclusions about the possible impact of the Hum→Nea migration has proved challenging due to the myriad ascertainment biases—known and unknown—that affect our power to detect introgressed regions. Even in the case of Nea→Hum migration, in which power to detect introgression is much higher, earlier claims of depletion near genes, as well as decreasing levels of introgression over time, have been recently called into question [46,47]. The strongest remaining pieces of evidence for negative selection against Nea→Hum introgression are the depletion on the X chromosome and several other genomic deserts. But for Hum→Nea, we see no depletion on the X, and while we do not have enough samples to detect deserts across Neanderthals, we confirm that previously identified Nea→Hum deserts are not depleted for introgression in the opposite direction. We do see a slight decrease in Hum→Nea introgression on the X chromosome in the Vindija Neanderthal compared to the Altai, which could be explained by weak negative selection removing some introgressed regions in the ~70ky that separate these fossils. An interesting question is whether this lack of selection is because human introgression introduced healthy variation into the Neanderthal genome, or because the Neanderthal population was too small for natural selection to act against anything but the most harmful variants.

However, without more archaic samples, these questions will be challenging to answer.

ARGweaver-D also identified 1% of the Denisovan genome as introgressed from a super-archaic hominin. Previous studies have estimated the total amount of Sup→Den as roughly 6% [54], but this is the first study to be able to identify specific introgressed regions. The fact that we only find a small fraction of the total amount suggests that the introgressing population was not too highly diverged from other hominins; this low power is much more consistent with a divergence time of 1Mya than 1.5Mya. Still, we report 27Mb of putative super-archaic sequence from this previously-unsequenced hominin, and we note that 15% of these regions have been passed on to modern humans through Den→Hum introgression. It may be possible to obtain more of this super-archaic sequence by applying ARGweaver-D to the set of 161 Oceanian genomes recently sequenced [67], looking for super-archaic segments passed through the Denisovans.

There have been several studies suggesting super-archaic introgression into various African poulations [55, 56, 68]. However, ARGweaver-D only detected a small amount of Sup→Afr introgression, which was somewhat lower than our estimated false positive rate. One aspect to note here is that the power to identify introgression from an unsequenced population is highly dependent on the population size of the recipient population. The larger the population, the deeper the coalescences are within that population, making it more difficult to discern which long branches might be explained by super-archaic introgression. In the case of Africans, we used a population size of 23,700, which was our best estimate from previous runs of GPhoCS [2, 10]. If we had used a smaller pop-

ulation size, ARGweaver-D would have produced more Sup→Afr predictions, but most of these would be false positives unless that smaller population size is closer to the truth. Overall, we caution that the problem of detecting super-archaic introgression into a large and structured population such as Africas is very difficult, and that claims of such introgression need to be robust to the demographic model used in analysis. It may not be possible to address the question of ancient introgression into Africans without directly sequencing fossils from the introgressing population.

We also explored some introgression events that do not have any support from previous literature; namely the Hum→Den and Sup→Nea events. *A priori*, we expected that levels predicted for these events would likely serve to confirm our false positive rates in real data. However, it is also possible that there is some amount of these types of gene flow, which has not been detected previously because it goes against the net direction of gene flow. For the Hum→Den event, we predicted a slightly smaller fraction (0.37%) than our predicted false positive rate from simulations (0.41%). For Sup→Nea, we predicted 0.75% of the genome introgressed, which is slightly higher than our predicted false positive rate for this event (0.65%). While these fractions are small, it seems entirely plausible that if there was admixture between *Homo erectus* and the Denisovans, there may have also been some with Neanderthals, perhaps in the Middle East; or genes may have passed from *Homo erectus* to Neanderthal through the Denisovans. Given the number of known interactions between ancient hominins, it may be more reasonable to assume that gene exchange likely occurred whenever these groups overlapped in time and space.

3.4 Materials and Methods

3.4.1 General ARGweaver-D settings

For all ARGweaver-D runs in this paper, the MCMC chain was run for 2000 iterations, with the first 500 discarded as burnin, and ARGs sampled every 20 steps thereafter. Except where otherwise noted, phase was randomized for all individuals and the phase integration feature of ARGweaver-D was used (`--unphased`). We used site compression throughout (`--compress 10`). We also used `--start-mig 100`, which disallows migrations for the first 100 iterations of the sampler, enabling ARGweaver-D to establish an ARG with a good general structure before exploring the migration space.

Recombination rate. Rather than use a recombination map calculated from modern humans, which may not be accurate for ancient hominins, we used a constant recombination rate of 5e-9/bp/generation for all analyses. This value was chosen for being somewhat between the mean and median genome-wide recombination rates (1.3e-8 and 1.7e-9 per bp per generation, respectively), and for providing reasonable power while still maintaining a low false positive rates in simulations (see Supplementary Text). Note that all simulated data sets were nonetheless created with a real human recombination map (see "Simulations", below).

Mutation rate. For real data analysis, the mutation rate map was based on primate divergence levels in 100kb sliding windows, using genome-wide alignments of human, chimp, gorilla, orangutan, and gibbon sequences (see Supplementary Text for details), and scaled to an average rate of 1.45e-

8/generation/site. Simulated data sets were generated by sampling rates from this map, and the same map was used for analysis.

Demographic model

The demographic model used in all ARGweaver-D analyses is depicted in Fig 3.3. The divergence times used were taken from [54], and population sizes from [2] (which were based on estimates from GPHoCS [10]). When analyzing chrX, population sizes were scaled by a factor of 0.75. When analyzing non-African humans, we only included the "recent" migration bands from Neanderthals and Denisovans into humans, whereas when looking for older introgression events, we excluded the "recent" bands as well as non-African humans.

Recall that ARGweaver uses a discrete-time model; 20 discrete times were chosen to span the range of relevant times, with more density near the leaves (where more coalescences occur) and to allow for coalescences between migration and population divergence events in the models. The discrete times (in kya) were: 0, 100, 200, 300, 400, 450, 500, 550, 600, 700, 950, 1200, 1450, 1700, 2000, 3000, 5000, 7000, 13,000, 15,000. Migration events occurred at half-time points including 50, 150, 250, and 350kya. Note that on this time scale, the European/African split is very recent, so that we did not model the population divergence among modern humans or recent growth in out-of-Africa populations. Similarly, we did not model the divergence between the Altai and Vindija Neanderthals, which are estimated to split only ~ 15ky before the Altai Neanderthal individual lived. Throughout, we assume a generation time of 29 years [69].

3.4.2 Calling introgressed regions

Once ARGweaver-D has been run, introgressed tracts can be identified for each migration event by scanning the resulting ARGs for local trees whose branches follow that migration band. Throughout this paper we use a probability threshold of 0.5 to identify introgressed regions, indicating that the region was introgressed in at least half of the sampled ARGs. To predict introgressed regions for a particular individual, we compute the posterior probability that either of the individual's two haploid lineages are introgressed. The probability of being in a heterozygous or homozygous introgressed state can be calculated as the fraction of ARG samples in which one or two lineages (respectively) from an individual are introgressed in the local tree.

The coverage of introgressed regions for an individual is computed as one-half times the coverage of heterozygous regions, plus the coverage of homozygous regions. In theory, this fails to account for sites that switch between the heterozygous and homozygous states without reaching the threshold for either, but in practice this occurs at a negligible fraction of sites.

3.4.3 Analysis of hominin data

Data preparation

We ran a series of ARGweaver-D analyses on freely available hominin data, described in Table 3.4. The panTro4 chimpanzee sequence was used as a haploid outgroup. The chimp alignment to hg19 was extracted from the alignments of 99 vertebrates with human available on the UCSC Genome Browser (`http://`

hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way). Any region which did not have an alignment for chimp is masked in the chimp sequence.

Table 3.4: **Hominin samples used in this study.**

name	region	source	ID	sex	coverage	age (ky)
Vindija Neanderthal	Europe	Max Planck	Vindija33.19	F	30x	52
Altai Neanderthal	Siberia	Max Planck	Altai	F	52x	115
Denisovan	Siberia	Max Planck	Denisova	F	31x	72
Papuan	Oceania	SGDP	LP6005441-DNA_B10	F	41x	0
French Basque	Europe	SGDP	LP6005441-DNA_D02	F	36x	0
Khomani San	Africa	SGDP	LP6005677-DNA_D03	F	44x	0
Mandenka	Africa	SGDP	LP6005441-DNA_F07	F	37x	0

We downloaded samples generated by investigators at the Max Planck Institute from: http://cdna.eva.mpg.de/neandertal/Vindija/VCF; this directory contains genotype calls for several ancient genomes using a consistent pipeline and genotype caller (snpAD) throughout.

SGDP: Simons Genome Diversity Panel.

Filtering

For each individual, we masked genotypes with quality scores less than 20 or sequencing depths outside the range $[20, 80]$. For each ancient individual, we also used the filters recommended by [54] and provided here: http://cdna.eva.mpg.de/neandertal/Vindija/FilterBed. We also masked (for all individuals): any site which belongs to a non-unique 35mer, according to the UCSC

Genome Browser table hg19.wgEncodeDukeMapabilityUniqueness35bp; "blacklisted" sites falling under the tables hg19.wgEncodeDacMapabilityConsensusExcludable or hg19.wgEncodeDukeMapabilityRegionsExcludable; ~ 9% of the genome for which SGDP genotype calls were not provided (85% of this set overlapped previously mentioned filters). ARGweaver-D was run in 2.2Mb windows, but we excluded any window for which any of the ancient filters, or the combined site filters, exceeds 50% of bases. In total we analyzed 1,166 autosomal windows and 52 windows on the X chromosome, covering 2.56Gb of hg19.

CRF calls

Introgression calls from [45] were downloaded from `https://sriramlab.cass.idre.ucla.edu/public/sankararaman.curbio.2016/summaries.tgz`. As recommended by the README contained therein, "set1" calls were used for Neanderthal ancestry in the Basque individual, whereas "set2" calls were used for Denisovan ancestry in both the Basque and Papuan, as well as for Neanderthal in the Papuan. For each individual, we took the set of regions with probability of introgression ≥ 0.5 in either haplotype.

F4 Ratio

The F4 ratio statistic F4(Altai, chimp; Basque, African)/F4(Altai, chimp; Vindija, African) was calculated across the autosomal genome, and for each individual chromosome. For the African samples, we used allele frequencies across 29 African individuals from the SGDP data set (this excludes 15 individuals with the highest Neanderthal ancestry according to [47]). For this analysis we

masked all sites that did not have a filter level (FL) field of 9 in the SGDP individuals. For the Neanderthals we used the same filters described previously.

3.4.4 Simulated data sets (deep introgression)

We performed a series of simulations to assess ARGweaver-D's ability to detect older migration events. Each simulated data set consists of a 2Mb region with 5 unphased diploid individuals and one haploid outgroup, mimicking the demographic histories and sampling dates of the individuals from the real data analysis. All simulations were produced with the software `msprime` [28].

The population tree used in the simulations is identical to the one depicted in Fig 3.3, and sampling dates correspond to the sample ages in Table 3.4. The human population size history also corresponds to the one in Fig 3.3. For the archaic hominins, we simulated a more detailed model of population size change, using piecewise-constant estimates produced by PSMC [8] and published in [54]. For the Neanderthal population history, we averaged the histories produced separately for the Altai and Vindija individuals, for the time periods when they overlap. Similarly, we averaged the Denisova and Neanderthal population size estimates during the time frame of their common ancestral population (415-575 kya).

For each data set, a random 2Mb region of the autosomal genome was chosen as a template region from which we chose recombination rates and mutation rates used to generate the simulated data. We used the recombination map estimated from African-American samples [70]. For the mutation map, we used the same map as in the real data analysis (based on primate divergence levels).

Missing data patterns were also taken from the template region; we applied the same ancient genome masks and mapability/blacklist masks to the simulated data. (We did not mimic the sequencing depth or quality score masks, which affected a relatively small fraction of sites).

Overall we produced several sets of simulations, each consisting of 100 2Mb regions. One set served as a control and contained no migration events. All other sets each had three types of migration (Hum→Nea at a rate of 8%, Sup→Den at 4%, and Sup→Afr at 0.5%). The rates of each event were chosen so as to have enough events per data set to be able to assess power, while still being less common than the non-migrant state. They were also chosen (by trial and error) to produce roughly similar levels of predicted introgression as observed in the real data. The simulated data sets varied in the demographic parameters used (migration time and super-archaic divergence time). A smaller set of additional simulations was produced with population sizes scaled by 0.75 to see how power might change on the X chromosome (see Fig 3.11).

All false positive and true positive rates were calculated basewise; separate false positive and true positive rates were calculated for each type of migration in the ARGweaver-D model. To be classified as a true positive, the method must infer the correct type of migration in the correct individual. False positives presented here were assessed using the simulated data set with no migration.

3.4.5 Simulated data sets (Nea→Hum introgression)

We also did a smaller simulation study to assess performance on the Nea→Hum event and compare performance to the CRF. Most of the settings were the same

as above, except that we sampled 86 haploid African lineages and 4 Europeans, along with the two diploid Neanderthals and a haploid chimpanzee outgroup. Demographic parameters were the same as above, except that a European population diverged from the African population 100kya and had a initial size of 2100; at 42kya it experienced exponential growth at a rate of 0.002, for a present-day population size of 37236. (These parameters were roughly adapted from [71], but modified to reflect current smaller estimates of the mutation rate in humans.) We then added 2% migration from Neanderthal into Europeans at 50kya. In some supplementary analysis we also included 5% migration from human to Neanderthal 250kya.

For this analysis only, we used true haplotype phases, in order to have a fair comparison with CRF, which assumes phased samples.

Annotations

CDS, 3'UTR, and 5'UTR annotations were taken from the ensGene (ensembl) track on the UCSC genome browser. Enhancers and promoters were extracted from the Ensembl regulatory build dated 2018-09-25. PhastCons elements came from the phastConsElements46wayPrimates track on the UCSC Genome Browser.

3.5 Acknowledgments

Thank you to Sriram Sankararaman for providing CRF software. MJH and AS were supported by US National Institutes of Health grant R35-GM127070

(to AS), and MJH was additionally supported by National Science Foundation GRFP DGE-1650441. ALW was supported by an Alfred P. Sloan Research Fellowship and a seed grant from Nancy and Peter Meinig. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The content is solely the responsibility of the authors and does not necessarily represent the official views of the US National Institutes of Health.

3.6 Supplementary Methods

3.6.1 Threading an ARG conditional on population structure

The main engine behind ARGweaver-D is the "threading" operation, which samples the coalescence points for a lineage that has been removed from the ARG. The threading operation uses an HMM in which the state space changes along the genome as the local tree changes. At a given genomic location, the state space is defined as the set of all possible coalescence points on the local tree, given by every time point along every branch. In the multiple population model, the state space is augmented with a third dimension indicating the "population path" of the new branch. Each population path is a vector of population assignments at every time point. This population path is sampled as part of the threading procedure, and retained in the new ARG, so that the full ARG defines the local genealogies as well as the population assignments for every lineage at each time point.

Recall that in the original ARGweaver model, time is discretized into $K + 1$

points, $t_0, ..., t_K$, with half-time points $t_{1/2}, ..., t_{K-1/2}$ between them. Coalescence and recombination events occur only at whole time points, based on their cumulative probabilities between the adjacent half-time points. In the multiple population model, both migration and population divergence are assumed to occur instantaneously at one of the discrete half-time points. This separates the coalescence process from the migration process, preventing ambiguities about the order of events in the ARG, and ensures that the number of lineages within a population is well-defined throughout a coalescence rounding interval.

Let each state be described by the vector (b, t, p), where b is a branch of the local tree, t is a time (looking in backwards in time; $t = 0$ is present-day), and p is a population path vector, such that p_i gives the population of path p at time point i. Each branch of the tree has its own population path, $p^{(b)}$. A state is only valid if $p_t = p_t^{(b)}$, since coalescence can only occur if the lineages are in the same population at the same time. We also assume throughout that each sample comes from a single known population (although this model could easily be extended to work for unknown or admixed samples). Therefore, for leaf branches, p_0 is fixed (or for an ancient sample with date a, p_a is fixed).

Due to these constraints, only a subset of all population paths are valid for each possible coalescent point. In the absence of migration bands, every coalescence point on the tree will be reachable by either zero or one population paths. Therefore, the size of the state space will be reduced compared to the single population model. However, as migration bands are added, more coalescence points become reachable, and some will be reachable by multiple distinct population paths. The result is that the size of the state space - and the computational complexity of the HMM - increases as more migration bands are added,

and can quickly exceed the single-population case. The cost of each migration band, in terms of run-time, depends on how many samples are in the receiving population, as well as the time of the migration band (recent migration events increase the state space more than older ones in a particular population). In order to improve both efficiency and mixing of the MCMC chain, we allow at most one migration event in any local tree in the ARG. So, "double migration" or "back migration" events are not allowed. When threading a branch into a local tree that already contains a migration, only states with non-migrating paths are valid. This assumption is reasonable when the rate of migration is low and the number of samples is modest.

Given this multiple population model, the threading algorithm proceeds similarly to the one described in the original ARGweaver paper [1]. The emissions probabilities (computed as the probability of the sequence data conditional on the local tree) are not affected by this model; nor are the probabilities of recombination at any point in the local tree. However, the probability of coalescence is now calculated conditional on the population path. The probability of the path also needs to be taken into account (as the product of migrating or not migrating as the branch passes through migration bands). Additionally, the symmetries exploited by ARGweaver for optimizing the forward algorithm also change. The algorithm takes advantage of the fact that the transition probabilities from state x_{i-1} to state x_i are not dependent on the branches assigned to each state, except for when the two branches are equal. Otherwise, only the coalescent times for each state matter, and the forward algorithm can be performed in $O(LnK^2)$ time, where L is the number of sites, n is the number of samples, and K the number of time points. In the demographic-aware model, the calculation also depends on the population path assigned to each state, so the complexity

increaes to $O(LnK^2P^2)$ where P describes the maximum number of population paths that any lineage is able to take under the specified demographic model.

Threading migrant lineages

One additional difference is that a new type of threading has been implemented for the population model. The original model has both leaf threading (which removes and re-threads a single haploid lineage from the entire ARG), and subtree threading (which removes and resamples a series of branches, both internal and leaf). We found that neither of these algorithms are sufficient to achieve good mixing of the MCMC chain when old migration events are present, because they are not able to add or remove entire migrant haplotypes in one step.

To remedy this, we have added a branch removal algorithm that focuses on lineages and time points which may potentially reach a migrant state. Recall that subtree threading uses a "branch graph" structure that is designed to choose a series of removal branches from adjacent local trees, so as to minimize constraints on how these branches need to be re-threaded to maintain consistency with the remaining ARG. Given a removal branch at site i, the choice of removal branch at site $i+1$ is often deterministic, as most branches have a single analog in the neighboring tree with the exact same set of descendants. But when a branch is involved in recombination and recoalescence, then there may be two possible analogous branches to choose from. The original subtree threading algorithm made this choice randomly, and also required a Metropolis-Hastings rejection step to correct for the differing numbers of possible un-threadings in different ARGs (which is related to the number of recombination events in each ARG).

In our modified threading algorithm, we start by randomly choosing a migration band and a haploid lineage from a population which may follow this band. (For example, we may have a migration band between the 4th and 5th time interval to represent human to Neanderthal introgression, and we would randomly choose one of the Neanderthal lineages in the tree.) At the first site, we choose the branch ancestral to our chosen lineage which crosses through the time of the migration band (whether it migrates or not). We then follow the branch graph procedure, with the additional constraint that we only choose branches which span the time of the migration. In this way, there is never a random choice to make; if a branch is split in two by recoalescence of another branch onto it, then only one of the resulting segments spans the time of the migration band. In some cases, the chosen branch may be broken by recombination, and recoalesce more recently than the migration band; in this case we go back to the default of choosing the ancestral branch of the chosen lineage which crosses the migration time.

This modified-subtree algorithm guarantees that any migration event undertaken by an ancestor to the chosen lineage is completely removed from the ARG, and helps the MCMC sampler move to likely migration states, and also prevents the MCMC chain from getting "stuck" in a migration state. Note that this procedure is agnostic to whether any migration events actually exist in the ARG, and that the choice of a lineage and migration band is independent of the current sampled ARG. Therefore, the Metropolis-Hastings acceptance ratio is always 1, and the rejection step is unnecessary.

The effect of this threading algorithm is demonstrated in Figure 3.19, which shows performance with and without the new algorithm on simulated data. The

algorithm appears to make very little difference in detecting the Nea→Hum migration event. We expect this is because for recent migration events, the leaf threading algorithm can effectively remove and replace entire introgressed regions. However, the algorithm improves the power for the deep introgression events significantly (except for Sup→Afr, which has very little power in either case). When the algorithm is not used, the power is much lower, though the false positive rate is still low. In this example at least, then, there is not a problem with getting "stuck" in a migration state, but in moving to the migration state.

All other analyses presented in this paper use this new threading algorithm. By default, when a migration model is being used, ARGweaver-D uses leaf threading algorithm for half the iterations, and equally divides the other iterations between the original and modified subtree pruning algorithms.

3.6.2 Ancient sample ages

Whereas the original implementation of ARGweaver assumed that all samples came from present day, there is now an option to specify sample ages. The option `--age-file` takes a file with the ages of ancient samples, and was used throughout this study to model the Neanderthal and Denisovan lineages. All sample ages are rounded to the nearest time point in the ARGweaver-D model.

Implementing this option required a few minor modifications to the software, including removing the implicit assumption that the distance from leaf to root is the same for all leaves. An ancient sample with age t has leaves that start at time t instead of zero, and when threading this lineage, the only valid

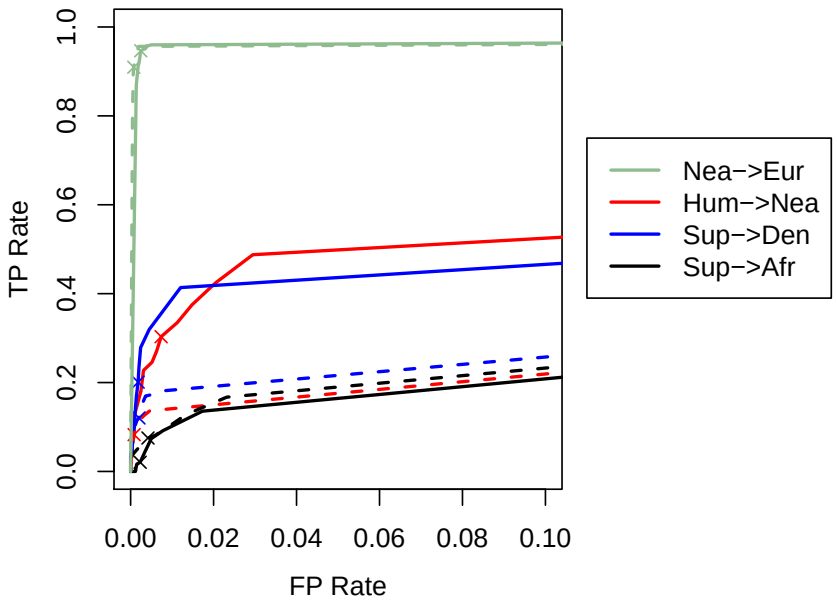

Figure 3.19: This ROC plot shows performance with (solid) and without (dashed) the modified subtree pruning algorithm. The Nea→Eur lines come from simulations in the main paper with a single migration at 50kya. The other lines come from "deep introgression" simulations with migration at 250kya and super-archaic divergence of 1Mya, and analyzed with the same model. The "x" on each line represents the performance when a posterior probability cutoff of 0.5 is used.

states are ones with ages $\geq t$. The code also had to be altered to ensure that lineages for an ancient branch do not contribute to coalescence probabilities in the time range $(0, t)$. Whereas the number of branches usually only decreases looking backwards in time, with ancient samples, branches may come into existence and the branch count can increase.

We did a simple simulation study to demonstrate that this option works as expected. Using `msprime` [28], we simulated 10 haploid lineages from a population of size 10000 across a 2Mb region. Four of the samples were modern day (sampled at $t = 0$); the other six were sampled at increasingly ancient times

(50kya, 100kya, 150kya, 200kya, 250kya, 300kya). We used a recombination rate of 1.5e-8/recomb/bp/generation and a mutation rate of 2.5e-8/bp/generation. For this demonstration we treated samples as haploid with known phase. We then compared properties of ARGs inferred with the `--age-file` option, to ARGs inferred without the option, and also to the true known ARG. Figure 3.20 shows that this option effectively corrects for biases in the estimates caused by ignoring sample ages. It also shows that, even in this fairly extreme example with several very old samples, many statistics of the ARG are not too badly skewed even when the sampling ages are not properly taken into account.

3.7 Supplementary simulation results

3.7.1 Out-of-Africa simulations with Hum→Nea

In addition to the out-of-Africa simulations with Nea→Eur migration presented in the main paper, we performed a second set of simulations that also included a true Hum→Nea migration event at 250kya at a rate of 0.02. We ran both CRF and ARGweaver-D on this data set. We tried two different ARGweaver-D models, one with a single Nea→Eur migration band at 50kya, and one that also included a Hum→Nea band at 250kya. The introgression predictions are summarized in Figure 3.21. When ARGweaver-D has only one band, it performs similarly to CRF, though with a somewhat lower false positive rate. Both methods mis-classify a small fraction of Hum→Nea regions as Nea→Eur (9.0% for CRF and 7.6% for ARGweaver-D). When a second band is added to the ARGweaver-D model, the true positive rate (34%) for correctly identifying Hum→Nea is

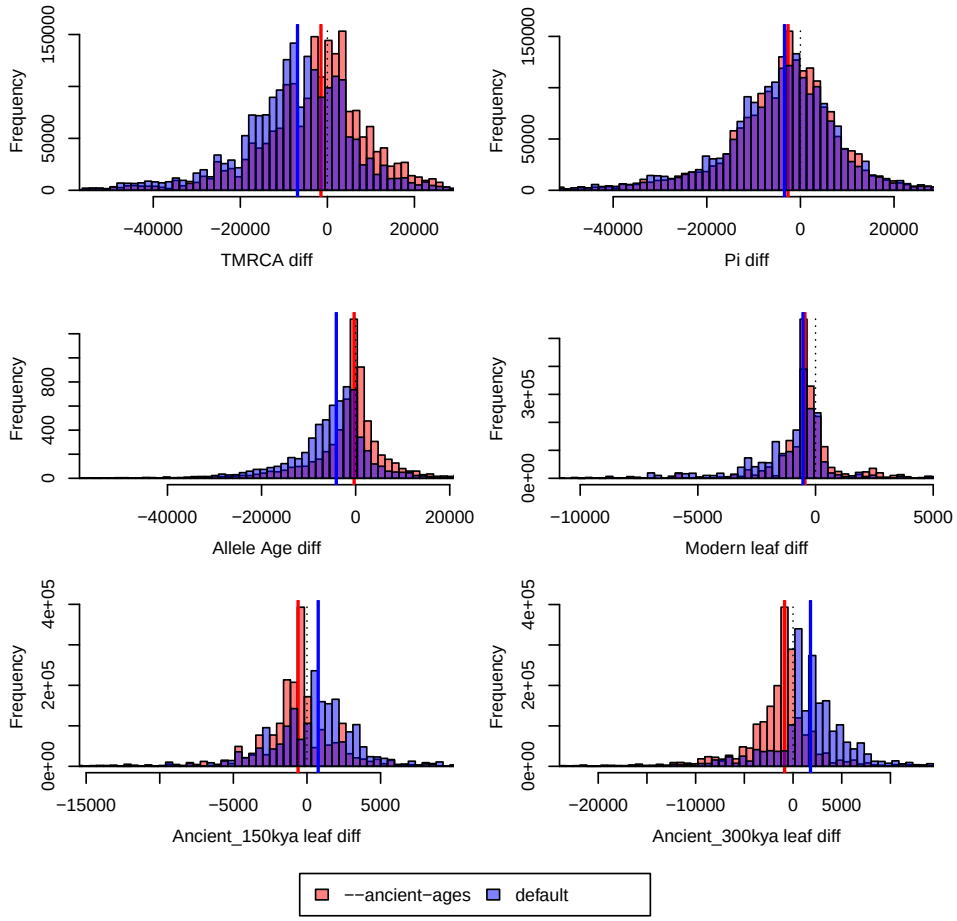

Figure 3.20: Looking at the ARG across a 2Mb region, at every base we compute the difference between the true statistic and the median from ARGs sampled across 2000 MCMC iterations. The pink distribution shows the ARGs inferred while accounting for ancient sampling dates; the blue uses the default parameter. (Purple is the overlap between the two). The dotted black line is at $x = 0$, and the red and blue lines are at the medians of the pink and blue distributions. The statistic for each plot is named in the x-axis, and the names are as follows: TMRCA (time to most recent common ancestor, in generations); Pi (average distance between two leaf nodes, in generations); Allele age (age of derived alleles); Modern leaf (coalescence time of a leaf node for a present-day sample); Ancient_150kya leaf (coalescence time of a leaf sampled 150kya), and Ancient_300kya leaf (coalescence time of a leaf sampled 300kya).

identical to what we saw in the "deep introgression" simulations presented in the main paper. However, a large fraction (38%) of the Nea→Eur regions are incorrectly classified as Nea→Hum. We are not sure why the misspecification is so much higher in one direction than the other. However, we have observed in general that getting the correct directionality when there are migrations between sister populations is difficult. This is one reason for excluding non-African populations when looking for older migration events.

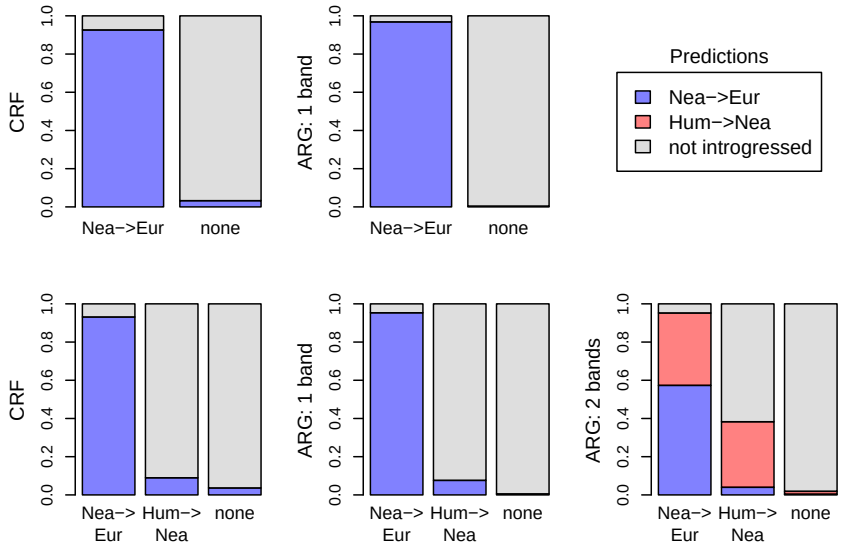

Figure 3.21: On the top is shown predictions for a set of simulations with only Nea→Hum introgression. Each bar represents a true (known) category; the colors show predictions for this category using the CRF (left) and ARGweaver-D with a single migration band (right). On the bottom are results where there are two true migration events. Here there are two sets of ARGweaver-D results; one with only the Nea→Hum band, and one that also has a Hum→Nea band.

Recombination rate analysis

Most of the analysis in this paper was done assuming a constant mutation rate across the genome. While this is unrealistic, little is known about the recombination map in Neanderthals and Denisovans (and nothing is known about possible super-archaic recombination maps).

To explore the effects of recombination rate misspecification, we compared the performance on our "deep introgression" simulations with four different recombination settings: one with the true map used in the simulations, one with a constant rate of 5e-9/recomb/generation/base-pair, another with a higher rate of 1e-8/recomb/generation/base-pair, and one with an incorrect recombination map. All simulations were created with a true recombination map sampled from some a random region of the human recombination map generated from African-American samples [70]; the incorrect maps were sampled from a different random region of the human genome. The results are shown in Figure 3.22.

Overall, the recombination rate seems to have the biggest impact on the false positive rate. Except for in Sup→Afr, where the power is low everywhere, the best performance is when the true recombination rate is used, and the worst are when a too-high or wrong rate is used. Using a constant rate of 5e-9 gave intermediate performance, and fewer false positive than other incorrect maps. Because hotspots identified in human may not apply to Neanderthal, Denisova, or super-archaic hominins, we chose to use the low recombination rate.

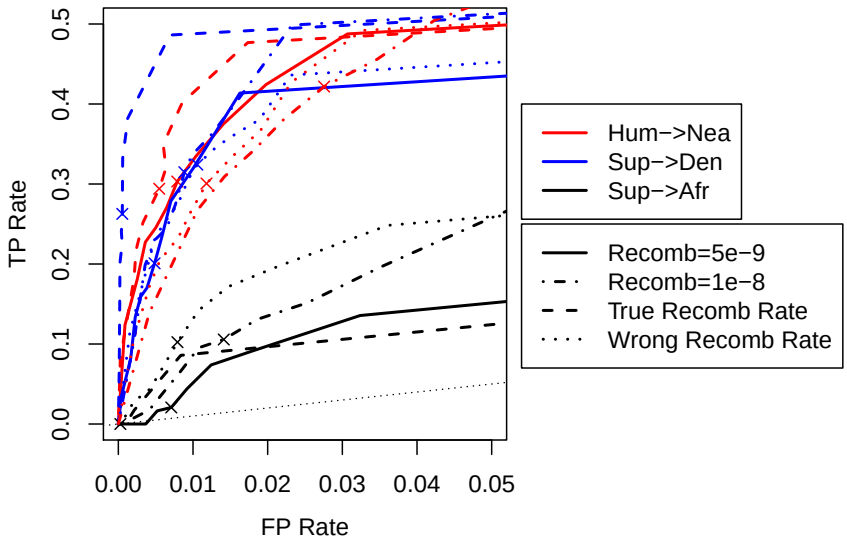

Figure 3.22: Performance on the "deep introgresion" simulations, in which the only difference between ARGweaver-D runs is the recombination map used for analysis.

Number of African samples

In theory, larger numbers of African samples might be expected to improve performance for finding Hum→Nea introgression events, as the introgressing ancient human population is most closely related to modern Africans, and seems to be equally related to various diverged African populations [2]. It is also possible that more African samples could boost the power of detecting Sup→Den regions, as they contribute more information about the ancestral archaic population. However, in practice we observed that power using 2 Africans was similar to using 4 or even 8 African individuals (Figure 3.23). We are not sure why this is, but suspect that the MCMC sampler does not mix as well as more individuals are added. Because ARGweaver-D is faster with fewer individuals and it does not seem to have much effect on performance, we did our main analysis with two Africans.

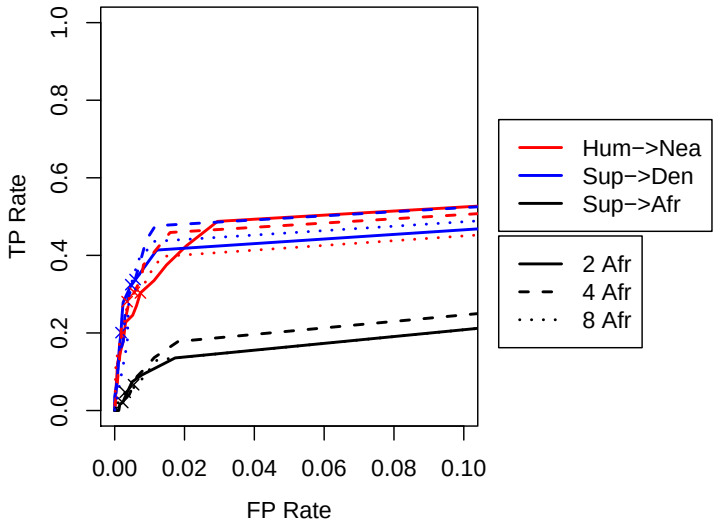

Figure 3.23: Effect of using more African individuals in the analysis. Here, we used the same "deep introgression" simulations as in the main paper, but sampled and used more African individuals in the ARGweaver-D analysis. Both the simulated data and the ARGweaver-D model had a migration band at 250kya and super-archaic divergence of 1Mya. The "x" on each line represents the performance when a posterior probability cutoff of 0.5 is used.

3.8 Supplementary analysis

3.8.1 Lengths of real vs simulated introgressed regions

The lengths of introgressed regions should be informative for the time of migration. However, there is more power to detect longer regions, creating a strong ascertainment bias that makes interpretation of the lengths difficult. The rate of recombination is also an important factor affecting the distribution of lengths, and the recombination rates in Neanderthal and Denisovan are not well characterized. Still, we looked at the distribution of lengths in our predicted set, and compared to both the true and predicted regions in simulations with different

migration times.

First, we wanted to see whether there is a difference in the length distribution of false positive regions compared to true ones.

We hoped that the lengths of predicted introgressed Hum→Nea elements might be informative for the time of migration. However, there is more power to find longer elements, and this ascertainment bias is so strong that information about the timing of migration is almost completely lost. Overall, our predicted elements were somewhat longer (median 105kb) than those observed in our simulations (median lengths ranging from 71kb-89kb), but the distributions were largely overlapping (Supp Fig 3.24). While all the ARGweaver-D analysis was done with a simple constant recombination model, the underlying recombination map is also an important factor, and little is known about the Neanderthal recombination map.

3.8.2 Validation of super-archaic regions in SGDP individuals

We further explored the 27Mb of the genome which was putatively identified as Sup→Den. This category had the strongest prior evidence for super-archaic introgression [16,54], and is the only super-archaic category for which the amount detected by ARGweaver-D (1%) significantly exceeds the false positive rate estimated from simulations (0.5%). We first identified variants in Sup→Den regions that map to the migrant lineage in our data set (which included the Denisovan, two Neanderthals, SGDP individuals Khomani_San_1 and Mandenka_2, and chimpanzee); there are 15,470 variants over 16.8Mb of unmasked Denisovan sequence. This suggests an average substitution rate is 9.2e-4/bp, which translates

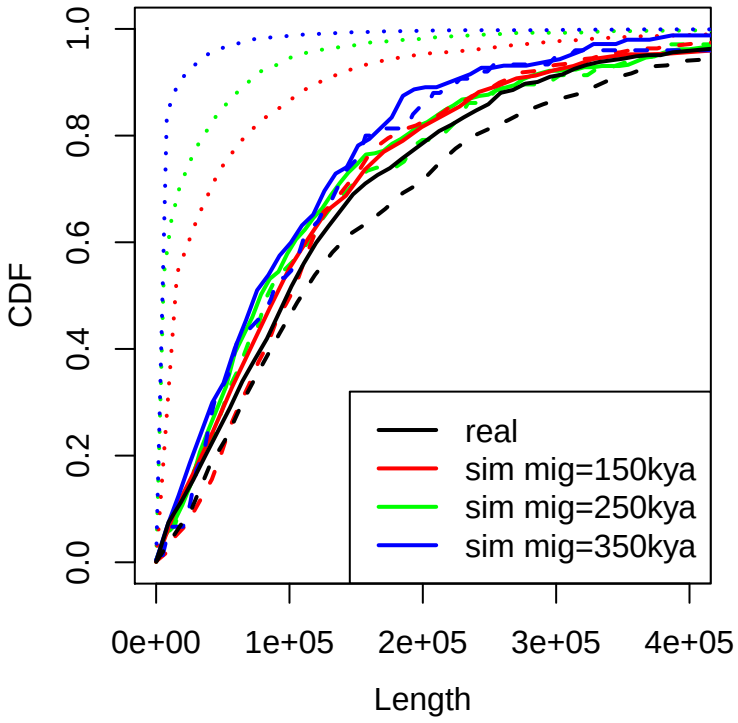

Figure 3.24: The distribution of lengths in real vs simulated Hum→Nea regions. The dotted lines show the true distribution of lengths in three simulated data sets, each produced with different migration times, indicated by the color. The solid lines show the distribution of regions found by ARGweaver-D in the simulated and real (black) data sets when analyzed with a model with a migration band at 250kya. The dashed lines show the same except with a migration band at 150kya.

to a branch length of 1.8My (using a mutation rate of 1.45e-8/bp/generation and a generation time of 29 years). However, we expect that this estimate is biased upwards, as Sup→Den regions with more variants are easier to detect. We compiled a VCF file containing all the substitutions on the super-archaic haplotype.

We next looked at all 279 individuals in the SGDP data set, comparing their divergence to Neanderthals and Denisovans in each region. If the Sup→Den

prediction is correct, then the Denisovan divergence should be high for all humans, and not just the two humans used in the ARGweaver-D analysis. Two example regions are shown in Figure 3.25. As described in the caption, and explored further below, most Sup→Den regions do indeed show higher divergence to the Denisovan across the SGDP individuals than to the Neanderthal, excepting individuals with Den→Hum introgression in that region. There are a small number of Sup→Den regions, such as the one shown in Figure 3.25B, where the Denisovan divergence does not exceed Vindija divergence in most humans and might be false positives of the ARGweaver-D approach.

While looking at example plots is helpful, we want to summarize the properties across all Sup→Den regions. We define a statistic f, which is the fraction of SGDP individuals in a given region for which the Denisovan divergence is greater than the divergence to either Neanderthal. This statistic can be visualized as the fraction of individuals that fall above the diagonal in plots such as those in Figure 3.25. In each region we exclude any individuals with Neanderthal or Denisovan introgression (as assigned by the CRF [45]). We computed f for each of the 161 putative Sup→Den regions with length $\geq$ 50kb, as well as for 262 putatitve Sup→Nea regions, 384 putatitve Sup→Afr regions, and 500 100kb regions randomly selected from regions of the genome without any ARGweaver-D introgression assignment. The distributions of f for each set of regions are shown in Figure 3.26. We see that for about 80% of the Sup→Den regions, f is close to 1. There are 29 Sup→Den regions with $f < 0.9$ (including the region shown in Figure 3.25B), and which might be best regarded as false positives.

For the Sup→Nea regions, where we would expect most individuals to have

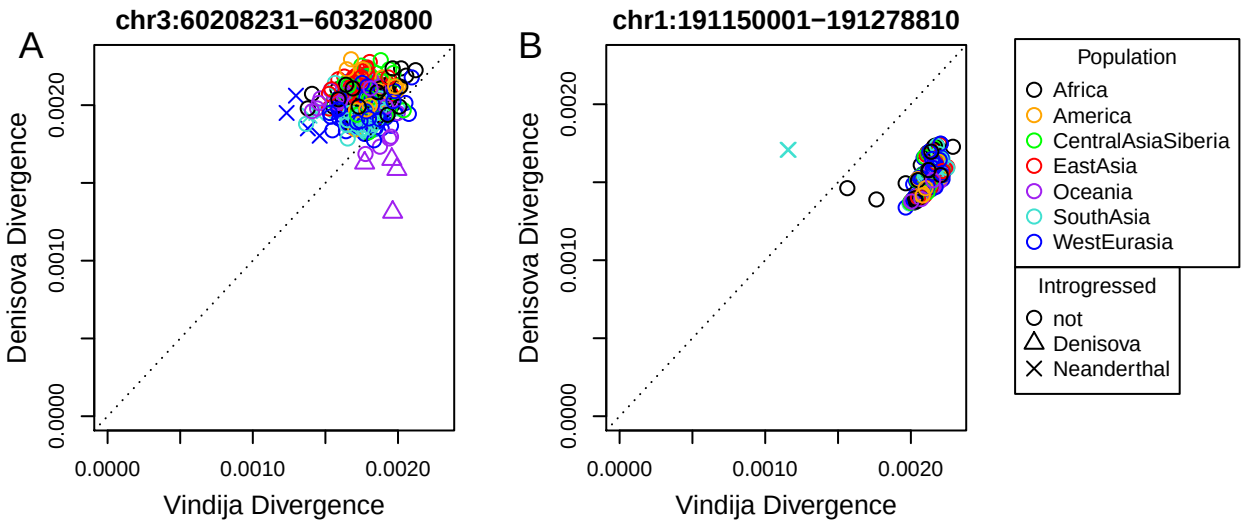

Figure 3.25: **Average divergence of SGDP individuals to Neanderthal and Denisovans in example Sup→Den regions** The color represents the population of each sample, whereas the symbol indicates whether introgression has been detected in this individual in at least half of the plotted region (according to the CRF method). A) This region shows the expected pattern for Sup→Den: most individuals have higher divergence to the Denisovan than to the Vindija Neanderthal. However, a few Oceanian individuals who have Denisovan introgression in this region have lower Denisovan divergence. Similarly, some European individuals with Neanderthal introgression also show a decreased Neanderthal divergence. B) This is a less typical Sup→Den region that is likely a false positive, as most SGDP individuals show lower divergence to the Denisovan than to Neanderthal. It is interesting that one of the outlying African dots represents an individual used in the ARGweaver-D analysis (Khomani_San_1).

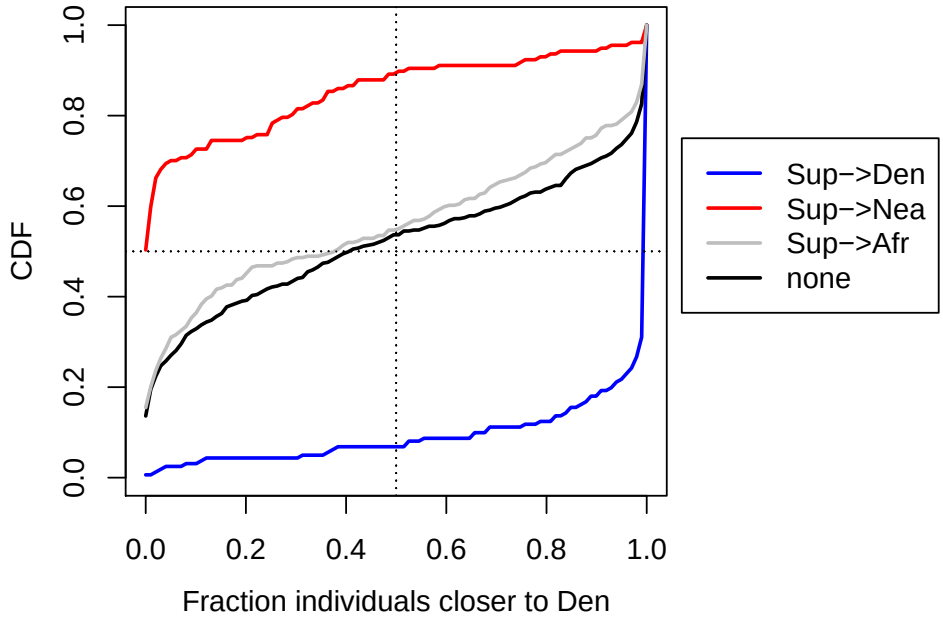

Figure 3.26: **Distribution of the fraction of individuals with higher Denisovan vs Neanderthal divergence**. Each cumulative distribution shown here is taken across a set of regions assigned to a particular introgression category by ARGweaver-D. For each region, we calculate the number of individuals for which the Denisovan divergence is higher than the Neanderthal divergence (excluding individuals with calls of Neanderthal or Denisovan introgression by the CRF). We see that Sup→Den regions have a high proportion of individuals more closely related to Neanderthal, and the opposite pattern in Sup→Nea regions. Both putative Sup→Afr and non-introgressed regions are very slightly biased towards Neanderthal ancestry.

higher Neanderthal than Denisovan divergence, we see a similar large shift towards small f; most SGDP individuals are closer to Denisovans than Neanderthals in these regions. In this case, 73% of the regions have at least 90% of individuals closer to the Denisovan.

It is important to note that while this analysis provides a check on

ARGweaver-D's predictions, and identifies some potential false positives, it does not imply that the remaining regions are true positives. Other scenarios, such as balancing selection, could also produce long regions of high divergence that may be virtually indistinguishable from super-archaic introgression. But this analysis does show that the signal identified using only two humans usually holds across a much larger set.

3.8.3 Analysis of Sup→Den regions passed to modern humans

Presumably, if there is super-archaic introgression into Denisovans, and later Denisovan introgression into Oceanian and Asian humans, then it seems likely that these modern humans harbor super-archaic alleles passed through the Denisovans. Indeed, 15% of our Sup→Den regions overlap regions with Den→Hum introgression calls in the SGDP (24 out of 161 regions, excluding regions with lengths <50kb). We looked into this by comparing the variants on the super-archaic lineage with those observed in individuals with Denisovan introgression (according to the CRF calls). Figure 3.27 shows the fraction of shared Denisovan variants vs. the number of hg19/Denisovan variants for individuals that are annotated with Hum→Den introgression by the CRF method.

The black points in Figure 3.27 show the fraction of shared variants in regions without any ARGweaver-D introgression calls in Africans or archaics. We see that the fraction of shared alleles is high (between 60-100%) for these regions, though the overall number of variants is moderate. The blue points show the same values in regions that have been identified as Sup→Den introgressed in both Denisovan lineages. For the most part, we also see high fraction of shared

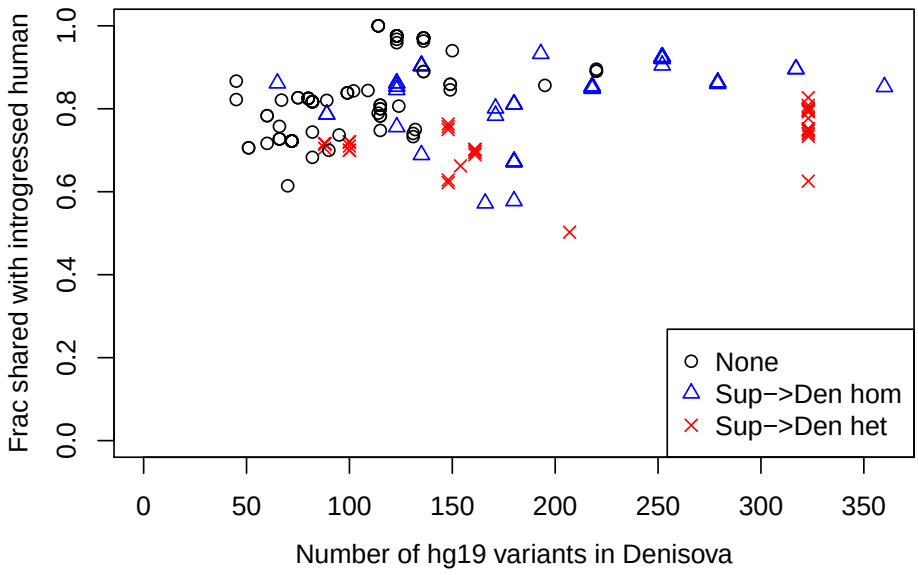

Figure 3.27: **Fraction of shared variants with Denisovan, for individuals with Denisovan introgression.** Each point is calculated for a particular genomic region and individual with Denisovan introgression in that region. The x-axis shows the number of Denisovan hg19 differences; the y-axis shows the fraction of these variants shared with the individual. The colors represent the type of region; blue regions are homozygous Sup→Den regions, red regions are heterozygous Sup→Den regions, and black are regions without any ARGweaver-D introgression calls in Africans or archaics.

variants, although the absolute number of variants is much higher overall. This suggests that the individual is sharing super-archaic alleles, as the majority of these variants occur on the super-archaic branch. Finally, the red points show a subset of Sup→Den where the super-archaic introgression is only found in one of the Denisovan lineages, so that our sampled Denisovan has both a "super-archaic" and a "Denisovan" haplotype. The red points with the lowest fraction shared may represent individuals who received the Denisovan haplotype.

One consideration here is that there is likely a bias towards identifying

Denisovan introgression in humans when the Denisovan and human both share the super-archaic haplotype, because introgression will be very easy to detect with such a large number of shared variants. This may explain why there are not many red points with lower fractions of shared Denisovan variants in Figure 3.27. Regardless, this analysis shows that many individuals with Denisovan introgression share alleles that are predicted introgressed into Denisovan from a super-archaic hominin.

3.8.4 Analysis of Sup→Nea regions passed to modern humans (and the hg19 reference sequence)

We did a similar analysis on regions identfied as Sup→Nea, this time looking at the overlap between these regions and Nea→Hum regions in SGDP humans. 35% of our Sup→Nea regions overlap regions with Nea→Hum introgression according to the CRF predictions (55 out of 157 regions, excluding regions with length <50kb). Figure 3.28 summarizes these regions; there are many more points in this plot than in Figure 3.27 because there are many more SGDP samples with Neanderthal introgression.

The first surprising aspect of these results is that there was one region (chr6:8450001-8563749) classified by ARGweaver-D as Sup→Nea, but which had only 13 hg19 differences across 79kb of unmasked Neanderthal sequence (giving an hg19/Neanderthal divergence of only 0.016%). After closer inspection, we suspect that this region has Neanderthal introgression in the hg19 reference sequence. Among SGDP individuals with annotated Neanderthal introgression in this region, there are between 1 and 18 homozygous hg19 dif-

ferences. Among the other SGDP individuals, there are between 68 and 369 homozygous hg19 differences. (This excepts one individual, Saharawi_1, fosr whom CRF introgression calls were not made and which has only one hg19 difference; we presume it is also Neanderthal introgressed. The Saharawi are a Northern African population that has been found to have almost as much Neanderthal introgression as non-African populations [54]). Given that the reference sequence is largely African-American, we should expect that it would contain some Neanderthal ancestry. While the theory of possible introgression from super-archaic introgression into Neanderthal does not yet have strong support, if the annotated Sup→Nea regions are correct, this would be an example of archaic hominin ancestry in the hg19 reference sequence, passed through Neanderthals. We also note that we found one other region (chr10:88106371-88206370), from our set of 500 randomly selected 100kb regions without ARGweaver-D introgression calls, in which the number of hg19 differences is < 4 for SGDP individuals with Neanderthal introgression, and much higher (median: 75) for other SGDP individuals. As our analysis in this section only spanned 2.4% of the genome, a genome scan for Neanderthal introgression on hg19 would discover many more such regions.

Beyond the observation of Neanderthal (and possibly super-archaic) ancestry in hg19, Figure 3.28 shows that there are indeed a number of regions annotated as Sup→Nea with a large number of hg19 variants, which are also shared to a high degree with Neanderthal-introgressed humans in the SGDP. Again, we see that the red points (representing non-fixed Sup→Nea regions) sometimes share fewer variants with Neanderthal, suggesting that these individuals are introgressed with the 'Neanderthal' haplotype rather than the 'super-archaic' haplotype.

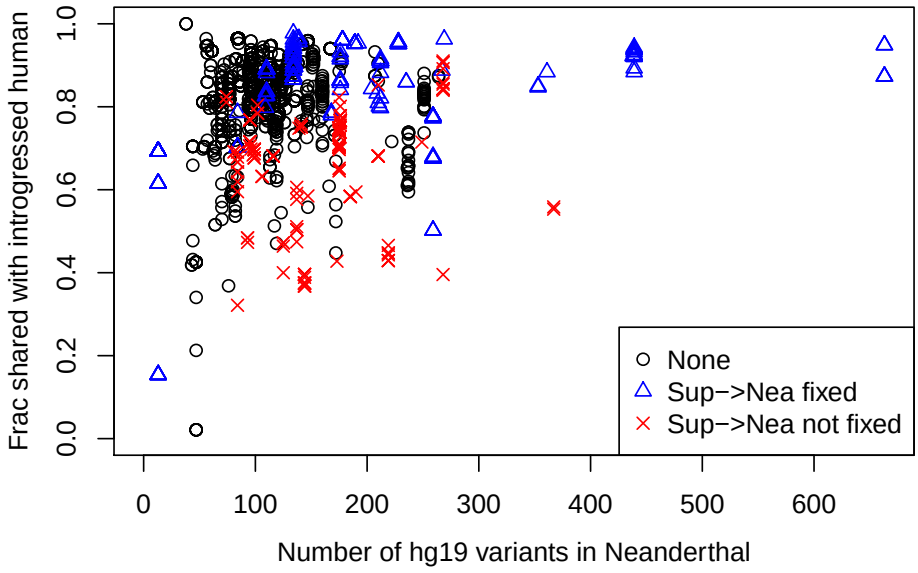

Figure 3.28: **Fraction of shared variants with Neanderthal, for individuals with Neanderthal introgression.** Each point is calculated for a particular genomic region and individual with Neanderthal introgression in that region. The x-axis shows the number of Neanderthal hg19 differences in the region; the y-axis shows the fraction of these variants shared with the individual. The colors represent the type of region; blue regions are Sup→Nea regions, red regions are a subset of Sup→Nea that are not fixed in our Neanderthal sample, and black are regions without any ARGweaver-D introgression calls in Africans or archaics.

Overall, the analysis of both Sup→Den and Sup→Nea regions show that these regions have a high number of variants compared to non-annotated regions, and that these variants are often shared with humans with introgression from Denisovans or Neanderthals (respectively). While the Sup→Nea event is not well supported, and the Sup→Nea regions may simply be highly diverged Neanderthal regions, there is stronger support for the Sup→Den migration, and it seems that humans with Denisovan ancestry must also harbor some variants from more diverged hominin species as well.

3.8.5 Functional enrichment analysis of introgressed regions

We checked for enrichments or depletions of various functional elements in our introgressed segments. However, the interpretation of these numbers is not straightforward, as the power to detect the segments is confounded by many factors, such as local population size, mutation rate, recombination rate, and sequence quality filters. Indeed, even for introgression into humans where the detection power is quite high, these depletions have been difficult to interpret (see Discussion). The enrichments are shown in Supp Fig 3.29. It is clear that biases for functional elements depend on the type of introgression event (from a sample population or super-archaic). We found a 1.15x enrichment of ensembl CDS regions in our Hum→Nea calls, which is most likely explained by higher power with lower effective population size. Perhaps the most interesting aspect is that the enrichment Hum→Nea in most functional categories (including CDS, phastCons, promoters) is larger in the Vindija Neanderthal than the Altai, despite the fact that the Vindija and Altai Hum→Nea regions are highly overlapping (55.6% of the combined set are called in both individuals). Again, it appears that functional elements were not lost in the duration between the Altai and Vindija individual's lifetimes, suggesting an absense of negative selection acting against these regions.

3.8.6 Deep introgression analysis with other models

In the main paper, we presented results using an ARGweaver-D model in which all migration times (t_{mig}) were set to 250kya and super-archaic divergence time (t_{div}) to 1Mya. We also ran ARGweaver-D genome-wide with t_{mig} =150kya and

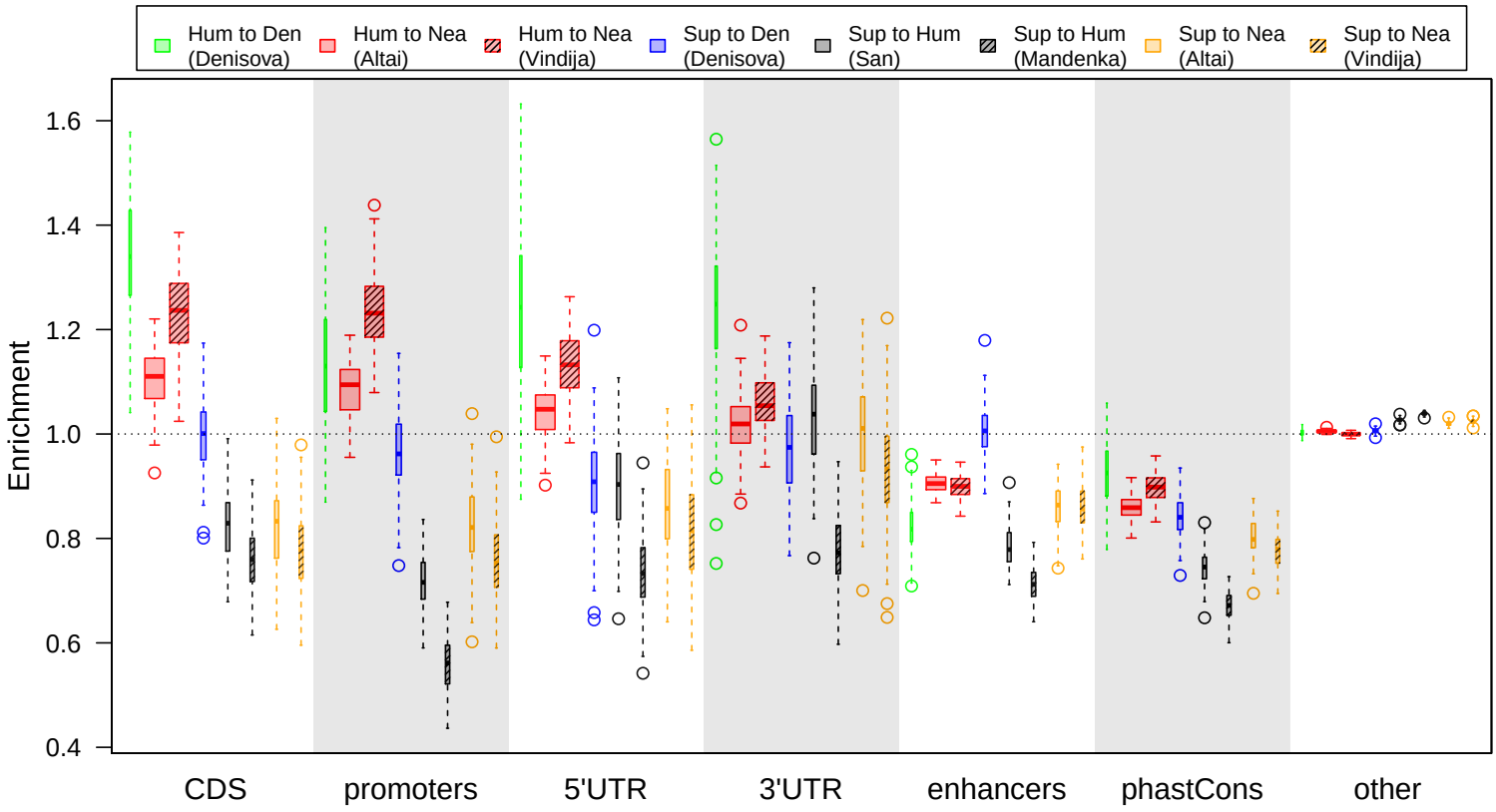

Figure 3.29: **Enrichment of predicted introgressed regions within different annotation groups.** Enrichment is calculated as the number of introgressed bases called within a particular category, divided by the number expected assuming that the introgression calls are independent of the annotations. The distributions shown in the box plots are calculated from 100 bootstrap replicates over the introgression calls. The width of each bar is proportional to the total coverage of the predicted introgressed elements genome-wide. These enrichments are likely highly influenced by factors affecting the power and false positive rates of the calling algorithm (See discussion in Supplementary Section XX).

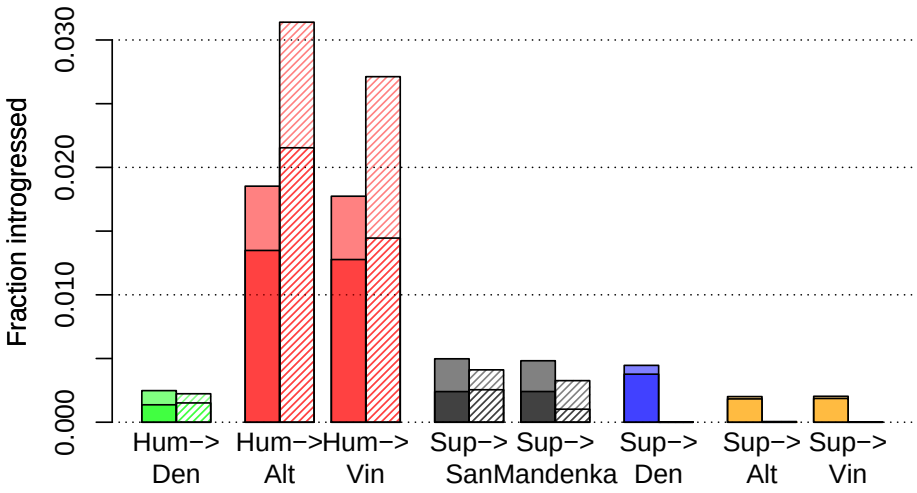

Figure 3.30: **Introgression coverages under alternative demographic model**. Genome-wide coverages using a migration time of 150kya and a super-archaic divergence of 1.5Mya. Solid bars show autosomal coverage; striped bars show X-chromosome coverage. The dark bottom halves of each bar represent regions that were also predicted by the model used in the main paper.

t_{div}=1.5kya. The coverage of the resulting elements is shown in Figure 3.30.

The results using this model are qualitatively similar to those presented in the main paper; the largest coverage is in Hum→Nea, with increased coverage on the X chromosome, and a somewhat smaller depletion from Altai to Vindija on the X. We again see similar low levels of all other introgression categories, and the same depletion for Sup→Den on the X chromosome. The coverages of predicted introgressed Hum→Nea and Sup→Den elements are about half what is presented in the main paper. This is consistent with our simulation results which show that power is much lower using this model when the true model more closely matches the one used in the paper, and further supports our claims that the Hum→Nea migration was quite old, and that the super-archaic

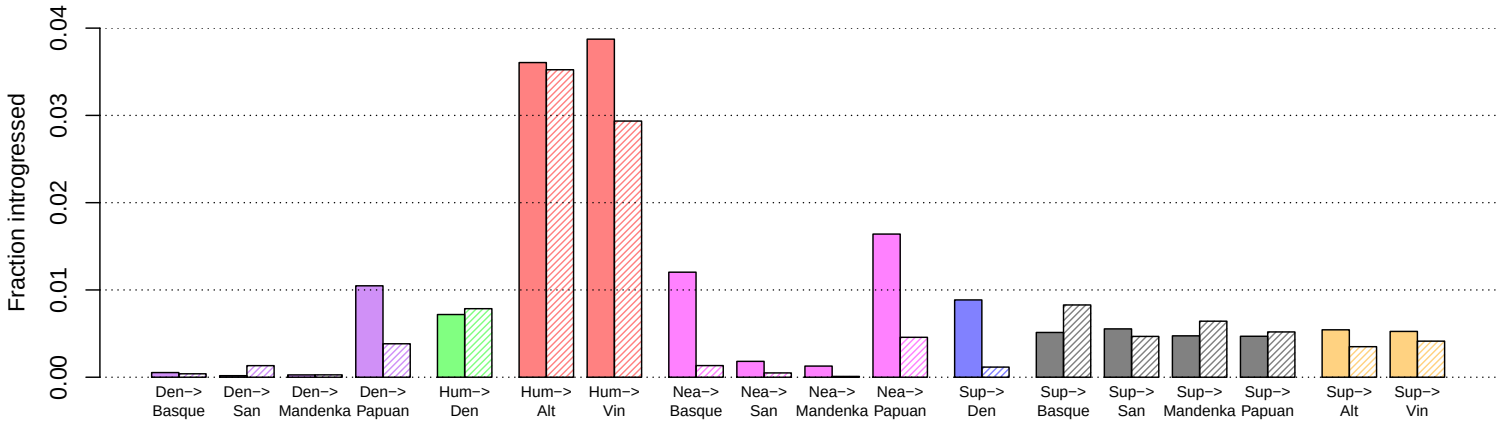

Figure 3.31: **Coverages using out-of-Africa model and full population tree. Notation as in Figure 3.30, with the addition of Papuan and Basque individuals.**

divergence was low.

We also did a genome-wide run in which we included the Papuan and Basque, along with the two Africans and archaic individuals. We analyzed the data as before, using t_{mig} = 250kya and t_{div} = 1Mya, but also added migration bands from Nea→Hum and Den→Hum at 50kya, for a total of 7 migration bands. The introgression coverages for this run are presented in Supplementary Figure 3.31. Although there are moderate differences in the absolute level of introgression predicted, the results agree well with those presented in the main paper. We do see an increase in the amount of Hum→Den and Hum→Nea regions, which most likely can be explained by the algorithm calling the incorrect direction of migration for some instances of Den→Hum and Nea→Hum.

Fraction of Neanderthal genome introgressed from humans

Using the true and false positive rates in from the deep introgression simulations in the main paper, we can make a rough estimate of the total amount of Neanderthal genome introgressed from ancient humans. If the fraction of the genome predicted to be introgressed is x , then the total amount is predicted to be $(x - FP)/TP$. These numbers are of course very rough because there are many unknown demographic parameters in the model. Using the results in the main paper, this would predict that 7.4% of the Altai autosomal genome is introgressed from Neanderthal, and 7.2% of the Vindija. If we instead use the simulation and real results from the previous supplemental section, where the demographic model has a migration at 150kya, we get predictions of 10.8% for Altai and 10.3% for Vindija.

3.8.7 Calculating the mutation rate map

We first extracted the hg19 (human), panTro4 (chimp), gorGor3 (gorilla), ponAbe2 (orangutan), and nomLeu3 (gibbon) sequences from the UCSC Genome Browser's 100-way vertebrate alignment. We then masked any segments of the alignment within 100bp of a phastCons element, using the union of all hg19 phastCons elements (phastConsElements46way, phastConsElements46wayPlacental, phastConsElements46wayPrimates, phastConsElements100way). We then ran phyloP on the alignment, using the options `--method LRT --features windows.bed --mode CONACC`, where the windows.bed file gives 100kb sliding windows of the human genome, staggered by 10kb. The substitution model used for the phyloP run was down-

loaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons100way/hg19.100way.phastCons.mod. The phyloP run produced tree scaling factors for every 100kb window. These were converted to mutation rates by rescaling all factors to achieve a mean mutation rate of 1.45e-8/bp/generation. The mutation rate for a particular base was then mean of the mutation rates in the sliding windows which overlap that base. For substitution rates on chromosome X, we used the same procedure, with the exception that we used a substitution model specific to chromosome X, downloaded from http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/primates.chrX.mod.

CHAPTER 4

CLOSING REMARKS AND FUTURE DIRECTIONS

There are two threads running throughout this dissertation: ancient hominins and ARGweaver, and I will comment on each of these in turn. The ancient DNA revolution is just beginning, and it will be very exciting to see what new data sets and discoveries are produced as new fossils are sequenced. It feels like evidence keeps pushing the image of Neanderthals closer to that of humans. Whereas Neanderthals used to be thought of as an entirely different, sub-human species, we now know that all Eurasian humans have some Neanderthal ancestry, and that Neanderthals have a sizeable chunk of human ancestry as well. The evidence that this ancestry is deleterious also seems to be shrinking as the science matures; it has recently been shown [47] that the apparent decline in Neanderthal ancestry observed across time in ancient human fossils was an artifact caused by unaccounted-for migration between European and African human populations, and is not caused by negative selection. Similarly, the observed depletion of Neanderthal ancestry near genes seems more likely due to bias in the methods than a real effect, and has not been replicated in more rigorous recent studies. The main piece of evidence left suggesting some level of deleteriousness of Neanderthal ancestry is the depletion on the X chromosome, and in a few "deserts of introgression".

I had hoped that investigating the many other introgression events between archaic hominins might shed light on possible hybrid incompatibilities between the species, or at least reveal whether other types of hominin introgression also show signs of negative selection. I did find that introgression from humans to Neanderthals is not depleted on the X, which is an interesting contrast to

the Nea→Hum case. However, overall, the biases in the ability to detect introgression is so strong and complicated that it is almost impossible to make any other conclusions about whether certain genomic regions are depleted for results. And the small sample sizes of Neanderthals and Denisovans do not allow us to detect deserts of introgression in those genomes.

The speculation about super-archaic introgression into the Denisovans, as well as into humans in Africa, is also a theory in its infancy. Until the source of introgression is discovered and sequenced, the evidence for these events remains indirect and therefore weak. It seems only a matter of time before the *Homo erectus* genome is sequenced, which is likely to answer a lot of questions about their relationship with the Denisovans and Neanderthals. The question in Africa is much more difficult, both due to the large population size and deep structure in African human populations, and the fact that the hot climate destroys ancient DNA.

Overall, this field is driven by data and is just waiting for the next fossil to reveal its secrets. However, methodological improvements and refinements, such as the ones presented in this paper, are also important. It is too easy to misinterpret the data by not accounting for all the factors that shape it, so sophisticated methods and careful analysis are required to make sense of our complex history.

As for ARGweaver, going into this dissertation, it was my hope to identify applications for this method where it outshines other approaches. ARGweaver can be difficult and slow to use, but produces a wealth of information that can be parsed in any way to look at population genetic questions from different angles. I do feel that the archaic hominin example showcases ARGweaver's strengths, and that it is most likely to be useful in data sets that are limited in sample size.

As the amount of information in a data set increases, it is more likely that faster, more approximate methods will perform very well. At the same time, convergence becomes an issue for ARGweaver as the sample size increases, so that the quality of results in large samples often decreases. It may be the case that ARGweaver is ultimately more useful for studying non-human data sets, where sample sizes are generally smaller, and low-quality sequencing is common. Its strength in distinguishing incomplete lineage sorting from migration, as well as its potential to classify selective sweeps, may be instrumental to understanding the speciation process in any number of closely related species.

APPENDIX A

SMC′ TRANSITION PROBABILITY DERIVATIONS

Here we show the derivations of the optimized transition probabilities for the SMC′ for the case where no previous recombination exists between the two adjacent sites under consideration.

When a recombination already does exist, there are three cases to consider, which are described in the Supplementary section of [1]. The ″deterministic″ and ″recombination-point″ cases do not differ between the SMC and SMC′. The ″recoalescent-point″ case can easily be adapted by adjusting the lineage counts, which is straightforward as the recombination event is already sampled.

A.1 Case 1: no previous recombination, different branches

Here we look at the case where tree T^{n-1} does not currently have a recombination between position $i-1$ and position i (i.e., $R_i^{n-1} = \emptyset$, so that $T_i^{n-1} = T_{i+1}^{n-1} \equiv T^{n-1}$). If the new branch ν coalesces to (branch x, time a) at position i and (branch y, time b) at position $i+1$, where $x \neq y$, then there must be a recombination on the new branch at time k, where with $0 \leq k \leq min(a, b)$ (refer to Figure 1.6 for an illustration).

In this case, the transition probability can be written as:

$$Pr(y, b|x \neq y, a, R_i^{n-1} = \emptyset, \Theta, T^{n-1}) = \sum_{k=0}^{min(a,b)} Pr(k|\Theta, T^{n-1}, a) Pr(b|k, \Theta, T^{n-1}) Pr(y|b, \Theta, T^{n-1}) \tag{A.1}$$

$$= Pr(y|a, b, \Theta, T^{n-1}) \sum_{k=0}^{min(a,b)} Pr(k|\Theta, T^{n-1}, a) Pr(b|k, a, \Theta, T^{n-1}) \tag{A.2}$$

The term $Pr(k|\Theta, T^{n-1}, a)$ represents the probability of recombining at a particular time, given the previous coalescence time (a), local tree (T^{n-1}), and demographic parameters (Θ). The term $Pr(b|k, \Theta, T^{n-1})$ is the probability of recoalescing at time b, and $Pr(y|b, \Theta, T^{n-1})$ is the probability that the recoalescence branch is y.

In the remainder of this section I will drop the conditioning on T^{n-1}, $R_i^{n-1} = \emptyset$, and Θ for brevity. The equation above can be further broken into the following three cases:

A.1.1 Case 1a: $a < b$

When $a < b$ then the recombination time k is in $0 \leq k \leq a$, so k must be strictly less than b. In this case, $P(y|b) = \frac{1}{l_b^C}$, where l_b^C is the number of branches available for coalescence at time b in tree T_i^{n-1}. Because $k < b$, the broken branch is not added to this count. The transition probability is:

$$Pr(y, b|a, x \neq y) = \frac{1}{l_b^C} \sum_{k=0}^{a} Pr(k|a) Pr(b|k) \tag{A.3}$$

When $k < a$, the probability $Pr(k|a)$ is the same as in the SMC, and is given by: is $\frac{(1-\exp(-\rho|T_a|))\Delta s_k(l_k+1)}{|T_a|(l_k^R+1)}$, where $|T_a|$ is the total branch length of tree T_{i-1}^n given that the new branch coalesces at time a, and Δs_k is the length of time interval k. The factor $1 - \exp(-\rho|T_a|)$ represents the probability of a recombination occurring anywhere on the tree; $\frac{\Delta s_k}{|T_a|}$ is the probability of choosing a branch segment of length Δs_k. $\frac{l_k+1}{l_k^R+1}$ is a corrective factor introduced to the discretized model to allow recombinations to occur on zero-length branches; l_k^R is the total count of branches existing in T^{n-1} at k (excluding the root branch, but including zero-length branches), whereas l_k is the number of those branches with length > 0. We add 1 to l_k and l_k^R to account for the new branch.

For notational simplicity, let $D_a \equiv \frac{1-\exp(-\rho|T_a|)}{|T_a|}$.

When $k = a$, the probability of recombination is slightly different. Because recombinations events are rounded down, the calculation considers the time interval after a, when the new branch has already coalesced. So the count in the numerator is simply l_k. But the coalescence at a creates a new node, which adds 2 possible branches for recombination to the denominator l_k^R, unless the coalescence is at the root of the tree (then we only add 1). So the denominator becomes $l_k^R + 1 + I(k < r)$, where r is the "root age" of the tree T_{i-1}.

$Pr(b|k)$ is the probability of coalescing at time b given a recombination at time k. This can be calculated by multiplying the probability of *not* coalescing between times k and $b - 1$, by the probability of coalescing at time b. Let Q_i represent the coalescence rate during the half time-interval i in tree T_{i-1}. This rate is given by:

$$Q_i = \frac{l_{\lfloor i/2 \rfloor}(s_{\frac{i+1}{2}} - s_{\frac{i}{2}})}{2N_{\lfloor \frac{i+1}{2} \rfloor}} \tag{A.4}$$

Let Q'_i represent the coalesence rate when an extra lineage is present in the tree:

$$Q'_i = \frac{(l_{\lfloor i/2 \rfloor} + 1)(s_{\frac{i+1}{2}} - s_{\frac{i}{2}})}{2N_{\lfloor \frac{i+1}{2} \rfloor}} \tag{A.5}$$

Also, define C_i and C'_i to represent cumulative coalescent rates from time zero to i, as follows:

$$C_i = \sum_{i=0}^{i} Q_i \tag{A.6}$$

$$C'_i = \sum_{i=0}^{i} Q'_i \tag{A.7}$$

(For convenience, let $C_{-1} = C'_{-1} = 0$).

Because the recombination time k is strictly less than the coalescence time b, the coalescence could have occurred in the half-time intervals immediately preceding or following time b, which are the half-intervals indexed by $2b - 1$ and $2b$. And because the previous coalesence time a is also strictly less than b, there is no extra lineage to account for, so the coalescence rate during these two intervals is $Q_{2b-1} + Q_{2b}$. Therefore the probability of coalescing during time b is $1 - \exp(-Q_{2b-1} - Q_{2b})$. However, there is an extra lineage until time a. The probability of not coalescing between time k and half-time interval $2b - 1$ is:

$$\exp\left(-\sum_{i=2k}^{2a-1} Q'_i - \sum_{i=2a}^{2b-2} Q_i\right) \tag{A.8}$$

$$= \exp\left(-\left(C'_{2a-1} - C'_{2k-1}\right) - \left(C_{2b-2} - C_{2a-1}\right)\right) \tag{A.9}$$

$$= \exp\left(-C'_{2a-1} + C'_{2k-1} - C_{2b-2} + C_{2a-1}\right) \tag{A.10}$$

Notice that when $a = k$, this term simplifies to $\exp(-C_{2b-2} + C_{2a-1})$.

We can now combine the above terms to arrive at the final transition probability for this case. It is:

$$Pr(b|a, a < b) = \frac{1}{l_b^C} D_a (1 - \exp(-Q_{2b-1} - Q_{2b})) \left[\sum_{k=0}^{a-1} \left(\frac{\Delta s_k (l_k + 1)}{l_k^R + 1} \exp\left(-C'_{2a-1} + C'_{2k-1} - C_{2b-2} + C_{2a-1}\right)\right)\right. \tag{A.11}$$

$$\left. + \frac{\Delta s_a l_a}{l_a^R + 1 + I(a < r)} \exp(-C_{2b-2} + C_{2a-1})\right] \tag{A.12}$$

Define the following substitutions for convenience. Note that these values can be pre-computed in $O(k)$ time:

$$B'_a = \sum_{k=0}^{a} \Delta s_k \frac{l_k + 1}{l_k^R + 1} \exp(C'_{2k-1}) \tag{A.13}$$

$$E_b = \frac{1 - \exp(-Q_{2b} - Q_{2b-1})}{l_b^C} \tag{A.14}$$

$$G'_a = \Delta s_a \frac{l_a}{l_a^R + 1 + I(a < r)} \tag{A.15}$$

Substituting these terms gives the transition probability:

$$Pr(b|a, a < b) = D_a E_b \left[\exp(-C'_{2a-1} - C_{2b-2} + C_{2a-1})B'_{a-1} + G'_a \exp(-C_{2b-2} + C_{2a-1})\right] \tag{A.16}$$

$$= D_a E_b \exp(-C_{2b-2} + C_{2a-1}) \left[\exp(-C'_{2a-1})B'_{a-1} + G'_a\right] \tag{A.17}$$

A.1.2 Case 1b: $a = b$

The remaining cases are performed similarly to the previous case, with rates and branch counts modified to reflect the placement of the branch at the previous tree. When $a = b$, the transition probability is as follows:

$$Pr(y, b|a = b, x \neq y) = \frac{1}{l_b^C + 2} \sum_{k=0}^{a} Pr(k|a)Pr(b|k) \tag{A.18}$$

Note that the coalescence at time $a = b$ increases the count of coalescence nodes at time b by 2. When $k < a$, the terms in the sum are:

$$Pr(k|a) = D_a \Delta s_k \frac{l_k + 1}{l_k^R + 1} \tag{A.19}$$

$$Pr(b|k) = (1 - \exp(-Q_{2b} - Q'_{2b-1})) \exp(-\sum_{i=2k}^{2b-2} Q'_i) \tag{A.20}$$

$$= (1 - \exp(-Q_{2b} - Q'_{2b-1})) \exp(C'_{2b-2} - C'_{2k-1}) \tag{A.21}$$

Note that since the extra lineage ends exactly at time b, the "prime" rate Q' is used for the half time-interval $2b - 1$, whereas the original rate Q is used for the half time interval $2b$.

When $k = a = b$, then the coalescence could only have occurred in the half time-interval $2b$, immediately after the recombination. In this case, the terms in the sum are:

$$Pr(k|a = k) = D_a \Delta s_a \frac{l_a}{l_a^R + 1 + I(a < r)} \tag{A.22}$$

$$Pr(b|k = b) = 1 - \exp(-Q_{2b}) \tag{A.23}$$

Combining these cases, the transition probability for $a = b$ becomes:

$$Pr(y, b|a = b, x \neq y) = \frac{1}{l_b^C + 2} D_a \left[\sum_{k=0}^{a-1} \Delta s_k \frac{l_k + 1}{l_k^R + 1} \left(1 - \exp(-Q_{2b} - Q'_{2b-1})\right) \exp(C'_{2b-2} - C'_{2k-1}) \right. \tag{A.24}$$

$$\left. + \Delta s_a \frac{l_a}{l_a^R + 1 + I(a < r)} \left(1 - \exp(-Q_{2b})\right) \right] \tag{A.25}$$

$$\tag{A.26}$$

Using previously defined terms, as well as the following:

$$E'_b = \frac{1 - \exp(-Q_{2b} - Q'_{2b-1})}{l_b^C + 2} \tag{A.27}$$

$$F'_a = \frac{1 - \exp(-Q_{2a})}{l_b^C + 2} \tag{A.28}$$

$$\tag{A.29}$$

the transition probability can be computed efficiently using:

$$Pr(y, b|a = b, x \neq y) = D_a \left[E'_b \exp(-C'_{2b-2}) B'_{a-1} + G'_a F'_a \right] \tag{A.30}$$

A.1.3 Case 1c: $a > b$

This case is simplest of all, because the extra lineage exists throughout. It is computed as:

$$Pr(y, b|a > b, x \neq y) = \frac{1}{l_b^C + 1} \sum_{k=0}^{b} Pr(k|a) Pr(b|k) \tag{A.31}$$

$$\tag{A.32}$$

where:

$$Pr(k|a) = D_a \Delta s_k \frac{l_k + 1}{l_k^R + 1} \tag{A.33}$$

$$\tag{A.34}$$

and

$$Pr(b|k) = \begin{cases} (1 - \exp(-Q'_{2b} - Q'_{2b-1}) \exp(-C'_{2b-2} + C'_{2k-1}) & \text{if } b < k \\ 1 - \exp(-Q'_{2b}) & \text{if } b = k \end{cases}$$

If we make the subsitutions:

$$E''_b = \frac{1 - \exp(-Q'_{2b} - Q'_{2b-1})}{l^C_b + 1} \tag{A.35}$$

$$F''_a = \frac{1 - \exp(-Q'_{2a})}{l^C_b + 1} \tag{A.36}$$

$$G''_a = \Delta s_a \frac{l_a + 1}{l^R_a + 1} \tag{A.37}$$

then the transition probability simplifies to:

$$Pr(y, b|a > b, x \neq y) = D_a \left[E''_b \exp(-C'_{2b-2}) B'_{b-1} + G''_b F''_b\right] \tag{A.38}$$

A.2 Case 2: no previous recombination, same branch ($x = y$)

A.2.1 Case 2a: $a \neq b$

When the coalescence times are different at adjacent sites, we still know there must be a recombination, and it could be on the new branch. Therefore, the probability computed in the previous section describes one possible occurrence. However, there is an additional probability that the recombination is on the branch being coalesced to (branch x). The probability associated with this possibility is computed similarly to the probabilities above, however the recombination cannot happen before time c, where c is the starting time of branch x. The transition is given by:

$$Pr(y,b|x=y,a\neq b) = Pr(y|b)\sum_{k=0}^{min(a,b)} Pr(k|a)Pr(b|k,a) + Pr(y|b)\sum_{k=c}^{min(a,b)} Pr(k|a)Pr(b|k,a) \tag{A.39}$$

$$= 2Pr(y|b)\sum_{k=0}^{min(a,b)} Pr(k|a)Pr(b|k,a) - Pr(y|b)\sum_{k=0}^{c-1} Pr(k|a)Pr(b|k,a) \tag{A.40}$$

The first term in this sum is just twice the term computed in Case 1, above. The second term is also familiar. c is the branch start time, so it must be less than or equal both a and b, so that $c-1$ is less than both a and b. Therefore this term is equivalent to the "summation" parts of the terms in the same-branch case, with the summation limit replaced by $c-1$ rather than $b-1$ or $a-1$. The final transition probabilities for the same-branch, different time case are then:

$$Pr(b|a,a\neq b,x=y) = \begin{cases} D_a E_b \left(\exp(-C'_{2a-1} - C_{2b-2} + C_{2a-1})(2B'_{a-1} - B'_{c-1}) \right. & \\ \left. +2G'_a \exp(-C_{2b-2} + C_{2a-1})\right) & \text{if } a < b \\ D_a E'_b \exp(-C'_{2b-2})(2B'_{a-1} - B'_{c-1}) + 2G'_a F'_a & \text{if } a = b^* \\ D_a E''_b \exp(-C'_{2b-2})(2B'_{b-1} - B'_{c-1}) + 2G''_b F''_b & \text{if } a > b \end{cases} \tag{A.41}$$

* Note that $a = b$ does not apply directly here, since this section deals with $a \neq b$, however this result will be used in the next section.

A.2.2 Case 1b: $a = b$

The final case with no previous recombination is when the coalescence is onto the same time and the same branch. There are several possibilities to consider here: 1) a recombination on the new branch or the coalescing branch which re-coalseces at the same point as the previous threading, and whose probability can be computed as described in the previous section 2) no recombination at all, which has a probability $\exp(-\rho|T_a|)$, or 3) any "self-recombination", such as is allowed by the SMC′ but not the original SMC. In this case, the recombination could be on any branch of the tree, so long as it coalesces back to that same branch. This is a bit different than the case of recombinations which change the tree topology. Those recombinations are stored as part of the local tree, and so are already annotated in T^{n-1}; if they are not in T^{n-1} then we know they did not occur. However, in this implementation, we will not store invisible recombinations in T^{n-1}, so that we must sum over the possibility of their existence on any branch when computing transition probabilities.

The total self recombination probability for a complete local tree T_i^n is given by:

$$Pr(\text{self recomb}|T_i^n) = \sum_{x \in \text{branches}} \left[\sum_{k=x_c}^{x_e} Pr(\text{recomb on branch x, time k}) \right. \tag{A.42}$$

$$\left. \left[\sum_{j=k}^{x_e} Pr(\text{coalescence on branch x, time j}) \right] \right] \tag{A.43}$$

$$= \sum_{x \in \text{branches}} \left[\sum_{k=x_c}^{x_e} Pr(k, x|T^n) \left[\sum_{j=k}^{x_e} Pr(y = x, j|k, x, T^n) \right] \right] \tag{A.44}$$

In the above, k represents the recombination time, x is the recombination

branch (which starts at time x_c and ends at time x_e), y is the re-coalescence branch (which is the same as x for self-recombinations), and j is the re-coalescence time. These terms are all straightforward to calculate. The challenge is calculating this efficiently; as it is written, the code must loop across all possible recombination and recoalescence points on every branch, and then do this for every possible coalescence point of a new branch v. Using some of the same strategies in the above derivations, we are able to pre-compute the inner two sums in $O(k)$ time. However, this value still needs to be calculated for every possible state, and every branch, which is $O(kn^2)$ (since there are $O(n)$ branches and $O(nk)$ states). As the number of samples increases, this calculation indeed becomes a bottleneck of the algorithm and can lead to a significant slowdown of the SMC′ compared to the SMC (see Section 1.5.3 and Figure 1.11).

A.3 Optimized SMC′ Transition probabilities

SMC′ transition probabilities

For reference, I present the formulas used in the code for the transition probabilities. This also provides a translation between variables in the above formulas, and the objects in the ARGweaver code where they are stored.

First, I summarize all the symbols/variables used in this section:

Symbol	**meaning**
N_i	population size in time interval i
n	number of samples
K	The number of time points

ν	The new branch currently being threaded
x	The coalescent branch of ν at site i
a	The coalescent time of ν at site i
y	The coalescent branch of ν at site $i+1$
b	The coalescent time of ν at site $i+1$
k	A recombination time
Θ	demographic model parameters
R_i^{n-1}	recombination events between T_i^{n-1} and T_{i+1}^{n-1}
ρ	recombination rate
T_i^n	tree at site i with n samples
T_i^{n-1}	tree at site i with one sample removed
$\lvert T_i^{n-1}\rvert$	the total branch length of tree T_i^{n-1}
$\lvert T_a\rvert$	the total branch length of a tree with ν coalescing at time a
s_i	the s^{th} discrete time point (looking backwards in time: $s_0 = 0$ = present day)
Δs_i	$s_{i+1} - s_i$
l_b	The number of linages passing through time b in T^{n-1}
l_b^C	The number of lineages available for coalescence in T^{n-1} at time b
l_b^R	The number of lineages available for recombination in T^{n-1} at time b

Table A.1: Variables used for transition probability calculations

The table below summarizes the pre-comptued values required for efficient computation of transition probabilities:

Symbol	**Name in Code**	**Equation**	**Description**

Q_i	Q0_prime	$\frac{l_{\lfloor i/2 \rfloor}(s_{\frac{i+1}{2}} - s_{\frac{i}{2}})}{2N_{\lfloor \frac{i+1}{2} \rfloor}}$	Coal rate in interval i with no extra linage
Q'_i	Q1_prime	$\frac{(l_{\lfloor i/2 \rfloor} + 1)(s_{\frac{i+1}{2}} - s_{\frac{i}{2}})}{2N_{\lfloor \frac{i+1}{2} \rfloor}}$	Coal rate in interval i with an extra lineage
C_i	C0_prime	$\sum_{i=0}^{i} Q_i$	Total coal rate from $t = 0$ to $t = i$, no extra lineages
C'_i	C1_prime	$\sum_{i=0}^{i} Q'_i$	Total coal rate from $t = 0$ to $t = i$, with an extra lineage
D_a	D	$\frac{1 - \exp(-\rho \lvert T_a \rvert)}{\lvert T_a \rvert}$	Recombination probability per unit time
B_a	B0_prime	$\sum_{k=0}^{a} \Delta s_k \frac{l_k}{l_k^R} \exp(C_{2k-1})$	Sum of recomb probabilities from $k = 0$ to $k = a$, normalized by probability of not coalescing before k, no extra lineage
B'_a	B1_prime	$\sum_{k=0}^{a} \Delta s_k \frac{l_k + 1}{l_k^R + 1} \exp(C'_{2k-1})$	Sum of recomb probabilities from $k = 0$ to $k = a$, normalized by probability of not coalescing before k, when an extra lineage exists
E_b	E0_prime	$\frac{1 - \exp(-Q_{2b} - Q_{2b-1})}{l_b^C}$	Probability of coalescing during time interval b, no extra lineage
E'_b	E1_prime	$\frac{1 - \exp(-Q_{2b} - Q'_{2b-1})}{l_b^C + 2}$	Probability of coalescing during time interval b when the extra lineage coalesces at $t = b$
E''_b	E2_prime	$\frac{1 - \exp(-Q'_{2b} - Q'_{2b-1})}{l_b^C + 1}$	Probability of coalescing in time interval b when an exta lineage always exists
F_a	F0_prime	$\frac{1 - \exp(-Q_{2a})}{l_b^C}$	Probability of coalescing in the top half of a time interval, no extra lineage

F'_a	F1_prime	$\frac{1-\exp(-Q_{2a})}{l^C_b+2}$	Probability of coalescing in the top half of a time interval, extra lineage coalesces at a
F''_a	F2_prime	$\frac{1-\exp(-Q'_{2a})}{l^C_b+1}$	Probability of coalescing in the top half of a time interval, extra lineage
G_a	G0_prime	$\Delta s_a \frac{l_a}{l^R_a}$	Recombination probability in a single time interval, no extra lineage
G'_a	G1_prime	$\Delta s_a \frac{l_a}{l^R_a+1+I(a<r)}$	Recombination factor in a single time interval, extra lineage coalesces at a
G''_a	G2_prime	$\Delta s_a \frac{l_a+1}{l^R_a+1}$	Recombination factor in a single time interval, with an extra lineage

Table A.2: A description of ARGweaver variables used to compute SMC′ transition probabilities in the code

Here are the transition probabilities:

Case	**Description**	**Equation**
1a	$a<b, v_i \neq v_{i+1}$	$D_a E_b \exp(-C_{2b-2}+C_{2a-1})\left[\exp(-C'_{2a-1})B'_{a-1}+G'_a\right]$
1b	$a=b, v_i \neq v_{i+1}$	$D_a\left[E'_b \exp(-C'_{2b-2})B'_{a-1}+G'_a F'_a\right]$
1c	$a>b, v_i \neq v_{i+1}$	$D_a\left[E''_b \exp(-C'_{2b-2})B'_{b-1}+G''_b F''_b\right]$
2a	$a<b, v_i = v_{i+1}$	2 [Case 1a prob] $-D_a E_b \exp(-C_{2b-2}+C_{2a-1})\exp(-C'_{2a-1})B'_{c-1}$
2c	$a>b, v_i = v_{i+1}$	2 [Case 1c prob] $-D_a E''_b \exp(-C'_{2b-2})B'_{c-1}$
2b	$a=b, v_i = v_{i+1}$	2 [Case 1b prob] $-D_a E'_b \exp(-C'_{2b-2})B'_{c-1}+\exp(-\rho\lvert T_a\rvert)+$ [self recomb prob]

Table A.3: SMC′ transition probabilties used in ARGweaver code

APPENDIX B

LIST OF ABBREVIATIONS

Abbreviation	Meaning
ARG	ancestral recombination graph
bp	base pair
CI	confidence interval
gen	generation
HMM	hidden Markov model
kb	kilo base
ky	kilo year
kya	kilo year ago
Mb	Mega base
MCMC	Markv chain Monte Carlo
My	million years
Mya	million years ago
TMRCA	Time to the most recent common ancestor
Hum→Den	Introgression from ancient humans into Denisovan genomes
Hum→Nea	Introgression from ancient humans into Neanderthal genomes
Den→Hum	Introgression from Densivans into human genomes
Nea→Hum	Introgression from Neanderthals into human genomes
Sup→Afr	Introgression from a super-archaic hominin into human populations from Africa
Sup→Den	Introgression from a super-archaic hominin into Denisovans
Sup→Nea	Introgression from a super-archaic hominin into Neanderthals

Table B.1: **Abbreviations used in this document**

BIBLIOGRAPHY

[1] Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 2014;10(5):e1004342.

[2] Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, et al. Ancient gene flow from early modern humans into Eastern Neanderthals. Nature. 2016;530(7591):429–433.

[3] Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. bioRxiv. 2019;doi:10.1101/592675.

[4] McVean GA, Cardin NJ. Approximating the coalescent with recombination. Philos Trans R Soc Lond, B, Biol Sci. 2005;360:1387–1393.

[5] Marjoram P, Wall JD. Fast "coalescent" simulation. BMC Genet. 2006;7:16.

[6] Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165:2213–2233.

[7] Terhorst J, Kamm JA, Song YS. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature Genetics. 2016;49:303–309.

[8] Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–496.

[9] Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nat Genet. 2014;46(8):919–925.

[10] Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011;43(10):1031–1034.

[11] Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics. 2013;194(3):647–662.

[12] Loh PR, Palamara PF, Price AL. Fast and accurate long-range phasing in a UK Biobank cohort. Nature Genetics. 2016;48:811–816.

[13] O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLOS Genetics. 2014;10(4):1–21. doi:10.1371/journal.pgen.1004234.

[14] Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics. 2007;81(5):1084–1097. doi:10.1086/521987.

[15] Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337–338. doi:10.1093/bioinformatics/18.2.337.

[16] Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014;505(7481):43–49.

[17] Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338(6104):222–226. doi:10.1126/science.1224344.

[18] Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, et al. The genomic history of southeastern Europe. Nature. 2018;555:197–203.

[19] Damgaard PdB, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, et al. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018;557(7705):369–374. doi:10.1038/s41586-018-0094-2.

[20] Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, et al. Early human dispersals within the Americas. Science. 2018;362(6419). doi:10.1126/science.aav2621.

[21] Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555(7695):190–196. doi:10.1038/nature25738.

[22] Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, et al. Reconstructing the deep population history of Central and South America. Cell. 2018;175(5):1185–1197.e22. doi:10.1016/j.cell.2018.10.027.

[23] Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, et al.

Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. Curr Biol. 2015;25(10):1395–1400. doi:10.1016/j.cub.2015.04.007.

[24] Barlow A, Cahill JA, Hartmann S, Theunert C, Xenikoudakis G, Fortes GG, et al. Partial genomic survival of cave bears in living brown bears. Nature Ecology & Evolution. 2018;2(10):1563–1570. doi:10.1038/s41559-018-0654-8.

[25] Gaunitz C, Fages A, Hanghøj K, Albrechtsen A, Khan N, Schubert M, et al. Ancient genomes revisit the ancestry of domestic and Przewalski's horses. Science. 2018;360(6384):111–114. doi:10.1126/science.aao3297.

[26] Campagna L, Repenning M, Silveira LF, Fontana CS, Tubaro PL, Lovette IJ. Repeated divergent selection on pigmentation genes in a rapid finch radiation. Science Advances. 2017;3(5). doi:10.1126/sciadv.1602404.

[27] Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981;17(6):368–376. doi:10.1007/BF01734359.

[28] Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput Biol. 2016;12(5):1–22. doi:10.1371/journal.pcbi.1004842.

[29] Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. bioRxiv. 2018;doi:10.1101/487983.

[30] DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43:491 EP –.

[31] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–2158. doi:10.1093/bioinformatics/btr330.

[32] Campagna L, Gronau I, Silveira LF, Siepel A, Lovette IJ. Distinguishing noise from signal in patterns of genomic divergence in a highly polymorphic avian radiation. Molecular Ecology. 2015;24(16):4238–4251. doi:10.1111/mec.13314.

[33] Wright S. The genetical structure of populations. Annals of Eugenics. 1951;15(1):323–354. doi:10.1111/j.1469-1809.1949.tb02451.x.

[34] Wilton PR, Carmi S, Hobolth A. The SMC' is a highly accurate approximation to the ancestral recombination graph. Genetics. 2015;200(1):343–355. doi:10.1534/genetics.114.173898.

[35] Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, et al. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science. 2003;302(5652):1960–1963. doi:10.1126/science.1088821.

[36] Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. PLOS Biology. 2005;3(6). doi:10.1371/journal.pbio.0030170.

[37] Moorjani P, Gao Z, Przeworski M. Human germline mutation and the erratic evolutionary clock. PLoS Biol. 2016;14(10):e2000744.

[38] Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, et al. Sequencing and analysis of Neanderthal genomic DNA. Science. 2006;314(5802):1113–1118. doi:10.1126/science.1131412.

[39] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710–722. doi:10.1126/science.1188021.

[40] Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014;507:354–357.

[41] Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010;468:1053 EP –.

[42] Juric I, Aeschbacher S, Coop G. The strength of selection against Neanderthal introgression. PLOS Genetics. 2016;12(11):1–25. doi:10.1371/journal.pgen.1006340.

[43] Harris K, Nielsen R. The genetic cost of Neanderthal introgression. Genetics. 2016;203(2):881–891. doi:10.1534/genetics.116.186890.

[44] Enard D, Petrov DA. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. Cell. 2018;175(2):360–371.e13. doi:10.1016/j.cell.2018.08.034.

[45] Sankararaman S, Mallick S, Patterson N, Reich D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. Current Biology. 2016;26:1241–1247.

[46] Steinrcken M, Spence JP, Kamm JA, Wieczorek E, Song YS. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. Molecular Ecology. 2018;27(19):3873–3888. doi:10.1111/mec.14565.

[47] Petr M, Pääbo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. Proceedings of the National Academy of Sciences. 2019;116(5):1639–1644. doi:10.1073/pnas.1814338116.

[48] Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999;27(2):573–580. doi:10.1093/nar/27.2.573.

[49] Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent Segmental Duplications in the Human Genome. Science. 2002;297(5583):1003–1007. doi:10.1126/science.1072047.

[50] Meacham F, Boffelli D, Dhahbi J, Martin DI, Singer M, Pachter L. Identification and correction of systematic error in high-throughput sequence data. BMC Bioinformatics. 2011;12(1):451. doi:10.1186/1471-2105-12-451.

[51] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;339(7164):851–861.

[52] Rambaut A, Grassly NC. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 1997;13:235–238.

[53] Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature. 2018;561(7721):113–116. doi:10.1038/s41586-018-0455-x.

[54] Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358(6363):655–658. doi:10.1126/science.aao1887.

[55] Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences. 2011;108(37):15123–15128. doi:10.1073/pnas.1109300108.

[56] Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, et al. Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Research. 2016;26(3):291–300. doi:10.1101/gr.196634.115.

[57] Plagnol V, Wall JD. Possible ancestral structure in human populations. PLOS Genetics. 2006;2(7):1–8. doi:10.1371/journal.pgen.0020105.

[58] Vernot B, Akey JM. Resurrecting surviving Neandertal lineages from modern human genomes. Science. 2014;343(6174):1017–1021. doi:10.1126/science.1245938.

[59] Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. Cell. 2018;173(1):53–61.e9. doi:10.1016/j.cell.2018.02.031.

[60] Hudson RR. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. 1990;7(1):44.

[61] Griffiths RC, Marjoram P. Ancestral Inference from Samples of DNA Sequences with Recombination. Journal of Computational Biology. 1996;3(4):479–502. doi:10.1089/cmb.1996.3.479.

[62] Griffiths R, Marjoram P. An ancestral recombination graph. In: Donnelly P, Tavaré S, editors. Progress in Population Genetics and Human Evolution. Springer Verlag; 1997. p. 257–270.

[63] Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–206.

[64] Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2013;30(7):1003–1005. doi:10.1093/bioinformatics/btt637.

[65] Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-

domain gene is mutated in a severe speech and language disorder. Nature. 2001;413(6855):519–523. doi:10.1038/35097076.

[66] Konopka G, Bomar JM, Winden K, Coppola G, Jonsson ZO, Gao F, et al. Human-specific transcriptional regulation of CNS development genes by FOXP2. Nature. 2009;462:213 EP –.

[67] Jacobs GS, Hudjashov G, Saag L, Kusuma P, Darusallam CC, Lawson DJ, et al. Multiple deeply divergent Denisovan ancestries in Papuans. Cell. 2019;177(4):1010–1021.e32. doi:10.1016/j.cell.2019.02.035.

[68] Durvasula A, Sankararaman S. Recovering signals of ghost archaic introgression in African populations. bioRxiv. 2019;doi:10.1101/285734.

[69] Fenner JN. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology. 2005;128(2):415–423. doi:10.1002/ajpa.20188.

[70] Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in African Americans. Nature. 2011;476(7359):170–175.

[71] Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5(10):e1000695.