Probabilistic Methods for the Inference of Selection and Demography from Ancient Human Genomes

by

Fernando Racimo

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Montgomery Slatkin, Chair
Professor Rasmus Nielsen
Professor Steven E. Brenner

Spring 2016

Abstract

Probabilistic Methods for the Inference of Selection and Demography from Ancient Human Genomes

by

Fernando Racimo

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Montgomery Slatkin, Chair

Recently developed technologies for the recovery and sequencing of ancient DNA have generated an explosion of paleogenomic data in the last five years. In particular, human paleogenomics has become a thriving field for understanding evolutionary patterns of different hominin groups over time. However, there is still a dearth of statistical tools that can allow biologists to discern meaningful patterns from ancient genomes. Here, I present three methods designed for inferring past demographic processes and detecting loci under selection using ancient and modern hominin genomes. First, I describe an algorithm to co-estimate the contamination rate, sequencing error rate and demographic parameters - including drift times and admixture rates - for an ancient nuclear genome obtained from human remains, when the putative contaminating DNA comes from present-day humans. The method is implemented in a C++ program called 'Demographic Inference with Contamination and Error' (DICE). Then, I present two methods for downstream analyses of paleogenomic samples, specifically tailored for detecting different types of positive selection. The first of these consists in a series of summary statistics for detecting adaptive introgression (AI). In particular, the number and allelic frequencies of sites that are uniquely shared between archaic humans and specific present-day populations are particularly useful for detecting adaptive pressures on introgressed haplotypes. The second approach for detecting selection is a composite likelihood ratio method called '3P-CLR', and is aimed at locating regions of the genome that were subject to selection before two populations split from each other. I use this method to look for regions under positive selection in the ancestral modern human population after its split from Neanderthals. I validate all of the above methods using simulations and real data, including present-day human genomes from the 1000 Genomes Project and several high- and low-coverage ancient genomes from archaic and early modern humans. I also recover potentially interesting candidate loci that may have been important for various phenotypic adaptations during recent human evolution.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I am grateful to my two advisors - Montgomery Slatkin and Rasmus Nielsen - who have provided me with invaluable advice and mentorship over the past four years. Monty's 'open-door' policy always meant I had direct access to one of the wisest minds in population genetics. He was always willing to discuss research questions, and help me understand difficult problems in coalescent and diffusion theory. Rasmus was a dynamic source of knowledge and advice. He was always enthusiastic about my projects, and eager to provide new ways of analyzing data to answer hard questions in human evolution, or putting me in contact with others that could.

I would like to thank my family - Mamá, Papá, Mary y Martín - for all the support they've given me throughout the years. My parents were always there when I needed them, even while living thousands of miles away from Berkeley. No hubiera podido completar este doctorado sin ustedes.

I would also like to thank Mickey Eagleson for never failing to cheer me up in my lowest moments, and reminding that there's more to life than grad school. I am forever grateful to you for all the great times we've had together.

Additionally, I am grateful to various lab members and collaborators without whom most of my PhD work would not have been possible, especially Joshua Schraiber, Emilia Huerta-Sánchez, Melinda Yang, Kelley Harris, Amy Ko, Gabriel Renaud, Benjamin Peter, Tyler Linderoth, Janet Kelso, Matteo Fumagalli, Svante Pääbo, Debora Brandt, Davide Marnetto, Martin Kuhlwilm, Joao Teixeira and Martin Kircher. I am also grateful to my thesis and qualifying committee members: Steve Brenner, Tim White, Yun Song, John Huelsenbeck and Doris Bachtrog.

1. Introduction

The last six years have seen enormous advances in the retrieval and sequencing of ancient DNA from fossils [1, 2, 3, 4, 5]. Arguably the most impressive insights stemming from these methods have been in the field of human evolutionary genomics [6]. Since the sequencing of the first ancient modern human genome [7] and the first Neanderthal genome [8] in 2010, the field of paleogenomics has exploded, and now hundreds of ancient human genomes are available for analysis [9, 10, 11, 12]. These have allowed geneticists to finely disentangle the last half million years of human evolutionary history, revealing complex patterns of population divergence, dispersal, selection and admixture. Most of these patterns would have remain hidden without ancient DNA, as it provides a unique temporal dimension that cannot be accessed with present-day sequences alone.

For these reasons, it is imperative to develop methods that can address the unique challenges of working with ancient DNA, so as to harness its potential for deriving meaningful evolutionary inferences. In this thesis, I will present three sets of methodological approaches for analyzing human paleogenomic data. Below, I will briefly introduce the biological and technical motivations for these approaches.

1.1 Challenges of ancient DNA

Unlike DNA obtained from living tissues, ancient DNA has a number of features that makes it especially challenging to analyze. First, it presents characteristic patterns of chemical damage, as a result of extended exposure to environmental conditions after biological death. The most prominent of these types of damages is generated by a process known as cytosine deamination and preferentially occurs at the ends of fragments [13, 1]. Awareness of these processes are important, as they will create mutations in sequenced reads that would not have been originally present in the living individual [14].

Another important concern is contamination. In general, ancient DNA libraries tend to be largely composed of exogenous DNA, be that from environmental microbes from the location where the fossil was found, or from individuals that handled the fossil - like archaeologists and laboratory technicians. As the proportion of ancient DNA in a fossil tends to be scarce, even a small number of exogenous DNA fragments may be overwhelmingly abundant relative to the number of endogenous DNA fragments [14]. Contamination from present-day humans is especially problematic when analyzing ancient human genomes, as the two types of DNA will look very similar to each other. This can confound downstream analyses: a researcher could, for example, wrongly infer that the present-day DNA in an archaic human genome signifies admixture between modern and archaic humans.

In the first part of this thesis, I will present a method to estimate contamination in an ancient DNA sample from a hominin fossil, along with other parameters of relevance, like sequencing error and chemical damage rates. Importantly, the method also estimates demogragraphic parameters, like admixture rates and the amount of genetic drift that separates the ancient genome from the contaminant source. While several methods already exist to estimate contamination [8, 2], this is the first to jointly estimate a demographic model along with the contamination rate. It is also particularly useful as it is one of a few to use data from the entire autosomal genome, rather than individual non-recombining sequences, like the mitochondrial genome or the Y chromosome, which may be unreliable indicators of contamination genome-wide [14, 15].

1.2 A complex history of admixture and divergence

Beside these difficulties, whole genome sequences from ancient hominins present unique advantages for the study of human history. Importantly, they can reveal ancient demographic processes, like gene flow between the ancestors of extant and extinct groups of humans. For example, analysis of the first Neanderthal genome - obtained from Vindija Cave, in Croatia - showed that modern humans and Neanderthals diverged from each other between 270,000 and 440,000 years ago (this date range was later revised to be between 275,000 and 765,000 years ago [16]). However, present-day non-African humans appear to derive a small percentage of their ancestry from post-divergence Neanderthal gene flow. This likely occurred via modern human - Neanderthal admixture event(s) in the Middle East, as modern humans were expanding out of Africa [8].

This archaic ancestry was subsequently confirmed with the sequencing of a high-coverage (50X) Neanderthal genome obtained from a finger bone found in the Altai Mountains, in Siberia (the "Altai Neanderthal") [16]. Using patterns of linkage disequilibrium, the time of the most recent admixture event into the ancestors of Eurasians was inferred to be $\sim 56,000$ years ago [17], though there is also evidence for additional episodes of Neanderthal admixture into East Asian populations [18, 19]. Later on, researchers were able to identify the admixture tracts along the human genome, using Markovian probabilistic models [20, 21]. More recently, several genomes recovered from fossils of ancient modern humans that lived shortly after the admixture event(s) have shown that these individuals carried longer Neanderthal admixture tracts than present-day humans [22, 23, 24]. This is consistent with the inferred timing of this event [17], as these tracts will tend to get shorter over time, via recombination [25].

The finding of Neanderthal-into-Non-African admixture was followed by the discovery of several other admixture events between diverged groups of humans. A toe bone found in the same cave as the Altai Neanderthal revealed the existence of a previously unknown sister group to Neanderthals, called "Denisovans". Intriguingly, individuals closely related to this group likely interbred with the ancestors of present-day Melanesians and East Asians, as these groups derive $2.3 - 3.7\%$ and $0.1 - 1.6\%$ of their ancestry, respectively, from this archaic population [26, 2, 27]. There may also have been interbreeding between Denisovans and eastern Neanderthals, as well as between Denisovans and an archaic group that diverged from present-day humans and Neanderthals about a million years ago [16]. Finally, a recent analysis of the Altai Neanderthal genome showed evidence for a fifth interbreeding event, between a basal modern present-day human population and eastern Neanderthals, which left detectable modern human admixture tracts in the Altai Neanderthal genome [27].

These discoveries suggest that admixture between diverged hominin groups in the Pleistocene was much more widespread than previously thought, and that these may be best described as a large metapopulation [28]. Interestingly though, none of the traces left by admixture involve more than a few percent of the entire human genome. In other words, only small amounts of genetic material were ultimately passed on to us from these groups. Hypotheses brought forward to explain these observations include Dobzhansky-Muller incompatibilities [20] and weakly deleterious alleles that were effectively neutral in Neanderthals, but became negatively selected after entering modern human populations with larger effective population sizes [29, 30].

1.3 Introgression as a shortcut to adaptation

In spite of the fact that most of the exchanged genetic material may have been deleterious, a growing body of evidence suggests that introgression of certain archaic genetic variants into the human gene pool may have accelerated the action of positive selection. In the past few years, biologists have found numerous haplotypes that appear to have originated in archaic populations - like Neanderthals - then introduced into modern humans and rapidly risen to high frequencies in particular locations of the world (reviewed in ref. [31]). Perhaps the prime example of this is the gene *EPAS1*. This gene was discovered to be under positive selection in Tibetans. The selected haplotype present at high frequency exclusively in Tibetans allows carriers to better respond to hypoxia at high altitudes [32]. More recently, researchers found that this haplotype was introduced into the ancestors of Tibetans via introgression from Denisovans. The Denisovan genome carries a haplotype that is almost identical to the Tibetan haplotype, which is highly divergent from almost all other haplotypes around the world [33, 34].

In the second part of this thesis, I will examine the effectiveness of different summary statistics to detect adaptive introgression. I will then use the most poweful of these to survey the landscape of adaptively introgressed variants across the human genome, in different populations around the world. Finally, I will focus on the haplotype structure and biological function of candidate genes showing strong evidence for adaptive introgression in different human populations. Some of these have been previously identified [35, 20, 21, 36], while some are novel and will require further attention in the future.

1.4 Searching for unique modern human selective events

Just as we may have shared adaptive variants with other groups of hominins, it is also plausible that some unique modern human variants set us phenotypically apart from these groups. Finding the genes that made us uniquely “modern" is a difficult challenge, as the genomic signatures of selection left by these adaptive events are likely very old and weak [37]. Methods aimed at detecting these types of events have consisted in looking at regions of the genome where Neanderthals fall outside of present-day human variation, which would be a hallmark of a selective sweep in the ancestors of present-day humans [8, 16].

In the last part of this thesis, I will present an alternative method, which is meant to scan the genome for selective sweeps that occurred in the distant past, before two populations split from each other, using a third outgroup population. My method also serves to distinguish between ancestral events and events that occurred specifically in one of the two daughter populations, after the split. I will then use this method to look for selective events that occurred in the ancestral population of Yoruba and Eurasians, after modern humans split from Neanderthals, as well as in the ancestral Eurasian population, after the split from Yoruba. I will present several interesting candidate regions, including one that contains a modern-human-specific nonsynonymous mutation that is fixed derived in all sampled present-day humans, but ancestral in Neanderthal and Denisova.

1.5 Conclusion

With the rise of ancient DNA studies, the universe of questions that can be answered about our distant and recent history has just exploded. However, as more ancient human genomes continue to be sequenced, it will become imperative to design computational tools that are tailored to the unique task of analyzing them efficiently. The three methods presented in this thesis aim to bridge the gap between raw sequence data and meaningful biological insights. These insights may help us understand the evolution of our species, and the adaptive processes that allowed us to expand around the world.

2. Joint estimation of contamination, error and demography from ancient DNA

Fernando Racimo*, Gabriel Renaud*, Montgomery Slatkin

*These authors contributed equally to this work.

2.1 Introduction

When sequencing a human genome using ancient DNA (aDNA) recovered from fossils, a common practice is to assess the amount of present-day human contamination in a sequencing library [8, 26, 2, 16, 22, 23]. Several methods exist to obtain a contamination estimate. First, one can look at 'diagnostic positions' in the mitochondrial genome at which a particular archaic population may be known to differ from all present-day humans. Then, one counts how many aDNA fragments support the present-day human base at those positions. This is the most popular technique and has been routinely deployed in the sequencing of Neanderthal genomes [38, 8]. However, contamination levels of the mitochondrial genome may sometimes differ drastically from those of the nuclear genome [14, 15].

A second technique involves assessing whether the sample was male or female using the number of fragments that map to the X and the Y chromosomes. After determining the biological sex, the proportion of reads that are non-concordant with the sex of the archaic individual are used to estimate contamination from individuals of the opposite sex (e.g. Y-chr reads in an archaic female genome are indicative of male contamination) [14, 8, 39, 16]. Another method uses a maximum-likelihood approach to estimate contamination, but is only applicable to single-copy chromosomes, like the X chromosome in individuals known *a priori* to be male [40, 41]. Finally, one last technique involves using a maximum-likelihood approach to co-estimate the amount of contamination, sequencing error and heterozygosity in the entire autosomal nuclear genome [8, 2], using an optimization algorithm such as L-BFGS-B [42].

Afterwards, if the aDNA library shows low levels of present-day human contamination ($< \sim 2\%$), demographic analyses are performed on the sequences while ignoring the contamination. If the library is highly contaminated, it is usually treated as unusable and discarded. Neither of these outcomes is optimal: contaminating fragments may affect downstream analyses, while discarding the library as a whole may waste precious genomic data that could provide important demographic insights.

One way to address this problem was proposed by [43], who developed a statistical framework to separate contaminant from endogenous DNA fragments by using the patterns of chemical deamination characteristic of ancient DNA. The method produces a score which reflects the odds that a particular fragment is endogenous or not, based on these chemical patterns. This approach is effective at isolating truly endogenous fragments from contaminant fragments, but at the cost of potentially discarding some fragments that may not have chemical damage and still be endogenous. This becomes more problematic the younger the ancient DNA sample is, because younger samples will tend to have a higher proportion of non-deaminated ancient DNA, and so the method will lead users to discard a larger fraction of endogenous material.

Instead of (or in addition to) attempting to separate the two type of fragments before performing a demographic analysis, one could incorporate the uncertainty stemming from the contaminant fragments into a probabilistic inference framework. Such an approach has already been implemented in the analysis of a haploid mtDNA archaic genome [44]. However, mtDNA represents a single gene genealogy, and, so far, no equivalent method has been developed for the analysis of the nuclear genome, which contains the richest amount of population genetic information. Here, we present a method to co-estimate the contamination rate, per-base error rate and a simple demography for an autosomal nuclear genome of an ancient hominin. We assume we have a large panel representing the putative contaminant population, for example, European, Asian or African 1000 Genomes data [45]. The method uses a Bayesian framework to obtain posterior estimates of all parameters of interest, including population-size-scaled divergence times and admixture rates.

2.2 Methods

Basic framework for estimation of error and contamination

We will first describe the probabilistic structure of our inference framework. We begin by defining the following parameters:

- r_c: contamination rate in the ancient DNA sample coming from the contaminant population

- ϵ: error rate, i.e. probability of observing a derived allele when the true allele is ancestral, or vice versa.

- i: number of chromosomes that contain the derived allele at a particular site in the ancient individual ($i = 0,\ 1\ or\ 2$)

- d_j: number of derived fragments observed at site j
- $\mathbf{d}$: vector of d_j counts for all sites $j = \{1, \ ..., \ N\}$ in a genome
- a_j: number of ancestral fragments observed at site j
- $\mathbf{a}$: vector of a_j counts for all sites $j = \{1, \ ..., \ N\}$ in a genome
- w_j: known frequency of a derived allele in a candidate contaminant panel at site j $(0 \leq w_j \leq 1)$
- $\mathbf{w}$: vector of w_j frequencies for all sites $j = \{1, \ ..., \ N\}$ in a genome
- K: number of informative SNPs used as input
- θ: population-scaled mutation rate. $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the per-generation mutation rate.

We are interested in computing the probability of the data given the contamination rate, the error rate, the derived allele frequencies from the putative contaminant population ($\mathbf{w}$) and a set of demographic parameters (Ω). We will use only sites that are segregating in the contaminant panel and we will assume that we observe only ancestral or derived alleles at every site (i.e. we ignore triallelic sites). In some of the analyses below, we will also assume that we have additional data ($\mathbf{O}$) from present-day populations that may be related to the population to which the sample belongs. The nature of the data in $\mathbf{O}$ will be explained below, and will vary in each of the different cases we describe. The parameters contained in Ω may simply be the population-scaled times separating the contaminant population and the sample from their common ancestral population. However, Ω may include additional parameters, such as the admixture rate - if any - between the contaminant and the sample population. The number of parameters we can include in Ω will depend on the nature of the data in $\mathbf{O}$.

For all models we will describe, the probability of the data can be defined as:

$$P[\ \mathbf{a}, \ \mathbf{d} \mid r_C, \epsilon, \mathbf{w}, \Omega, \mathbf{O}] = \prod_{j=1}^{K} P[a_j, d_j | r_C, \epsilon, w_j, \Omega, \mathbf{O}] \tag{2.1}$$

where

$$P[a_j, d_j | r_C, \epsilon, w_j, \Omega, \mathbf{O}] = \sum_{i=0}^{2} P[a_j, d_j \mid i, r_C, \epsilon, w_j] P[i \ |\Omega, \mathbf{O}] \tag{2.2}$$

Here, i is the true (unknown) genotype of the ancient sample, and $P[i \ |\Omega, \mathbf{O}]$ is the probability of genotype i given the demographic parameters and the data.

We focus now on computation on the likelihood for one site j in the genome. In the following, we abuse notation and drop the subscript j. Given the true genotype of the ancient individual, the number of derived and ancestral fragments at a particular site follows a binomial distribution that depends on the genotype, the error rate and the rate of contamination [8, 2]:

$$P[a, d|i, r_C, \epsilon, w] = \binom{a+d}{d} q_i^d (1-q_i)^a \tag{2.3}$$

where

$$q_2 = r_C\left(w(1-\epsilon) + (1-w)\epsilon\right) + (1-r_C)(1-\epsilon) \tag{2.4}$$

$$q_1 = r_C\left(w(1-\epsilon) + (1-w)\epsilon\right) + (1-r_C)\left((1-\epsilon)/2 + \epsilon/2\right) \tag{2.5}$$

$$q_0 = r_C\left(w(1-\epsilon) + (1-w)\epsilon\right) + (1-r_C)\epsilon \tag{2.6}$$

In the sections below, we will turn to the more complicated part of the model, which is obtaining the probability $P[i|\mathbf{\Omega}, \mathbf{O}]$ for a genotype in the ancient sample, given particular demographic parameters and additional data available. We will do this in different ways, depending on the kind of data we have at hand.

Diffusion-based likelihood for neutral drift separating two populations

First, we will work with the case in which $\mathbf{O} = \mathbf{y}$, where $\mathbf{y}$ is a vector of frequencies y_j from an "anchor" population that may be closely related to the population of the ancient DNA sample. An example of this scenario would be the sequencing of a Neanderthal sample that is suspected to have contamination from present-day humans, from which many genomes are available.

For all analyses below, we restrict to sites where $0 < y_j < 1$. Note that it is entirely possible (but not required) that $\mathbf{y} = \mathbf{w}$, meaning that, aside from the ancient DNA sample, the only additional data we have are the frequencies of the derived allele in the putative contaminant population, which we can use as the anchor population too. However, it is also possible to use a contaminant panel that is different from the anchor population (Figure 2.1.A). We will assume we have sequenced a large number of individuals from a panel of the contaminant population (for example, The 1000 Genomes Project panel) and that the panel is large enough such that the sampling variance is approximately 0. In other words, the frequency we observe in the contaminant panel will be assumed to be equal to the population frequency in the entire contaminant population. In this case, $\mathbf{\Omega} = \{\tau_C, \tau_A\}$, where τ_A and τ_C are defined as follows:

τ_A: drift time (i.e. time in generations scaled by twice the haploid effective population size) separating the population to which the ancient individual belongs from the ancestor of both populations

τ_C: drift time separating the anchor population from the ancestor of both populations

We need to calculate the conditional probabilities $P[i|\mathbf{\Omega}, \mathbf{O}] = P[i|y, \tau_C, \tau_A]$ for all three possibilities for the genotype in the ancient individual: $i = 0$, 1 or 2. To obtain these expressions, we rely on Wright-Fisher diffusion theory (reviewed in [46]), especially focusing on the two-population site-frequency spectrum (SFS) [47]. The full derivations can be found in Appendix A, and lead to the following formulas:

$$P[\ i = 0 \mid y, \tau_C, \tau_A\ ] = 1 - y * e^{-\tau_C} - \frac{1}{2} * y * e^{-\tau_A - \tau_C} + y\left(y - \frac{1}{2}\right) e^{-\tau_A - 3\tau_C} \tag{2.7}$$

$$P[\ i = 1 \mid y, \tau_C, \tau_A\ ] = y * e^{-\tau_A - \tau_C} + y\,(1 - 2y)\, e^{-\tau_A - 3\tau_C} \tag{2.8}$$

$$P[\ i = 2 \mid y, \tau_C, \tau_A\ ] = y * e^{-\tau_C} - \frac{1}{2} * y * e^{-\tau_A - \tau_C} + y\left(y - \frac{1}{2}\right) e^{-\tau_A - 3\tau_C} \tag{2.9}$$

We generated 10,000 neutral simulations using msms [48] for different choices of τ_C and τ_A (with $\theta = 20$ in each simulation) to verify our analytic expressions were correct (Figure 2.2). The probability does not depend on θ, so the choice of this value is arbitrary.

The above probabilities allows us to finally obtain $P[i \mid y_j, \mathbf{\Omega}, \mathbf{O}]$.

Estimating drift and admixture in a three-population model

Although the above method gives accurate results for a simple demographic scenario, it does not incorporate the possibility of admixture from the ancient sample to the contaminant population. This is important, as the signal of contamination may mimic the pattern of recent admixture. We will assume that, in addition to the ancient DNA sample, we also have the following data, which constitute $\mathbf{O}$:

1) A large panel from a population suspected to be the contaminant in the ancient DNA sample. The sample frequencies from this panel will be labeled $\mathbf{w}$, as before.

2) Two panels of genomes from two "anchor" populations that may be related to the ancient DNA sample. One of these populations - called population Y - may (but need not) be the same population as the contaminant and may (but need not) have received admixture from the ancient population (Figure 2.1.B). The sample frequencies for this population will be labeled as $\mathbf{y}$. The other population - called Z - will have sample frequencies labeled $\mathbf{z}$. We will assume the drift times separating these two populations are known (parameters τ_Y and τ_Z in Figure 2.1.B). This is a reasonable assumption as these parameters can be accurately estimated without the need of using an ancient outgroup sample, as long as admixture is not extremely high.

We can then estimate the remaining drift parameters, the error and contamination rates and the admixture time (β) and rate (α) between the archaic population and modern population Y. The diffusion solution for this three-population scenario with admixture is very difficult to obtain analytically. Instead, we use a numerical approximation, implemented in the program ∂a∂i [49].

Markov Chain Monte Carlo method for inference

We incorporated the likelihood functions defined above into a Markov Chain Monte Carlo (MCMC) inference method, to obtain posterior probability distributions for the contamination rate, the sequencing error rate, the drift times and the admixture rate. Our program - which we called 'DICE' - is coded in C++ and is freely available at: `http://grenaud.github.io/dice/`. We assumed uniform prior distributions for all parameters, and the boundaries of these distributions can be modified by the user.

For the starting chain at step 0, an initial set of parameters $X_0 = \{ \; r_{C0}, \epsilon_0, \Omega_0 \; \}$ is sampled randomly from their prior distributions. At step k, a new set of values for step $k+1$ is proposed by drawing values for each of the parameters from normal distributions. The mean of each of those distributions is the value for each parameter at state X_k and the standard deviation is the difference between the upper and lower boundary of the prior, divided by a constant that can be increased or decreased to achieve a desired rate of acceptance of new states [50]. By default, this constant is equal to 1,000 for all parameters. The new state is accepted with probability:

$$P[accept] = min\left(1, \frac{P[\mathbf{a}, \mathbf{d} \mid X_{k+1}]}{P[\mathbf{a}, \mathbf{d} \mid X_k]}\right) \tag{2.10}$$

where $P[\mathbf{a}, \mathbf{d} \mid X_k]$ is the likelihood defined in Equation 2.1.

Unless otherwise stated below, we ran the MCMC chain for 100,000 steps in all analyses, with a burn-in period of 40,000 and sampling every 100 steps. The sampled values were then used to construct posterior distributions for each parameter.

Multiple error rates and ancestral state misidentification

[22] showed that, when estimating contamination, ancient DNA data can be better fit by a two-error model than a single-error model. In that study, the authors co-estimate the two genome-wide error rates along with the proportion of the data that is affected by each rate. Therefore, we also included this error model as an option that the user can choose to incorporate when running our program.

Furthermore, we developed an alternative error estimation method that allows the user to flag transition polymorphisms, which are more likely to have occurred due to cytosine deamination in ancient DNA. These sites are therefore likely to be subject to different error rates than those common in present-day sequencing data [13, 1]. Our program can then estimate two error rates separately: one for transitions and one for transversions. Finally, we incorporated an option to include an ancestral state misidentification (ASM) parameter, which should serve to correct for mispolarization of alleles [51].

BAM file functionality

The standard input for DICE is a file containing counts of particular ancestral/derived base combinations and SNP frequencies (see README file online). As an additional feature, we also developed a module for the user to directly input a BAM file and a file containing population allele frequencies for the anchor and contaminant panels, rather than the standard input. The user can either choose to convert the BAM file to native DICE format using a program provided with the software package and then run the program, or run it directly on the BAM file. In the latter case, instead of calculating genome-wide error parameters, the program will calculate error parameters specific to each sequenced fragment, based on mapping qualities, base qualities and estimated deamination rates at each site (see Appendix B).

2.3 Results: two-population method

Simulations

We first used DICE to obtain posterior distributions from simulated data, under the two-population inference framework. We simulated two populations (i.e. an archaic and a modern human population) with constant population size that split a number of generations ago. For each demographic scenario tested, we generated 20,000 independent replicates (theta=1) in *ms* [52], making sure each simulation had at least one usable SNP. In general, this yielded ∼80,000 usable SNPs in total. We then proceeded to sample derived and ancestral allele counts using the same binomial sampling model we use in our inference framework, under different sequencing coverage and contamination conditions. In all simulations, the contaminant panel was the same as the anchor population panel. We then applied our method to the combined set of ∼80,000 SNPs.

Figure 2.3 and 2.4 show parameter estimation results from various demographic and contamination scenarios for a low-coverage (3X) and a high-coverage (30X) archaic genome, respectively, with low sequencing error (0.1%), and a contaminant/anchor population panel of 100 haploid genomes. In both cases, the method accurately estimates the error rate, the contamination rate and the drift parameters. All parameters are also accurately estimated for the same scenarios even if the sequencing error rate is high (10%) (Figure 2.5).

Figures 2.6, 2.7, 2.8, 2.9 show how well the method does at estimating parameters over a wide range of contamination and drift scenarios, by displaying the absolute difference between simulated parameters and their corresponding posterior modes. So long as coverage is high (for example, 5X or 30X), the contamination and anchor drift parameters are accurately estimated even at 75% contamination. The method performs well even if the drift times on both sides of the tree are as small as ≈ 0.001 or as large as ≈ 5, but starts becoming inaccurate when contamination is extremely high. In general, the contamination rate and anchor drifts are easier to determine than the drift corresponding to the ancient population.

We find that for samples of very low coverage (0.5X, 1X, 1.5X) we require a larger number of sites to obtain accurate estimates (Figures 2.10, 2.11, 2.12). For example, for a sample of 0.5X coverage, we tried different numbers of independent replicate simulations and found that at 800,000 replicates, we obtained approximately 1.6 million valid SNPs for inference, which was enough to reach reasonable levels of accuracy (Figure 2.13). We note that this number of SNPs is approximately the same as what is available, for example, in the low-coverage (0.5X) Mezmaiskaya Neanderthal genome [16], which contains about 1.55 million valid sites with coverage ≥ 1, and which we analyze below. We also observed that the MCMC chain in some of these simulations needed a longer time to converge than when testing samples of higher coverage, especially when contamination is very high, and so in this set of simulations, we ran it for 1 million steps instead of 100,000, with a burn-in of 940,000 steps and sampling every 100 steps. Finally, we note that our failure to recover the true parameters under low coverage in a single MCMC run is partly due to the chain failing to converge. Indeed, when we run the MCMC 10 times and recover the estimates from the chain with the highest posterior probability, we are able to obtain increased accuracy relative to the single run, especially when the drift parameters are extremely low and when the contamination rate is extremely high (Figures 2.14, 2.15, 2.16).

Finally, we tested the method on simulations in a more realistic scenario, in which we generated ancient and contaminant fragments based on empirical fragment sizes and then mapped them to a simulated reference genome using BWA [53] with default parameters. We produced DNA sequences from the output of msms [48] via seq-gen v.1.3.3 [54] with the HKY substitution model [55]. This allows for multiple substitutions to occur at the same site since the split from chimpanzee (which could cause ASM). We then simulated ancient DNA fragments that had a fragment size distribution emulating empirical distributions. Contaminant fragments were also sampled from the contaminant population. We used the deamination rates from the single-stranded library from the Loschbour ancient individual [9] ($\sim 8\%$ at the 5' end and $\sim 34\%$ at the 3' end with a residual deamination rate of $\sim 1\%$ along the whole fragment) to artificially deaminate the ancient fragments. We simulated sequencing errors on both the ancient and contaminant fragments using empirical sequencing error rates from a PhiX library (Illumina Corp.) sequenced at the Max Planck Institute for Evolutionary Anthropology on an Illumina HiSeq, basecalled using freeIbis[56]. With the same empirical PhiX dataset distribution, we generated quality scores for each nucleotide. Fragments were mapped back to a random individual from the contaminant panel. Figure 2.17 shows DICE's performance on this scenario with different error models. In all cases, we find that the parameters are estimated with high accuracy. As expected, the ts/tv model infers a higher error rate at transitions, due to the additional errors introduced by deamination on the ends of the ancient fragments.

Performance under violations of model assumptions

We evaluated the consequences of different violations of model assumptions. We started by observing the effects of using a small modern human panel. Figure 2.18 shows results for cases in which the contaminant/anchor panel is made up of only 20 haploid genomes. In this case, all parameters are estimated accurately, with only a slight bias towards overestimating the drift parameters, presumably because the low sampling of individuals acts as a population bottleneck, artificially increasing the drift time parameters estimated.

Additionally, we simulated a scenario in which only a single human contaminated the sample. That is, rather than drawing contaminant fragments from a panel of individuals, we randomly picked a set of two chromosomes at each unlinked site and only drew contaminant fragments from those two chromosomes. Figure 2.19 shows that inference is robust to this scenario, unless the contamination rate is very high (25%). In that case, the drift of the archaic genome is substantially under-estimated, but the error, contamination and anchor drift parameters only show slight inaccuracies in the estimate.

We then investigated the effect of admixture in the anchor/contaminant population from the archaic population, occurring after their divergence, which we did not account for in the simple, two-population model (Figure 2.20). In this case, the error and the contamination rates are accurately estimated, but both drift times are underestimated. This is to be expected, as admixture will tend to homogenize allele frequencies and thereby reduce the apparent drift separating the two populations.

Identifying the contaminant population

We sought to see whether we would use our method to identify the contaminant population, from among a set of candidate contaminants (for example, different present-day human panels). Because our MCMC samples are samples from the posterior distribution of the parameters and not the marginal likelihood of the data over the entire parameter space, we cannot perform proper Bayesian model selection. Instead, we used the posterior mode as a heuristic statistic that may suggest which panel is most likely to have contaminated the sample. We validated this choice of statistic using simulations under a variety of demographic scenarios (Figure 2.21). We simulated 5-population trees of varying drift times. The outgroup was chosen to be the ancient population and the rest were chosen to be the present-day human populations (A, B, C and D). One of the populations (A) was the true contaminant. To add another layer of complexity, we also allowed for admixture (at 0%, 5% and 50% rate) from the ancient population to the ancestral population of A and B. We then ran our MCMC method four times on each of these demographic scenarios, using D as the anchor and different panels as the putative contaminant in each run.

Figure 2.22 shows that the highest posterior mode always corresponds to the run that uses the true contaminant (A), and that the mode decreases the farther the tested contaminant is from the true contaminant in the tree. Additionally, Figures 2.23, 2.24, 2.25 show the effect of misspecifying the contaminant panel for different admixture scenarios. The error rate and the anchor drift time are correctly estimated, even when the candidate contaminant is highly diverged from the true contaminant, while the other two parameters are more sensitive to misspecification. In general, the correct candidate contaminant produces the highest posterior probability and yields the best parameter estimates.

Empirical data

We first applied our method to published ancient DNA data from a high-coverage genome (52X) from Denisova cave in Siberia (the Altai Neanderthal) [16], and visually ensured that the chain had converged. The demographic, error and contamination estimates are shown in Table 2.1. We used the African (AFR) 1000 Genomes Phase 3 panel [45] as the anchor population. The drift times estimated for both samples are consistent with the known demographic history of Neanderthals and modern humans, and the contamination rates largely agree with previous estimates (see Discussion below).

We ran our method with different putative contaminant panels: Africans (AFR), East Asians (EAS), Native Americans (AMR), Europeans (EUR), South Asians (SAS). For the Altai sample, we observe a contamination rate of $\sim 1\%$ and an error rate of $\sim 0.1\%$, regardless of which panel we use. Furthermore, the drift on the Neanderthal side of the tree seems to be 6 times as large as the drift on the modern human side of the tree, reflecting the smaller effective population size of Neanderthals after their divergence. The EUR panel is the one with the highest posterior mode (Table 2.1).

We then tested a variety of ancient DNA nuclear genome sequences at different levels of coverage, obtained via different methods (shotgun sequencing and SNP capture) and from different hominin groups (modern humans and Neanderthals). We used AFR as the anchor panel and either AFR (Table 2.2) or EUR (Table 2.3) as the contaminant panel. For samples of high and medium average coverage, the MCMC converges to reasonable values for all parameters. For example, we estimate the ancient population drift parameter (τ_A) to be larger in Neanderthals than in various modern humans sampled across Eurasia, as the effective population size of the former was smaller and their split time to Africans was larger.

However, for samples of very low coverage, we observe a failure of some of the parameters to properly converge, as the MCMC seems to get stuck in the boundaries of parameter space. We tested different boundaries and the problem remains. This appears to be less of a problem when using AFR as the putative contaminant panel than when using EUR as the putative contaminant panel, presumably because of the larger amount of SNPs that may be informative for inference. In the former case, we only observe this problem when samples are at lower than ~ 0.5X coverage. In the latter case, we observe the problem for samples at lower than ~ 3X coverage.

For example, the low-coverage Neanderthal genome (0.5X) from Mezmaiskaya Cave in Western Russia [16] seems to converge to parameters within the prior boundaries when using AFR as the contaminant panel but the ancient population drift gets stuck in the upper limit of parameter space when any of the other panels are used as contaminants (Table 2.4). Regardless of which contaminant panel is used, there is good agreement with the modern human drift parameter obtained when using the Altai Neanderthal genome. However, we note that when using non-African populations as the contaminants, we obtain a higher ($\sim 5\%$) contamination rate in the Mezmaiskaya Neanderthal than in the Altai Neanderthal. It is currently unclear to us whether this is due to the MCMC failing to properly converge or to a real feature of the data.

We sought to determine the robustness of our results to different levels of GC content. We did this because we initially hypothesized that endogenous DNA might be preserved at lower rates when GC content is low, leading to the presence of proportionally more contaminant DNA. We partitioned the Altai Neanderthal genome into three different regions of low ($0\% - 30\%$), medium ($31\% - 69\%$) and high ($70\% - 100\%$) GC content, using the 'GC content' track downloaded from the UCSC genome browser [57]. We then used the two-population method to infer contamination, error and drift parameters, using Africans as the anchor population and Europeans as the contaminant population (Figure 2.26). We observe that contamination rates are higher in low-GC regions than in medium-GC regions (Welch one-sided t-test on the posterior samples, P $<$ 2.2e-16), which in turn have higher contamination rates than high-GC regions (P $<$ 2.2e-16). The opposite trend occurs in the error estimates, while the drift parameters are largely unaffected. However, we find that the differences we observe across GC levels are almost entirely eliminated by removing CpG sites from the input dataset (Figure 2.26), as CpG sites are known to have higher mutation rates than the rest of the genome. For this reason, we recommend filtering them out when testing for contamination on ancient DNA datasets, which is what was done in Tables 1 and 2.

Finally, we tested a present-day Yoruba genome (HGDP00936) sequenced to high coverage [16], which should not contain any contamination. Indeed, when applying our method, we find this to be the case (Figure 2.27). We infer 0% contamination, regardless of whether we use EUR or AFR as the candidate contaminant. Furthermore, the anchor drift time is very close to 0 when using AFR as the anchor population (as the sample belongs to that same population), while it is non-zero (= 0.22) when using EUR, which is consistent with the drift time separating Europeans from the ancestor of Europeans and their closest African sister populations [58].

2.4 Results: three-population method

Simulations

We applied our three-population method to estimate both drift times and admixture rates. We simulated a high-coverage (30X) archaic human genome under various demographic and contamination scenarios. Each of the two anchor population panels contained 20 haploid genomes. The admixture time was 0.08 drift units ago, which under a constant population size of 2N=20,000 would be equivalent to 1,600 generations ago. When running our inference program, we set the admixture time prior boundaries to be between 0.06 and 0.1 drift units ago.

We find that the admixture time is inaccurately estimated under this implementation - likely due to lack of information in the site-frequency spectrum - so we do not show estimates for that parameter below. For admixture rates of 0%, 5% or 20%, the error and contamination parameters are estimated accurately in all cases (Figures 2.28, 2.29 and 2.30, respectively). The method is less accurate when estimating the demographic parameters, especially the admixture rate which is sometimes under-estimated. Importantly though, the accuracy of the contamination rate estimates are not affected by incorrect estimation of the demographic parameters.

We also tested what would happen if the admixture time was simulated to be recent: 0.005 drift units ago, or 100 generations ago under a constant population size of 2N=20,000. When estimating parameters, we set the prior for the admixture time to be between 0 and 0.01 drift units ago. In this last case, we observe that the drift times and the admixture rate (20%) are more accurately estimated than when the admixture event is ancient (Figure 2.31).

As before, we also verified that the posterior mode was a good proxy to identify the true contaminant (A), when running the MCMC using different contaminant panels (A, B, C and D). In all cases, we used D as the unadmixed anchor panel and B as the admixed anchor panel. Results are shown in Figure 2.32 for all the demographic scenarios from Figure 2.21. Again, we observe that the true contaminant (A) is always the one that corresponds to the highest posterior probability, though we again caution that because we do not have the marginal probabilities, we cannot formally perform model selection to favor a particular panel. Furthermore,the admixture rate from the ancient population into the ancestors of A and B is robustly estimated unless the true contaminant (A) is highly diverged from the candidate contaminant (Figures 2.33, 2.34, 2.35, for admixture rates of 0%, 5% and 50%, respectively).

Empirical data

We also applied the three-population inference framework to the high-coverage Altai Neanderthal genome. We first estimated the two drift times specific to Europeans and Africans after the split from each other (τ_Y and τ_Z, respectively), using ∂a∂i and the L-BFGS-B likelihood optimization algorithm [42], but without using the archaic genome ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$). Then, we used our MCMC method to estimate the rest of the drift times, the archaic admixture rate and the contamination and error parameters in the Neanderthal genome. We set the admixture time prior boundaries to be between 0.06 and 0.1 drift units ago, which is a realistic time frame given knowledge about modern human - Neanderthal cohabitation in Eurasia [59]. The error rate and contamination rates we obtain are similar to those obtained under the two-population method, and we estimate an admixture rate from Neanderthals into modern humans of 1.72% for the choice of contaminant panel with the highest posterior mode - which is again EUR (Table 2.5).

We also applied the method to the low-coverage Mezmaiskaya Neanderthal genome. As before, we are able to reach convergence for all parameters (including the admixture rate) with the exception of the Neanderthal drift, which gets stuck in the upper boundary of parameter space (Table 2.6).

2.5 Discussion

We have developed a new method to jointly infer demographic parameters, along with contamination and error rates, when analyzing an ancient DNA sample. The method can be deployed using a C++ program (DICE) that is easy to use and freely downloadable. We therefore expect it to be highly applicable in the field of paleogenomics, allowing researchers to derive useful information from previously unusable (highly contaminated) samples, including archaic humans like Neanderthals, as well as ancient modern humans.

Applications to simulations show that the error and contamination parameters are estimated with high accuracy, and that demographic parameters can also be estimated accurately so long as enough information (e.g. a large panel of modern humans) is available. The drift time estimates reflect how much genetic drift has acted to differentiate the archaic and modern populations since the split from their common ancestral population, and can be converted to divergence times in generations if an accurate history of population size changes is also available (for example, via methods like PSMC, [60]). Although we cannot perform proper model testing, we found via extensive simulations that the posterior mode of an MCMC run was a robust heuristic statistic to help detect which panel was most likely to have contaminated the sample. We caution, however, that the fact that a particular panel yields a higher posterior mode than another is no guarantee that it is a better fit to the data for demographic scenarios that may be different from the ones we simulated.

We also applied our method to empirical data, specifically to two Neanderthal genomes at high and low coverage, a present-day high-coverage Yoruba genome, and several ancient genome sequences of varying degrees of coverage, some obtained via shotgun-sequencing and some via SNP capture. For the high-coverage Yoruba genome, we infer no contamination, as would be expected from a modern-day sample, and drift times indicating the Yoruba sample indeed belongs to an African population.

The contamination and sequencing error estimates we obtained for the Altai Neanderthal are roughly in accordance with previous estimates [16]. The drift times we obtain under the three-population model for the African population ($\tau_C + \tau_{Afr}$) are approximately $0.411 + 0.009 = 0.42$ drift units. The geometric mean of the history of population sizes from the PSMC results in [16] give roughly that $N_e \approx 21,818$ since the African population size history started differing from that of Neanderthals, assuming a mutation rate of $1.25 * 10^{-8}$ per bp per generation. If we assume a generation time of 29 years, and use our drift time in the equation relating divergence time in generations to drift time ($t/(2N_e) \approx \tau$), this gives an approximate human-Neanderthal population divergence time of 531,486 years. This number roughly agrees with the most recent estimates obtained via other methods [16]. Additionally, the Neanderthal-specific drift time is approximately 6.5 times as large as the modern human drift time, which is expected as Neanderthals had much smaller population sizes than modern humans [61, 16]. The admixture rate from archaic to modern humans that we estimate is 1.72%, which is consistent with the rate estimate obtained via methods that do not jointly model contamination ($1.5-2.1\%$) [16]. In the case of the Altai Neanderthal, we observe that the sample was probably contaminated by one or more individuals with European ancestry.

When testing modern human and Neanderthal ancient genomes of lower coverage than the Altai Neanderthal, we obtain reasonable parameter estimates for samples of medium to high-coverage. However, we run into problems in estimation when the samples are of low coverage. For these reasons, and from our simulation results, we recommend that our method should be used on nuclear genomes with > 3X coverage. The method may converge under certain conditions at coverages as low as 0.5X (for example, in the case of the Mezmaiskaya genome under the two-population model when using AFR as the anchor and contaminant panel), but, in such cases, we caution the user to check convergence is achieved before drawing any conclusions from the estimates. For SNP capture data, we obtain reliable estimates for samples with a minimum coverage of 500,000 sites that are polymorphic in the anchor panel.

The demographic models used in our approach are simple, involving no more than three populations and a single admixture event. This is partly due to limitations of known theory about the diffusion-based likelihood of an arbitrarily complex demography for the 2-D site-frequency spectrum - in the case of the two-population method - and to the inability of ∂a∂i [49] to handle more than 3 populations at a time. In recent years, several studies have made advances in the development of methods to compute the likelihood of an SFS for larger numbers of populations using coalescent theory [62, 63, 64], with multiple population size changes and admixture events. We hope that some of these techniques could be incorporated in future versions of our inference framework.

2.6 Acknowledgments

We thank Kelley Harris, Philip Johnson, Graham Coop, Nicolas Duforet-Frebourg, Joshua Schraiber, Sergi Castellano, Christoph Theunert, Janet Kelso, Rasmus Nielsen and members of the Slatkin and Nielsen labs for helpful advice and discussions. This work was supported by NIH grant R01-GM40282 to Montgomery Slatkin.

2.7 Tables

Table 2.1. Posterior modes of parameter estimates under the two-population inference framework for the Altai Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. Africans were the anchor population in all cases, so the modern human drift is with respect to Africans. Values in parentheses are 95% posterior quantiles.

Contaminant panel	Anchor panel	Error rate	Contamination rate	Modern human drift	Neanderthal drift	Log-posterior mode
EUR	AFR	0.12% (0.119% – 0.12%)	0.952% (0.949% – 0.956%)	0.414 (0.411 – 0.414)	2.497 (2.49–2.504)	-6476175.868
AMR	AFR	0.118% (0.118% – 0.118%)	0.964% (0.963% – 0.967%)	0.414 (0.411 – 0.414)	2.499 (2.494 – 2.506)	-6484270.973
SAS	AFR	0.12% (0.12% – 0.121%)	0.95% (0.946% – 0.951%)	0.411 (0.411 – 0.414)	2.496 (2.493 – 2.5)	-6489357.978
EAS	AFR	0.13% (0.129% – 0.13%)	0.888% (0.888% – 0.891%)	0.414 (0.412 – 0.414)	2.493 (2.488 – 2.493)	-6521082.384
AFR	AFR	0.112% (0.111% – 0.112%)	0.969% (0.966% – 0.973%)	0.412 (0.41–0.413)	2.495 (2.495 – 2.504)	-6574080.092

Table 2.2. We applied the two-population method to ancient Neanderthal and modern human genomes ranging from 52X to 0.054X coverage. We tested both shotgun-sequencing data and SNP capture data. We used AFR as both the anchor panel and the putative contaminant panel. Samples are sorted by decreasing mean coverage. We define Convergence (Conv.) to be true (T) if all the parameters stably converged in a region of parameter space that does not include the upper parameter boundary. Otherwise Convergence is false (F). A line separates the two Convergence classes. SNPs = number of SNPs overlapping with anchor panel. Obs. = total number of base observations analyzed. SC = SNP capture. SS = shotgun sequencing. HG = hunter-gatherer. LBK = Linear Pottery culture. MN = Middle Neolithic. LN = Late Neolithic. NEA = Neanderthal. MH = Modern Human. LogPos = Log-posterior mode. Cov. = Mean read coverage reported in corresponding study. For SNP capture, this is the mean coverage of the targeted SNPs.

ID	Study	Group	Type	Description	Cov.	SNPs	Obs.	Conv.	Error	Cont.	τ_C	τ_A	LogPos
Altai	[16]	NEA	SS	Altai Nea.	52	9500771	495741350	T	0.11%	0.97%	0.412	2.495	-6574080.092
Loschbour	[9]	MH	SS	Loschbour	22	8733958	181642481	T	0.17%	0.00%	0.025	0.634	-5905688.289
Stuttgart	[9]	MH	SS	LBK	19	8720170	157538109	T	0.14%	0.00%	0.019	0.392	-5921770.389
I0100	[10]	MH	SC	LBK	6.727	1017124	4608980	T	0.09%	0.00%	0.025	0.361	-483569.9795
I0061	[10]	MH	SC	Karelia (HG)	5.272	729066	3189601	T	0.11%	0.00%	0.026	0.438	-394527.9859
I0104	[10]	MH	SC	Corded Ware (LN)	4.184	912245	2714837	T	0.13%	0.00%	0.022	0.325	-423668.0231
I0406	[10]	MH	SC	Spain (MN)	3.947	545379	3204204	T	0.12%	0.00%	0.024	0.367	-341002.1352
I0014	[10]	MH	SC	Motala (HG)	2.709	497524	2164912	T	0.12%	0.05%	0.031	0.445	-295108.9014
Kennewick	[65]	MH	SS	Kennewick	1	5725599	9648018	T	0.90%	2.13%	0.021	0.603	-1951145.65
MA-1	[66]	MH	SS	Mal'ta	1	6969896	13578653	T	0.26%	3.64%	0.026	0.603	-2375835.916
I0111	[10]	MH	SC	Bell Beaker (LN)	0.731	300636	456149	T	0.20%	0.16%	0.033	0.386	-127455.6442
I0013	[10]	MH	SC	Motala (HG)	0.657	349019	788739	T	0.24%	2.20%	0.033	0.464	-179713.5404
Mezmaiskaya	[16]	NEA	SS	Mezmaiskaya Nea.	0.48	4896677	6811727	T	0.52%	0.01%	0.406	1.756	-889165.6704
I0439	[10]	MH	SC	Yamnaya	0.26	176088	194152	F	0.28%	15.60%	0.04	3.495	-61178.22162
I0060	[10]	MH	SC	Bell Beaker (LN)	0.105	67741	73195	F	0.24%	12.03%	0.045	3.344	-24561.12309
I0804	[10]	MH	SC	Unetice	0.054	34069	35522	F	0.32%	7.23%	0.042	1.566	-11980.98331

Table 2.3. We applied the two-population method to ancient Neanderthal and modern human genomes ranging from 52X to 0.054X coverage. We tested both shotgun-sequencing data and SNP capture data. We used AFR as the anchor panel and EUR as the putative contaminant panel. Samples are sorted by decreasing mean coverage. We define Convergence (Conv.) to be true (T) if all the parameters stably converged in a region of parameter space that does not include the upper parameter boundary. Otherwise Convergence is false (F). A line separates the two Convergence classes. SNPs = number of SNPs overlapping with anchor panel. Obs. = total number of base observations analyzed. SC = SNP capture. SS = shotgun sequencing. HG = hunter-gatherer. LBK = Linear Pottery culture. MN = Middle Neolithic. LN = Late Neolithic. NEA = Neanderthal. MH = Modern Human. LogPos = Log-posterior mode. Cov. = Mean read coverage reported in corresponding study. For SNP capture, this is the mean coverage of the targeted SNPs.

ID	Study	Group	Type	Description	Cov.	SNPs	Observations	Conv.	Error	Cont.	τ_C	τ_A	LogPos
Altai	[16]	NEA	SS	Altai Nea.	52	9500771	495741350	T	0.13%	0.92%	0.455	2.479	-3546071.79
Loschbour	[9]	MH	SS	Loschbour	22	8733958	181642481	T	0.17%	0.03%	0.025	0.631	-5905605.85
Stuttgart	[9]	MH	SS	LBK	19	8720170	157538109	T	0.14%	0.00%	0.019	0.393	-5921740.055
I0100	[10]	MH	SC	LBK	6.727	1017124	4608980	T	0.05%	1.00%	0.025	0.382	-482622.2504
I0061	[10]	MH	SC	Karelia (HG)	5.272	729066	3189601	T	0.06%	1.69%	0.027	0.472	-393315.7875
I0104	[10]	MH	SC	Corded Ware (LN)	4.184	912245	2714837	T	0.03%	14.75%	0.027	0.685	-416006.3196
I0406	[10]	MH	SC	Spain (MN)	3.947	545379	3204204	T	0.08%	0.98%	0.025	0.387	-340329.3866
I0014	[10]	MH	SC	Motala (HG)	2.709	497524	2164912	T	0.05%	3.45%	0.033	0.542	-293020.4266
Kennewick	[65]	MH	SS	Kennewick	1	5725599	9648018	F	0.53%	45.42%	0.031	4.999	-1800155.257
MA-1	[66]	MH	SS	Mal'ta	1	6969896	13578653	F	0.06%	43.76%	0.04	5	-2133722.285
I0111	[10]	MH	SC	Bell Beaker (LN)	0.731	300636	456149	F	0.00%	50.00%	0.068	4.999	-109596.2711
I0013	[10]	MH	SC	Motala (HG)	0.657	349019	788739	F	0.01%	39.06%	0.051	4.965	-161692.482
Mezmaiskaya	[16]	NEA	SS	Mezmaiskaya Nea.	0.48	4896677	6811727	F	0.30%	5.57%	0.425	4.984	-883632.4637
I0439	[10]	MH	SC	Yamnaya	0.26	176088	194152	F	0.00%	50.00%	0.086	3.731	-50398.43106
I0060	[10]	MH	SC	Bell Beaker (LN)	0.105	67741	73195	F	0.00%	49.97%	0.113	3.685	-20210.34403
I0804	[10]	MH	SC	Unetice	0.054	34069	35522	F	0.00%	50.00%	0.08	4.636	-9780.085474

Table 2.4. Posterior modes of parameter estimates under the two-population inference framework for the Mezmaiskaya Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. AFR were the anchor population in all cases, so the modern human drift is with respect to Africans. Values in parentheses are 95% posterior quantiles. Except when using AFR as the contaminant, the Neanderthal drift parameter gets stuck at the upper boundary (5 drift units) of parameter space.

Contaminant panel	Anchor panel	Error rate	Contamination rate	Modern human drift	Neanderthal drift	Log-posterior mode
EUR	AFR	0.295% (0.284% – 0.306%)	5.568% (5.472% – 5.673%)	0.425 (0.423 – 0.429)	4.984 (4.95 – 5)	-883632.4637
AMR	AFR	0.316% (0.3% – 0.322%)	5.333% (5.261% – 5.48%)	0.426 (0.422 – 0.428)	4.994 (4.952 – 4.999)	-884312.5366
SAS	AFR	0.328% (0.317% – 0.341%)	5.203% (5.097% – 5.313%)	0.426 (0.422 – 0.428)	4.996 (4.946 – 4.999)	-884684.3521
EAS	AFR	0.393% (0.379% – 0.402%)	4.53% (4.48% – 4.684%)	0.423 (0.421 – 0.426)	4.99 (4.887 – 4.999)	-885493.7081
AFR	AFR	0.515% (0.5% – 0.525%)	0.007% (0.002% – 0.126%)	0.406 (0.403 – 0.409)	1.756 (1.701 – 1774)	-889165.6704

Table 2.5. Posterior modes of parameter estimates under the three-population inference framework for the Altai Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. In all cases, Africans were the unadmixed anchor population and Europeans were the admixed anchor population. The ancestral human drift refers to the drift in the modern human branch before the split of Europeans and Africans. The post-split European-specific and African-specific drifts were estimated separately without the archaic genome ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$).

Contaminant panel	Unadmixed anchor panel	Admixed anchor panel	Error rate	Contamination rate	Ancestral human drift	Neanderthal drift	Admixture rate	Log-posterior mode
EUR	AFR	EUR	0.119% (0.119% – 0.12%)	0.967% (0.954% – 0.967%)	0.411 (0.405 – 0.414)	2.669 (2.656 – 2.689)	1.72% (1.682% – 1.805%)	-7452958.125
AMR	AFR	EUR	0.119% (0.118% – 0.12%)	0.967% (0.962% – 0.974%)	0.407 (0.402 – 0.412)	2.677 (2.651 – 2.708)	1.661% (1.618% – 1.696%)	-7461041.325
SAS	AFR	EUR	0.122% (0.122% – 0.123%)	0.95% (0.944% – 0.955%)	0.399 (0.398 – 0.406)	2.682 (2.677 – 2.695)	1.469% (1.422% – 1.48%)	-7465214.726
EAS	AFR	EUR	0.13% (0.129% – 0.132%)	0.896% (0.884% – 0.903%)	0.421 (0.413 – 0.428)	2.702 (2.658 – 2.706)	2.388% (2.009% – 2.447%)	-7509504.053
AFR	AFR	EUR	0.117% (0.117% – 0.119%)	0.957% (0.945% – 0.964%)	0.409 (0.409 – 0.418)	2.681 (2.66−2.702)	1.837% (1.766% – 1.961%)	-7554080.773

Table 2.6. Posterior modes of parameter estimates under the three-population inference framework for the Mezmaiskaya Neanderthal autosomal genome. We used different 1000G populations as candidate contaminants. In all cases, Africans were the unadmixed anchor population and Europeans were the admixed anchor population. The ancestral human drift refers to the drift in the modern human branch before the split of Europeans and Africans. The post-split European-specific and African-specific drifts were estimated separately without the archaic genome ($\tau_{Afr} = 0.009$, $\tau_{Eur} = 0.255$). In all cases, the Neanderthal drift parameter gets stuck at the upper boundary (5 drift units) of parameter space.

Contaminant panel	Unadmixed anchor panel	Admixed anchor panel	Error rate	Contamination rate	Ancestral human drift	Neanderthal drift	Admixture rate	Log-posterior mode
AFR	AFR	EUR	0.517% (0.502% – 0.526%)	4.663% (4.564% – 4.787%)	0.428 (0.426 – 0.432)	4.999 (4.989 – 5)	1.609% (1.585% – 1.63%)	-1025944.516
EAS	AFR	EUR	0.71% (0.697% – 0.721%)	2.471% (2.403% – 2.564%)	0.415 (0.412 – 0.418)	4.997 (4.985 – 5)	1.486% (1.462% – 1.508%)	-1028456.347
AMR	AFR	EUR	0.727% (0.71% – 0.733%)	2.288% (2.208% – 2.361%)	0.414 (0.412 – 0.417)	4.999 (4.985 – 5)	1.482% (1.459% – 1.501%)	-1028866.312
SAS	AFR	EUR	0.724% (0.709% – 0.732%)	2.315% (2.219% – 2.375%)	0.414 (0.412 – 0.418)	4.998 (4.984 – 5)	1.479% (1.458% – 1.5%)	-1028823.568
EUR	AFR	EUR	0.761% (0.745% – 0.77%)	1.875% (1.784% – 1.928%)	0.413 (0.41 – 0.415)	4.998 (4.984 – 2.5)	1.463% (1.457% – 1.495%)	-1029429.156

2.8 Figures

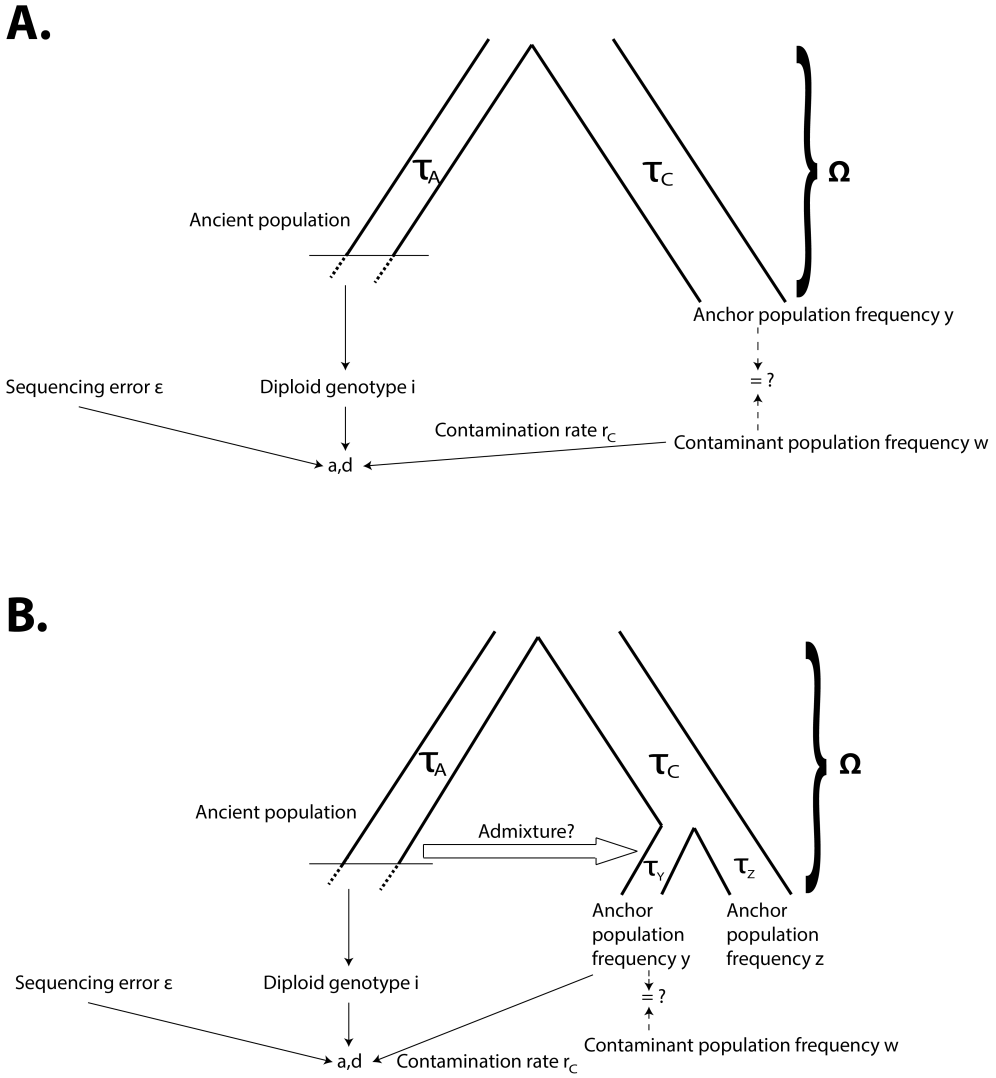


Figure 2.1. A) Schematic of two-population modeling framework: at each site, derived and ancestral fragments (a, d) are binomially sampled from the true genotype of the archaic individual, with some amount of contamination and error. In turn, the true genotype depends on a demographic model, which can include the contaminant population. B) Schematic of three-population modeling framework, incorporating admixture between the archaic population and one of two anchor populations.

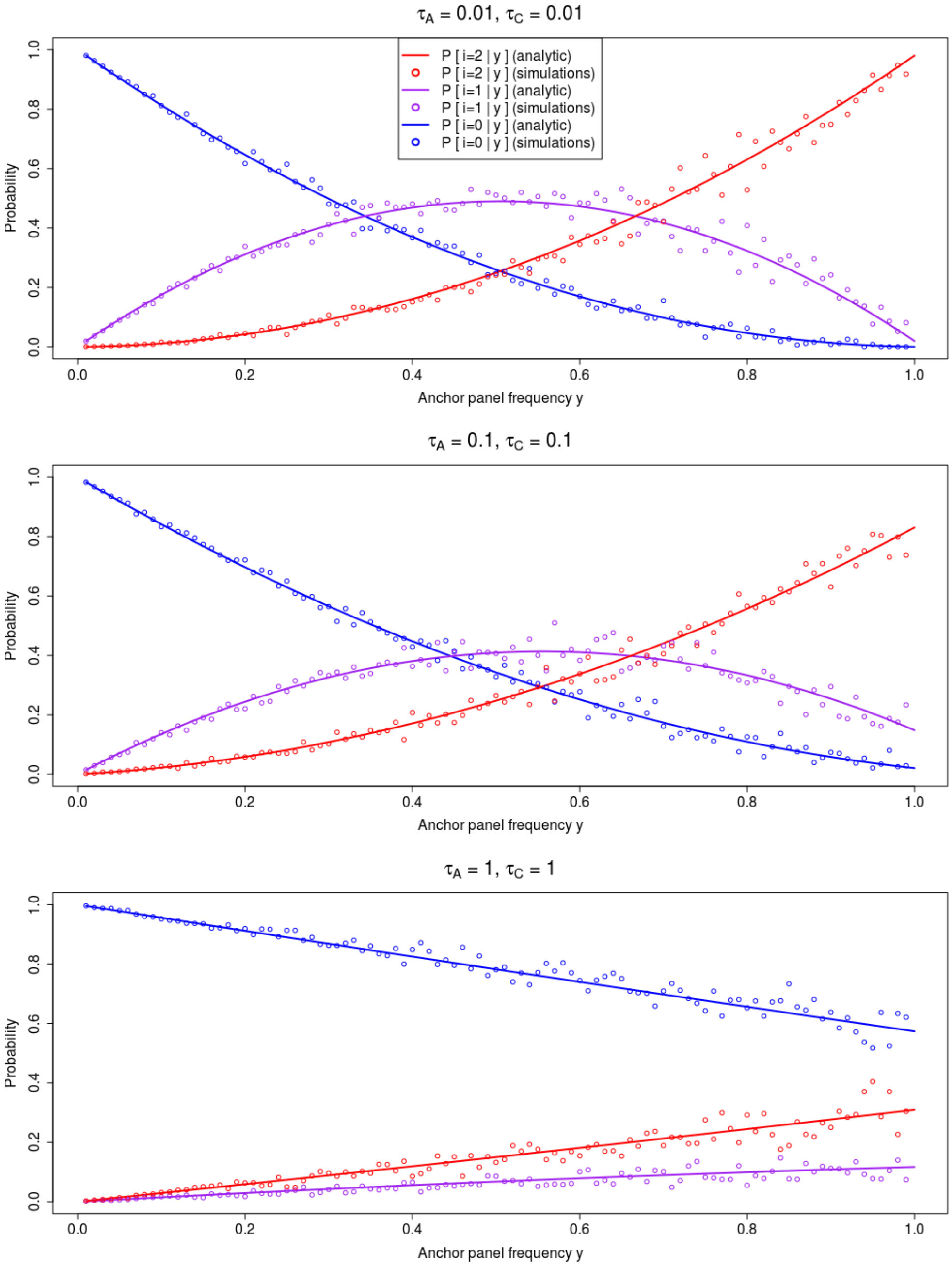


Figure 2.2. Comparison of analytic solutions to $P[i|y, \tau_C, \tau_A]$ and simulations under neutrality from msms, for different choices of τ_A and τ_C.

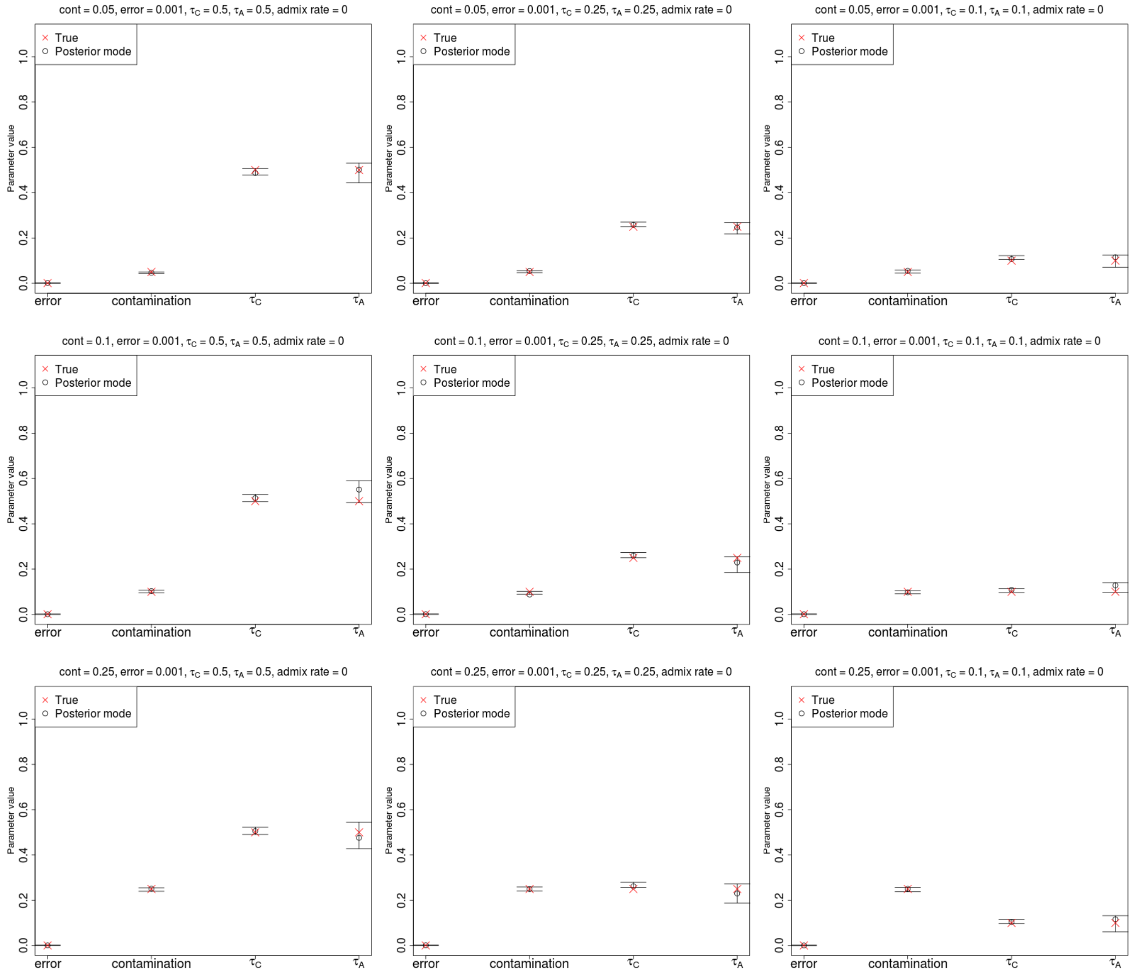


Figure 2.3. Estimation of parameters for a low-coverage ancient DNA genome (3X) with low sequencing error (0.1%), no admixture and a large anchor population panel (100 haploid genomes). Error bars represent 95% posterior intervals.

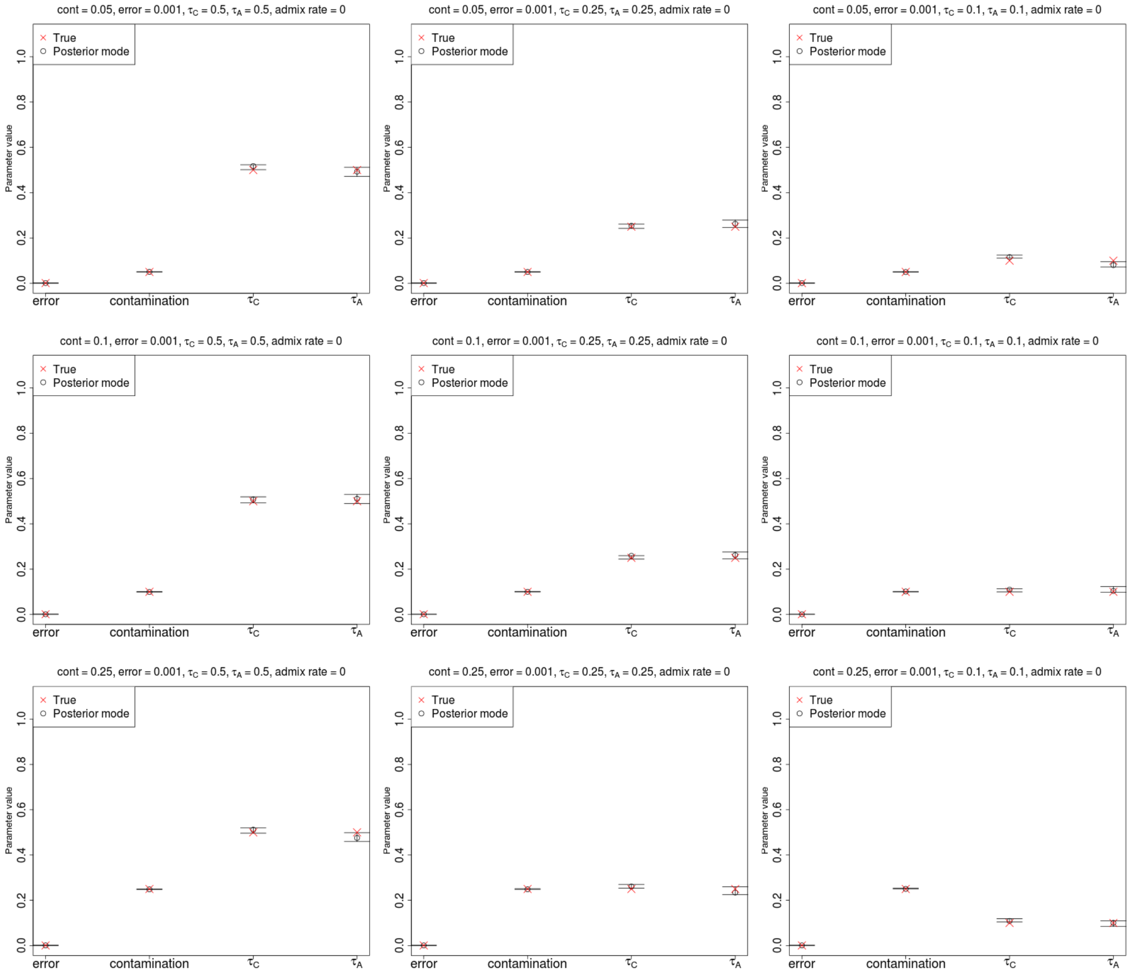


Figure 2.4. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), no admixture and a large anchor population panel (100 haploid genomes). Error bars represent 95% posterior intervals.

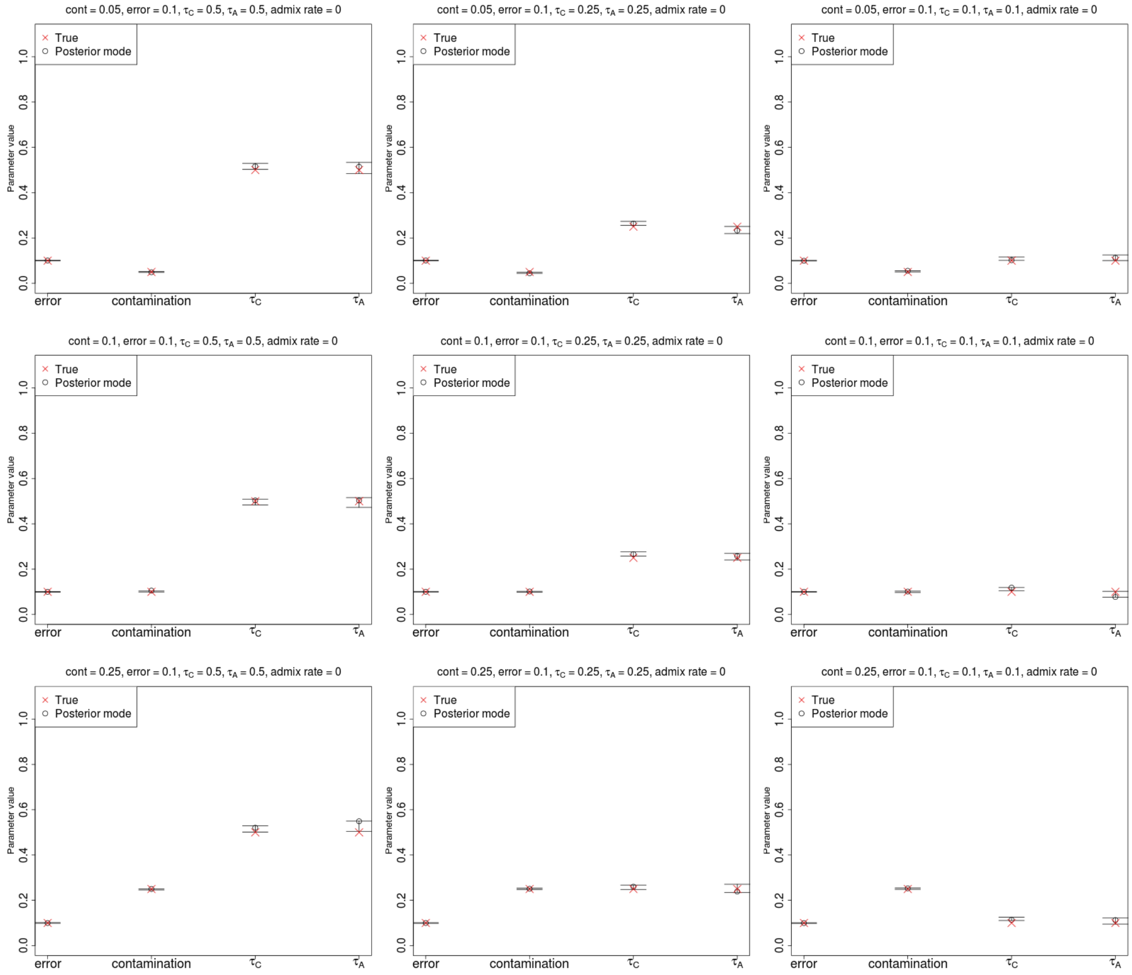


Figure 2.5. Estimation of parameters for a high-coverage ancient DNA genome (30X) with high sequencing error (10%), no admixture and a large anchor population panel (100 haploid genomes). Error bars represent 95% posterior intervals.

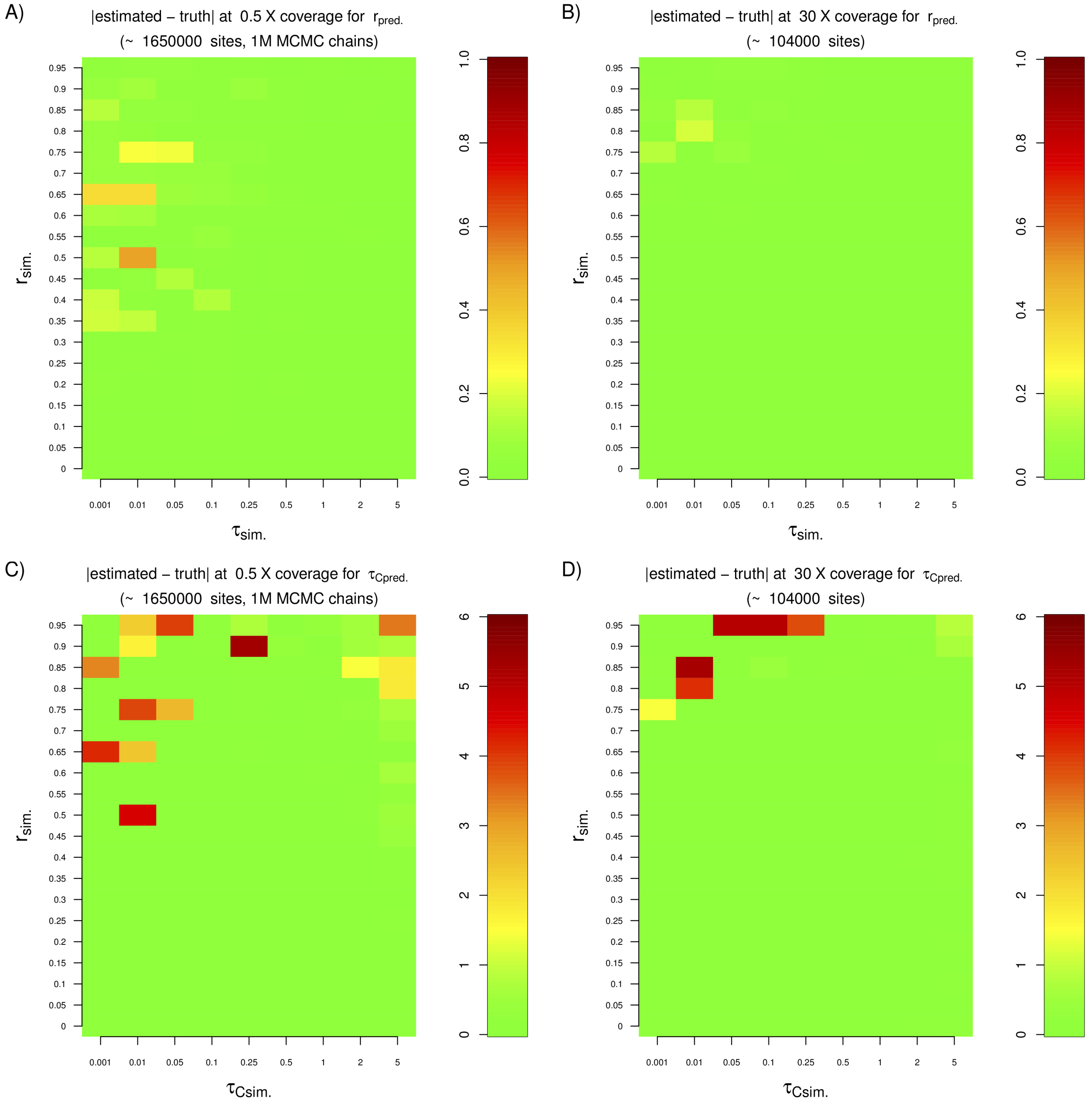


Figure 2.6. We tested the performance of the two-population method under a variety of drift and contamination scenarios for a sample of very low (0.5X) or very high (30X) coverage. We found that we needed more sites (≈ 1.6 million) to obtain accurate estimates from the low coverage sample. The MCMC chain was also run for a longer time (1 million steps). A) Absolute difference between the estimated and the simulated contamination rate for a 0.5X genome. B) Absolute difference between the estimated and estimated and the simulated contamination rate for a 30X genome. C) Absolute difference between the estimated and the simulated anchor drift for a 0.5X genome. D) Absolute difference between the estimated and the simulated anchor drift for a 30X genome. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

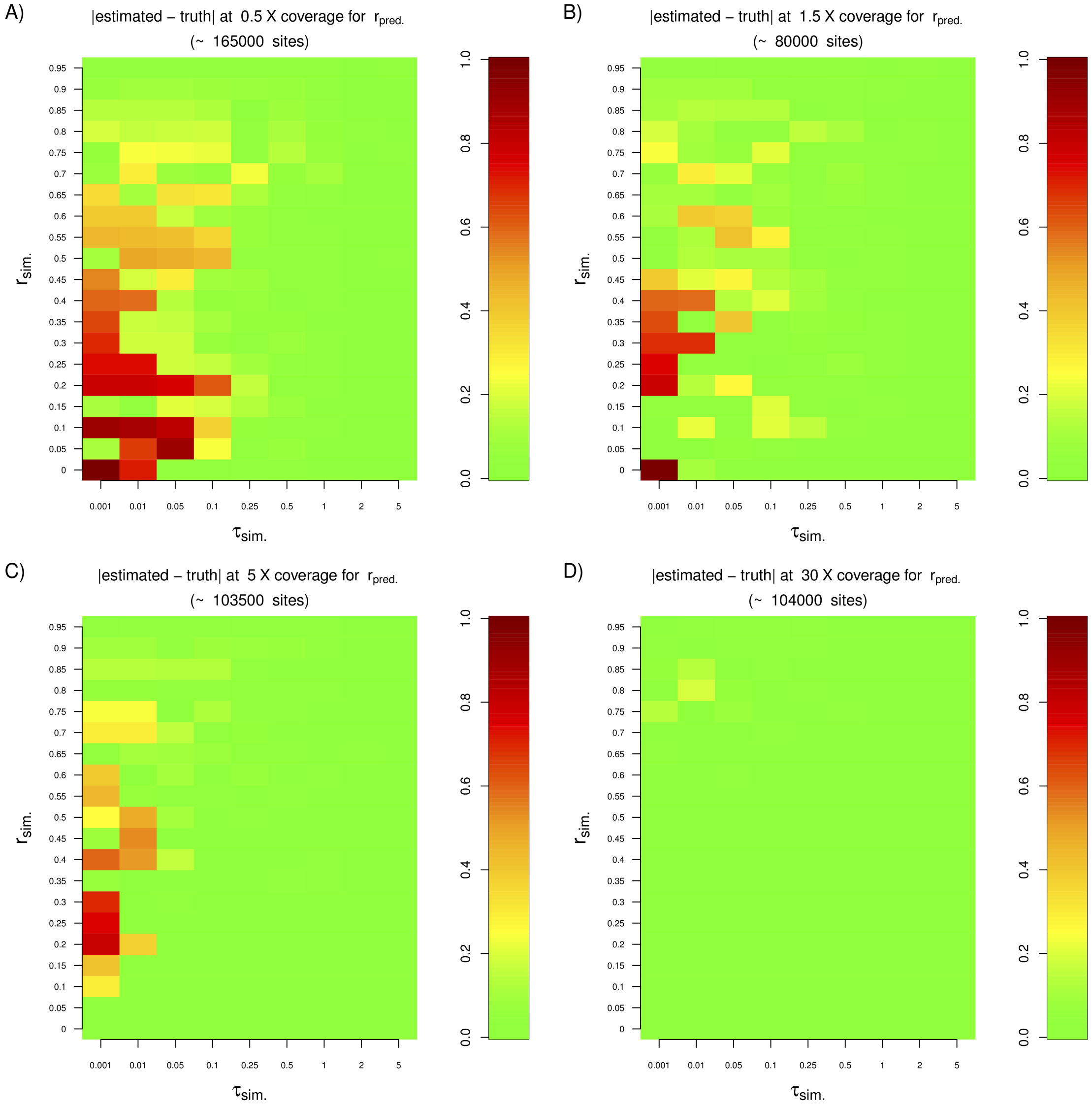


Figure 2.7. Absolute difference between estimated and simulated contamination rates for a variety of anchor drift and contamination scenarios, for different levels of coverage. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

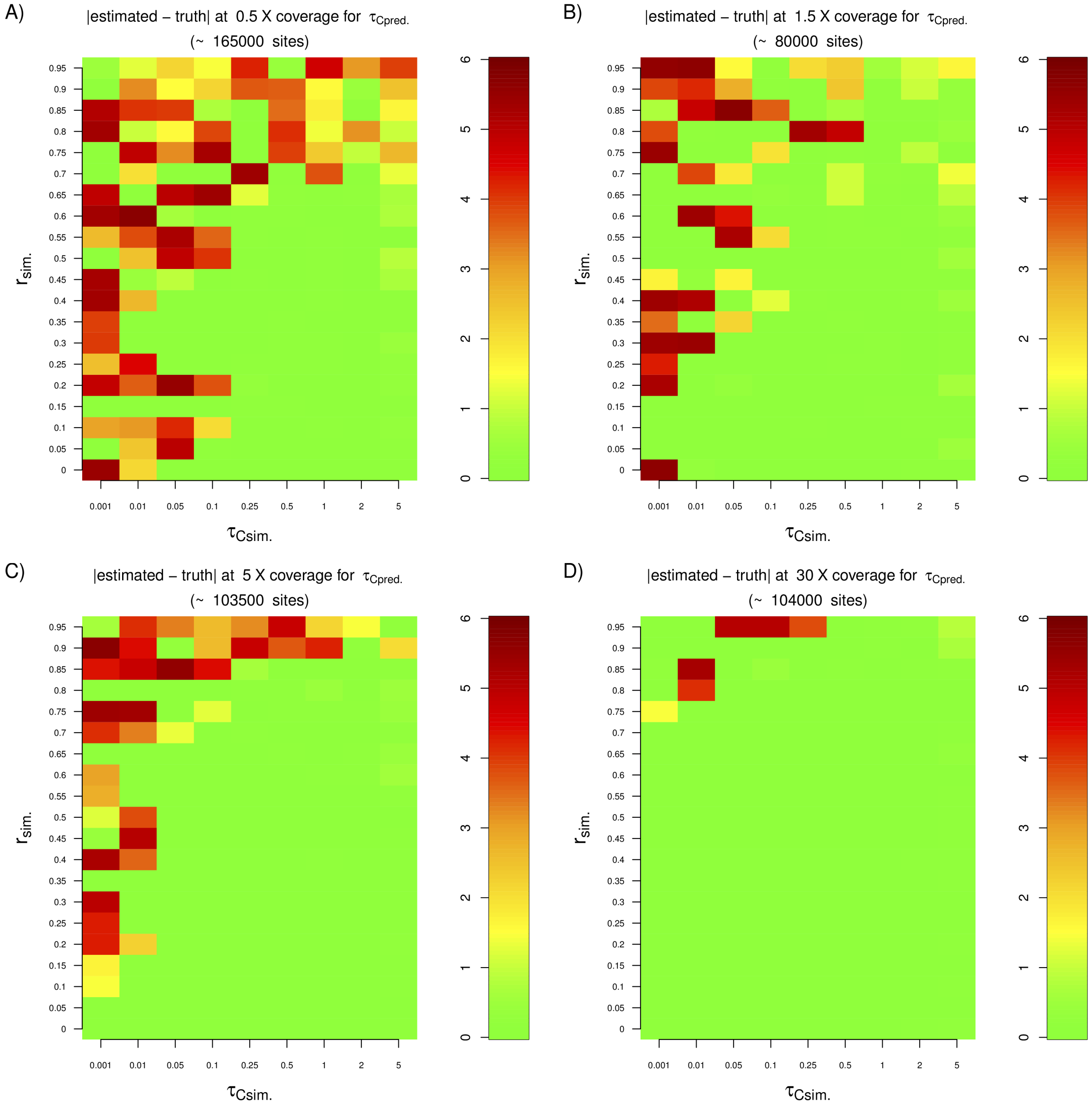


Figure 2.8. Absolute difference between estimated and simulated anchor drifts for a variety of anchor drift and contamination scenarios, for different levels of coverage. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

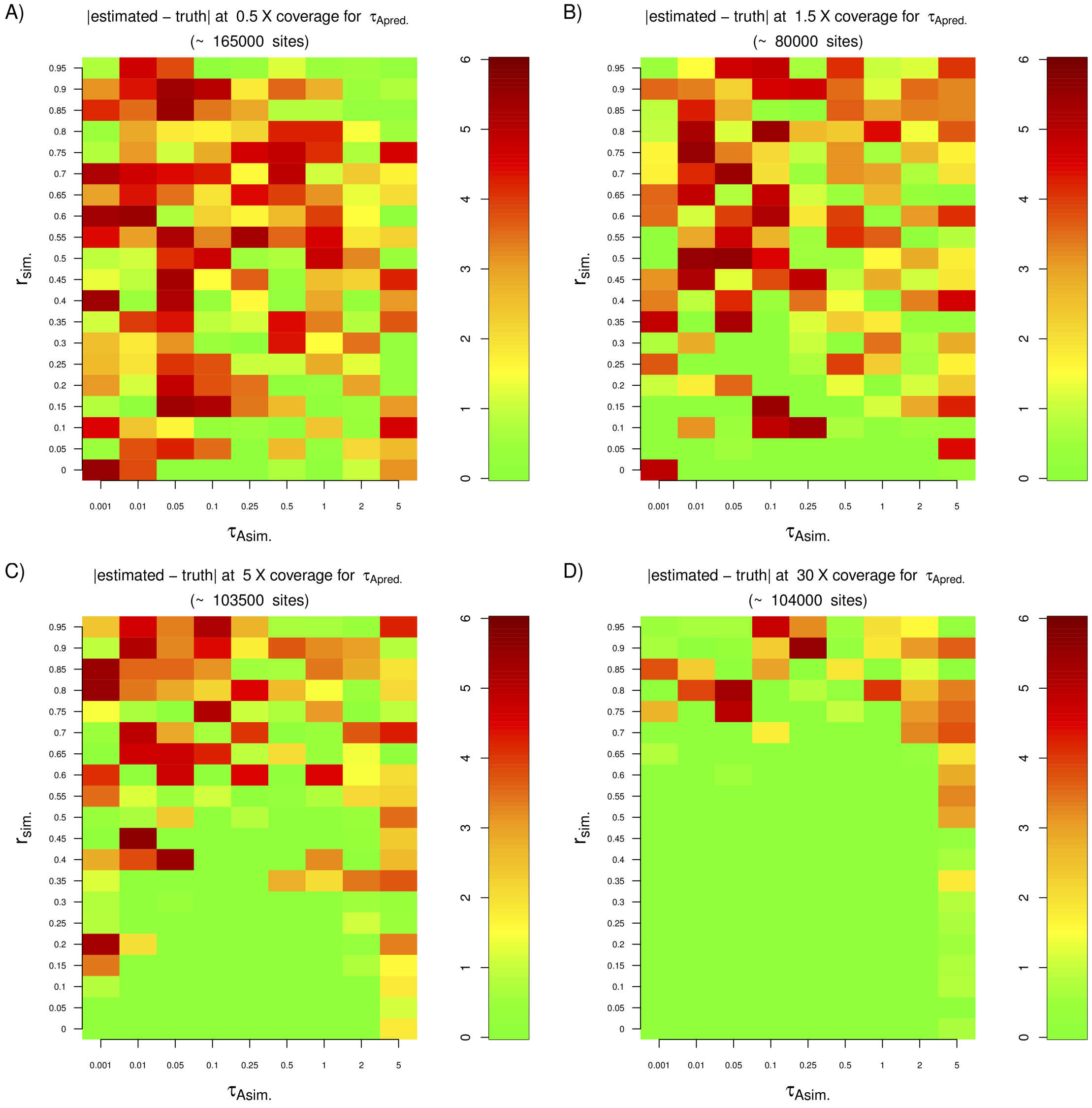


Figure 2.9. Absolute difference between estimated and simulated ancient sample drifts for a variety of anchor drift and contamination scenarios, for different levels of coverage. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

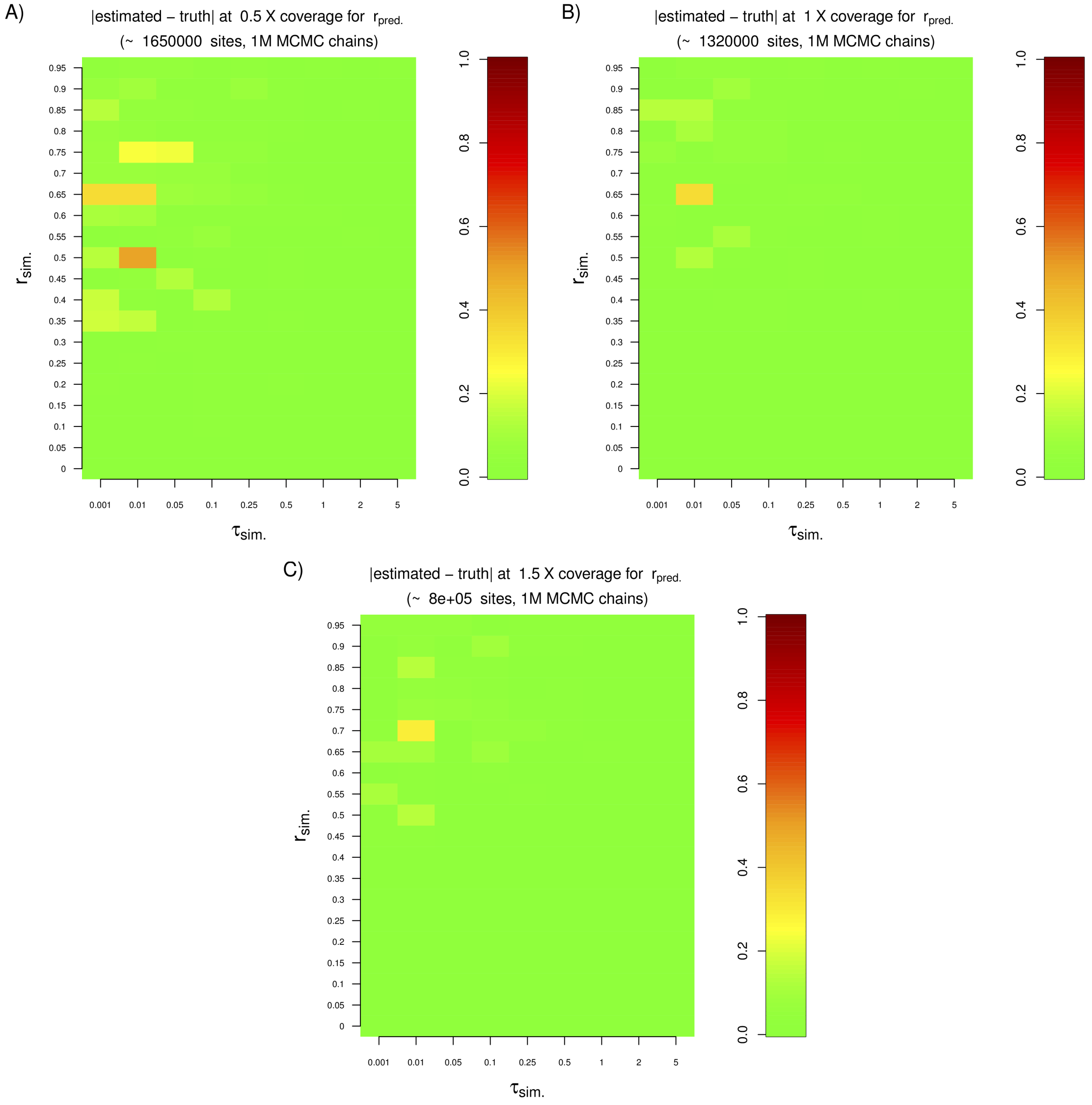


Figure 2.10. Absolute difference between estimated and simulated contamination rates for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). Here, we used a large number of sites and run the MCMC chain for 1 million steps. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

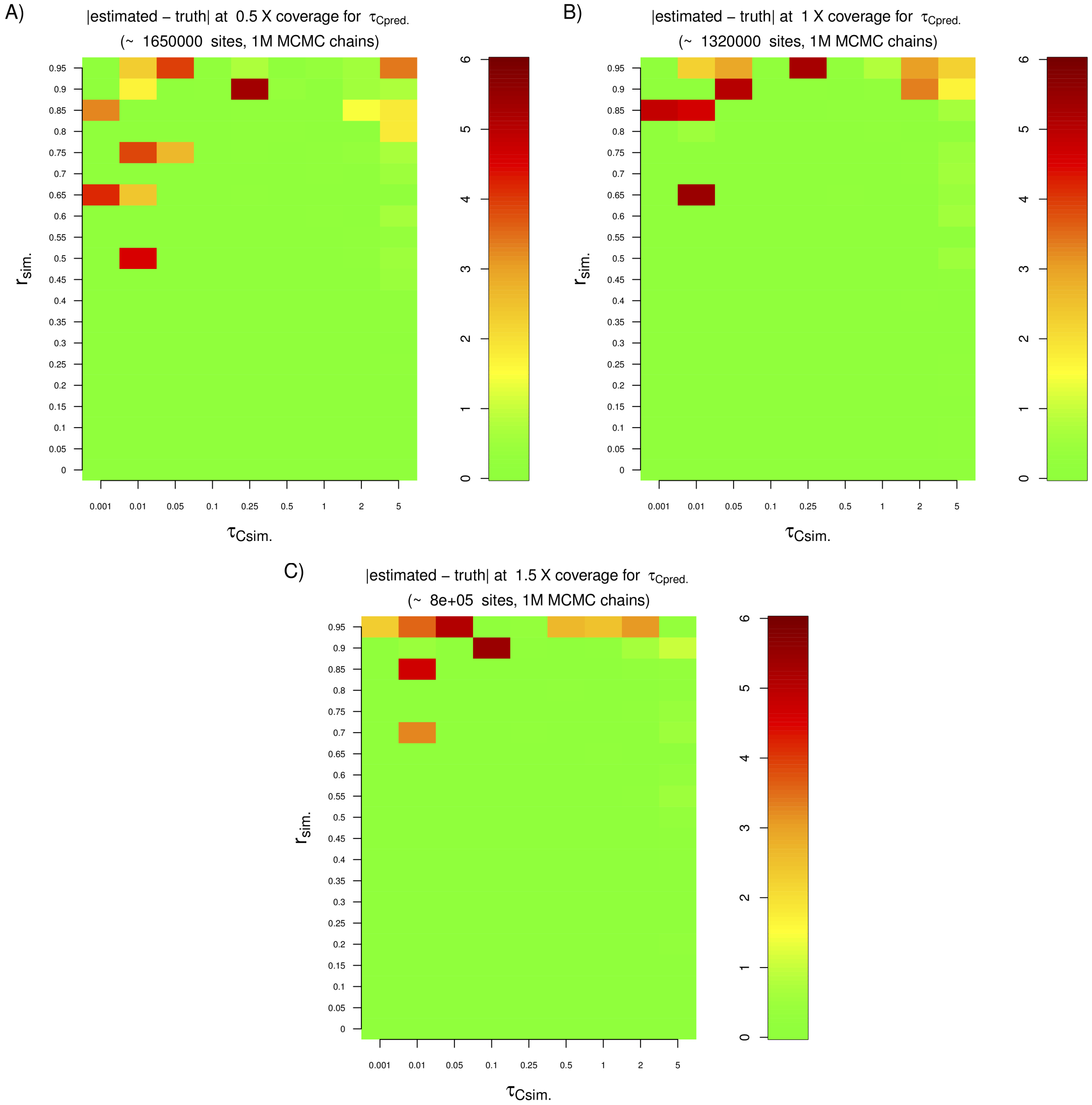


Figure 2.11. Absolute difference between estimated and simulated anchor drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). Here, we used a large number of sites and run the MCMC chain for 1 million steps. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

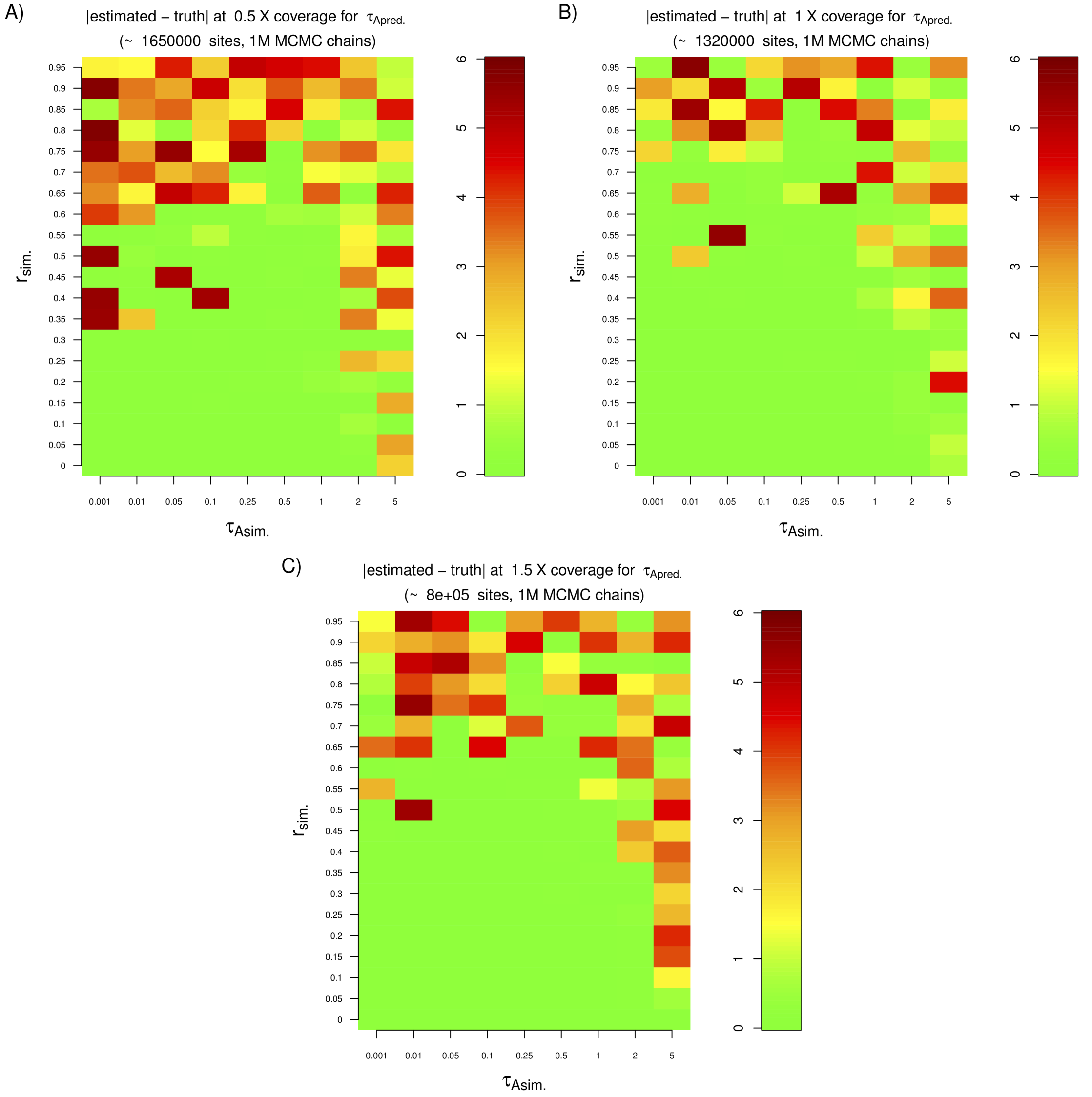


Figure 2.12. Absolute difference between estimated and simulated ancient sample drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). Here, we used a large number of sites and run the MCMC chain for 1 million steps. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

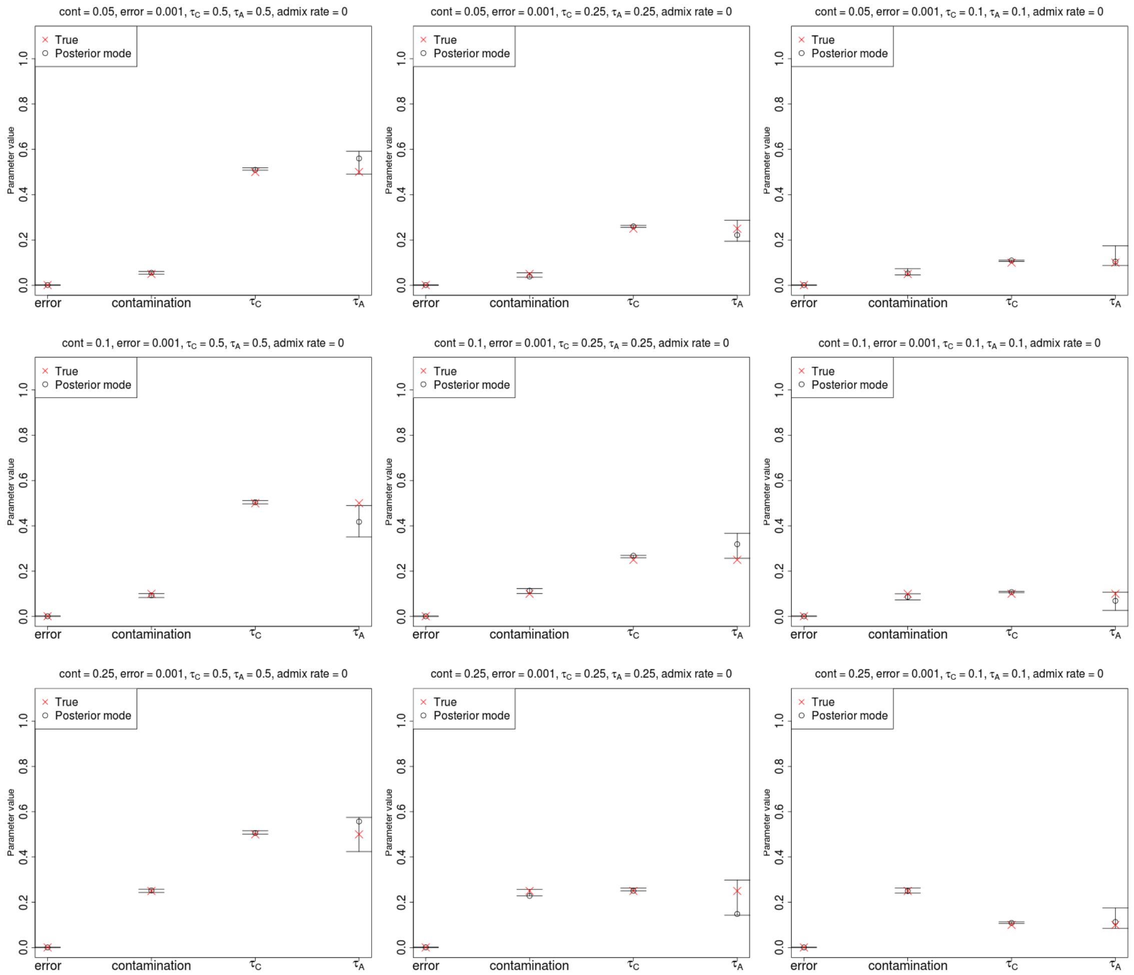


Figure 2.13. Estimation of parameters for an ancient DNA genome of very low coverage (0.5X) with low sequencing error (0.1%) and a large anchor population panel (100 haploid genomes). Note that unlike the rest of the simulations, the number of SNPs used in this case was approximately 1.6 million instead of 80,000, and the MCMC chain was run for 1 million steps instead of 100,000. Using a lower number of SNPs or running the chain for a shorter time resulted in inaccurate inferences. Error bars represent 95% posterior intervals.

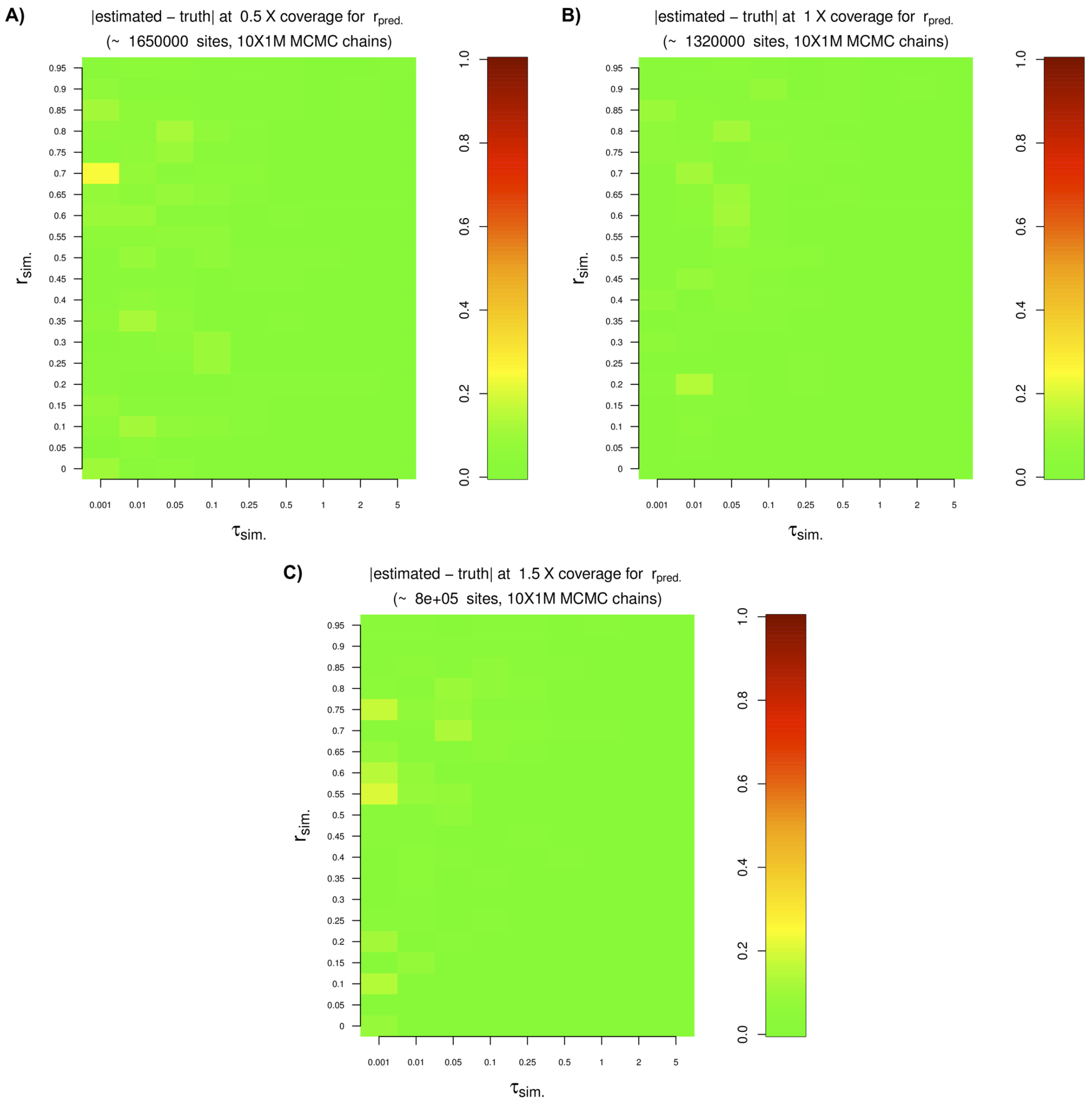


Figure 2.14. Absolute difference between estimated and simulated contamination rates for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). We used a large number of sites and run 10 MCMC chains for 1 million steps each. To ensure convergence, we then selected the chain with the highest posterior probability, and here show estimates from that chain. In all simulations, the anchor drift was set to be equal to the ancient sample drift

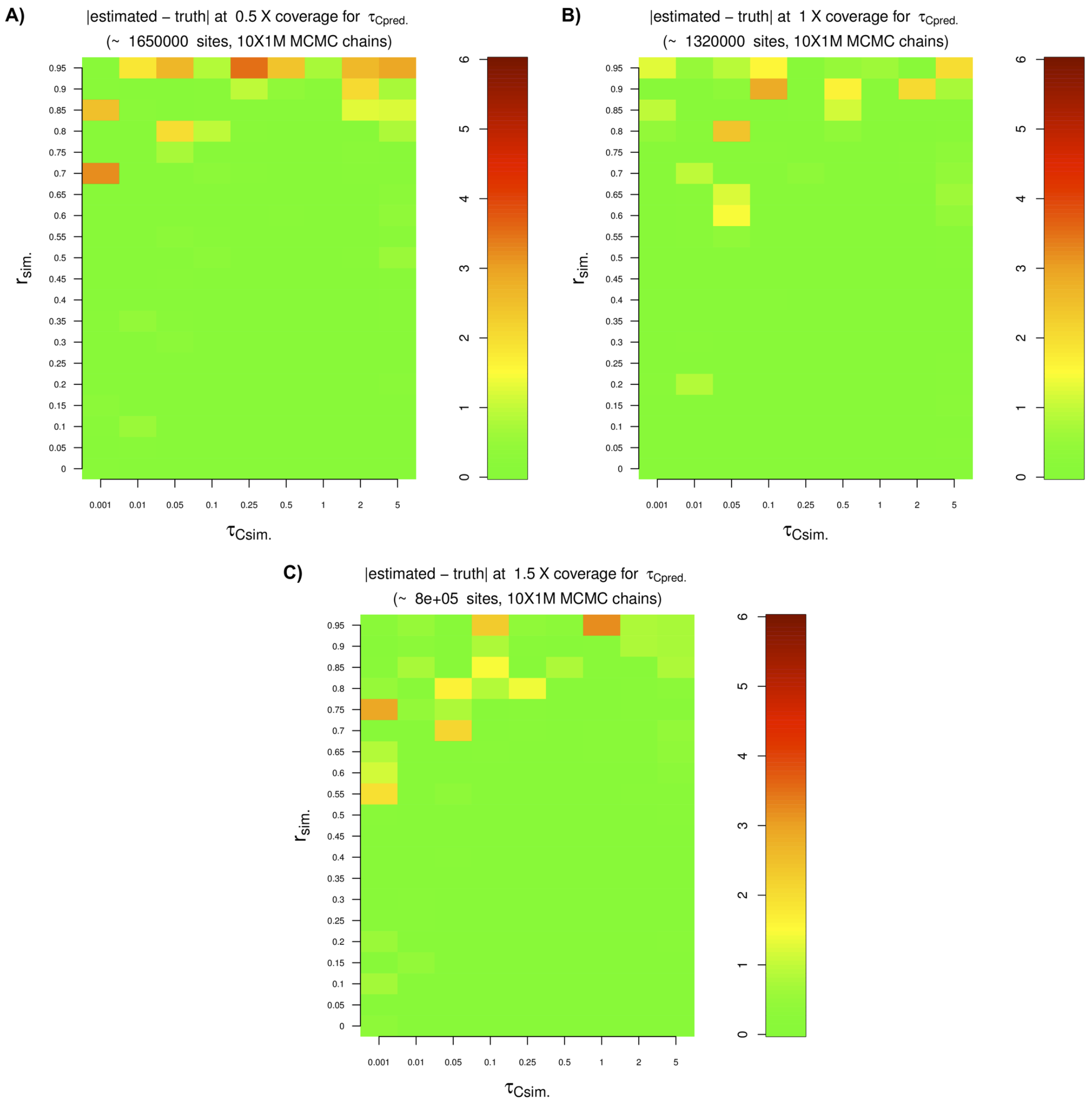


Figure 2.15. Absolute difference between estimated and simulated anchor drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). We used a large number of sites and run 10 MCMC chains for 1 million steps each. To ensure convergence, we then selected the chain with the highest posterior probability, and here show estimates from that chain. In all simulations, the anchor drift was set to be equal to the ancient sample drift

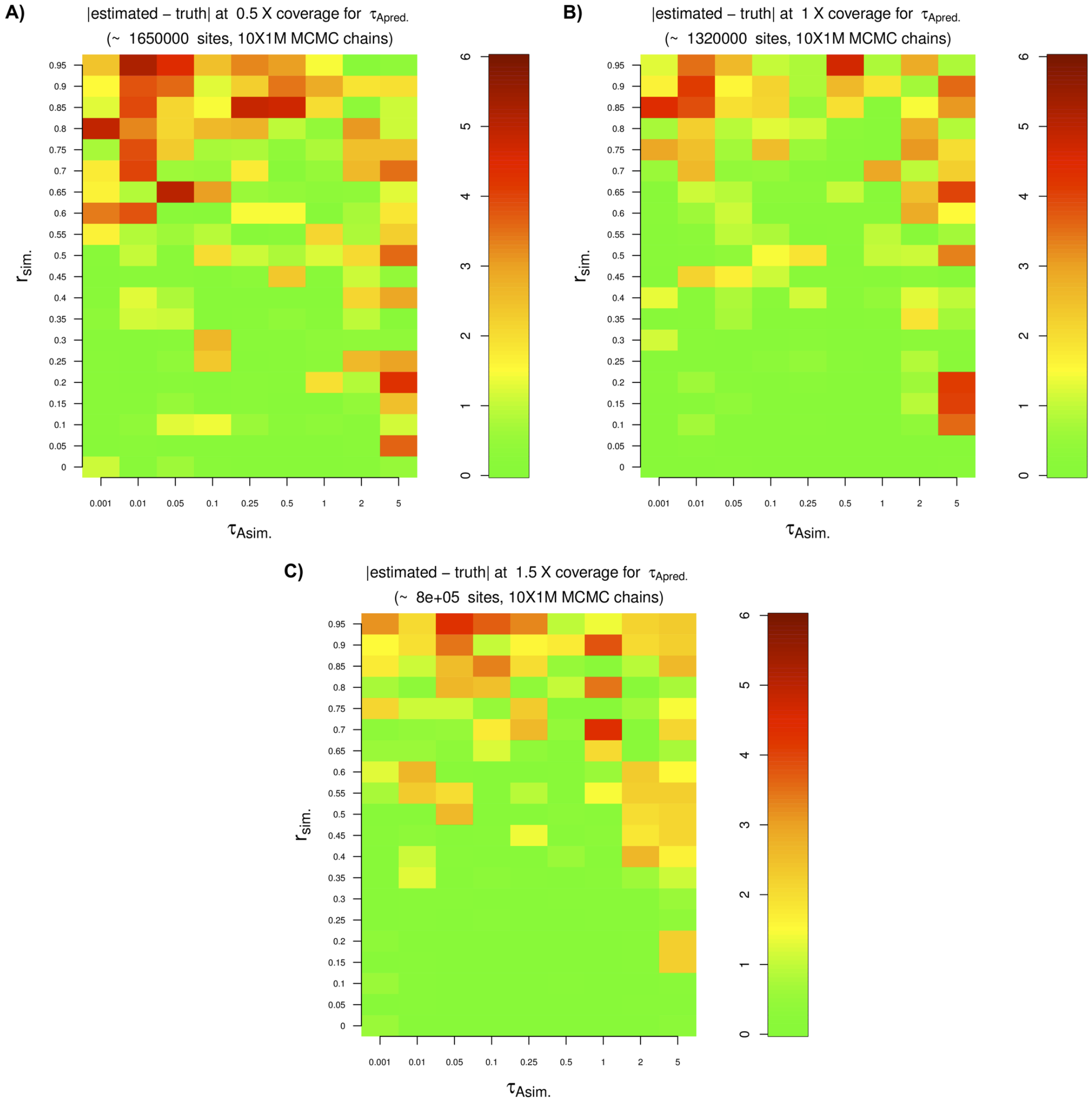


Figure 2.16. Absolute difference between estimated and simulated ancient sample drifts for a variety of anchor drift and contamination scenarios, when coverage is low (0.5X, 1X or 1.5X). We used a large number of sites and run 10 MCMC chains for 1 million steps each. To ensure convergence, we then selected the chain with the highest posterior probability, and here show estimates from that chain. In all simulations, the anchor drift was set to be equal to the ancient sample drift.

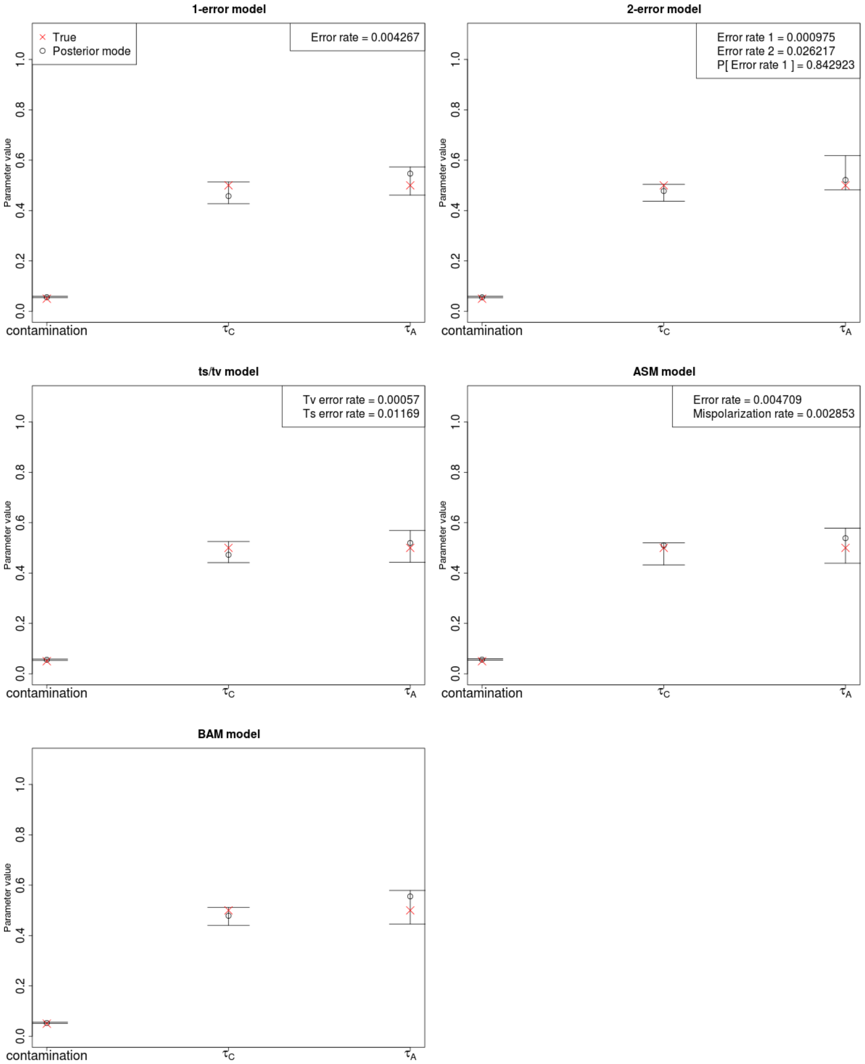


Figure 2.17. Estimation of parameters for a high-coverage ancient DNA genome (30X) simulated under a realistic scenario in which fragments from the ancient and contaminant genome were generated and then mapped to a reference genome. We allowed for multiple substitutions at the same site after the split from chimp, as well as sequencing errors and post-mortem deamination errors at the ends of the fragments. The five panels show results from inferring parameters under five different error rate models. Top-left: single-error model. Top-right: two-error model [22]. Middle-left: model with separate errors for transitions (ts) and tranversions (tv). Middle-right: single-error model with an ancestral state misidentification parameter. Bottom-left: Model in which errors were inferred individually at each site, using base and mapping qualities obtained from the simulated BAM file. Error bars represent 95% posterior intervals.

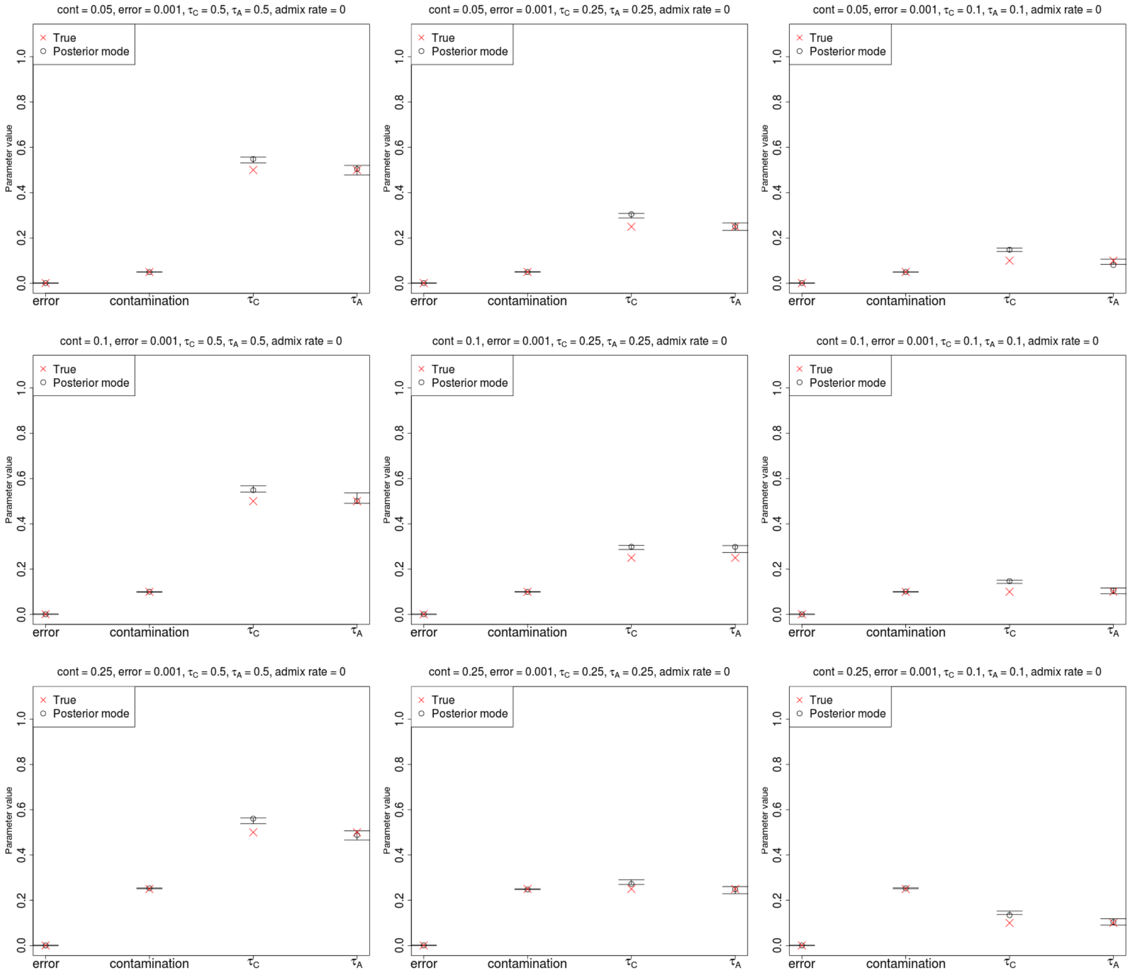


Figure 2.18. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), no admixture and a small anchor population panel (20 haploid genomes). Error bars represent 95% posterior intervals.

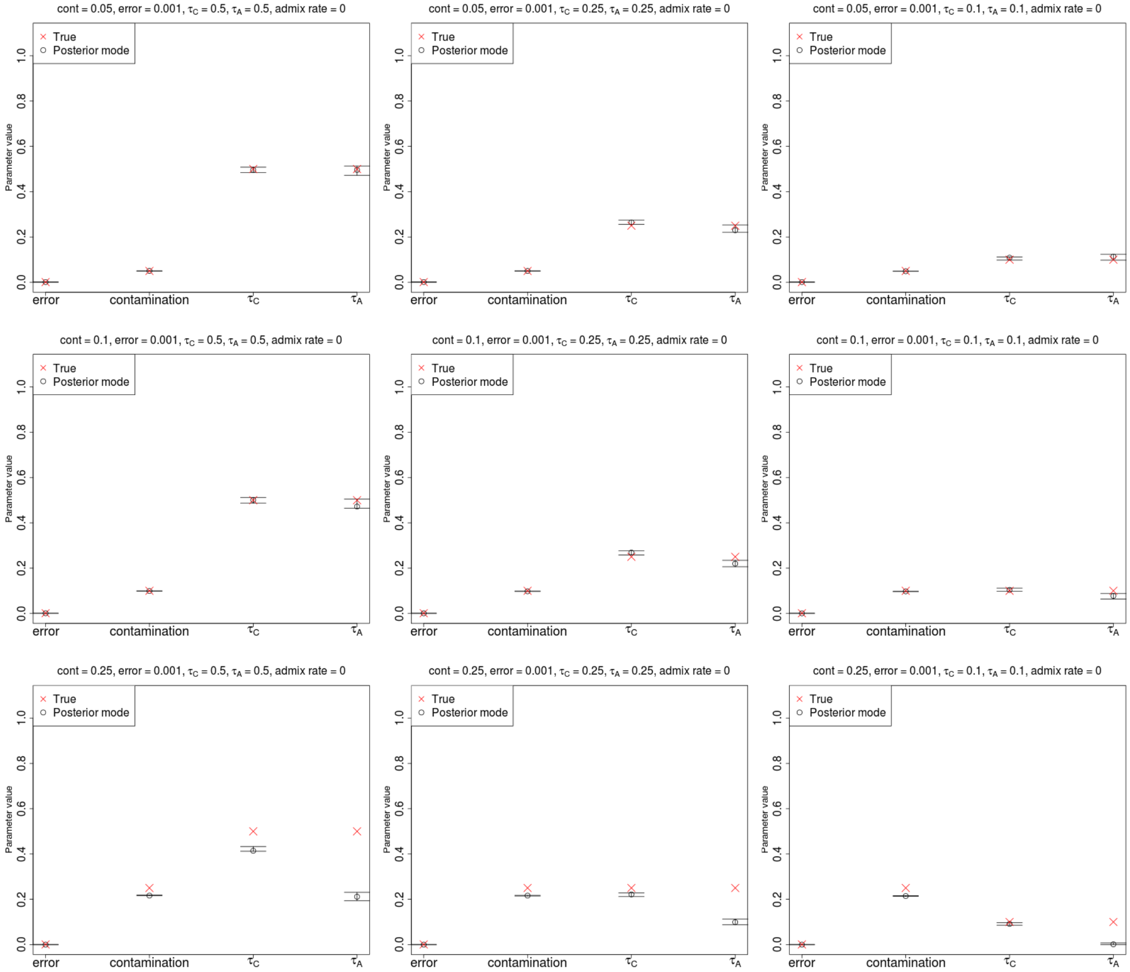


Figure 2.19. Estimation of parameters for a high-coverage ancient DNA genome (30X), when the contaminant fragments are exclusively drawn from a single diploid individual from the contaminant panel. Error bars represent 95% posterior intervals.

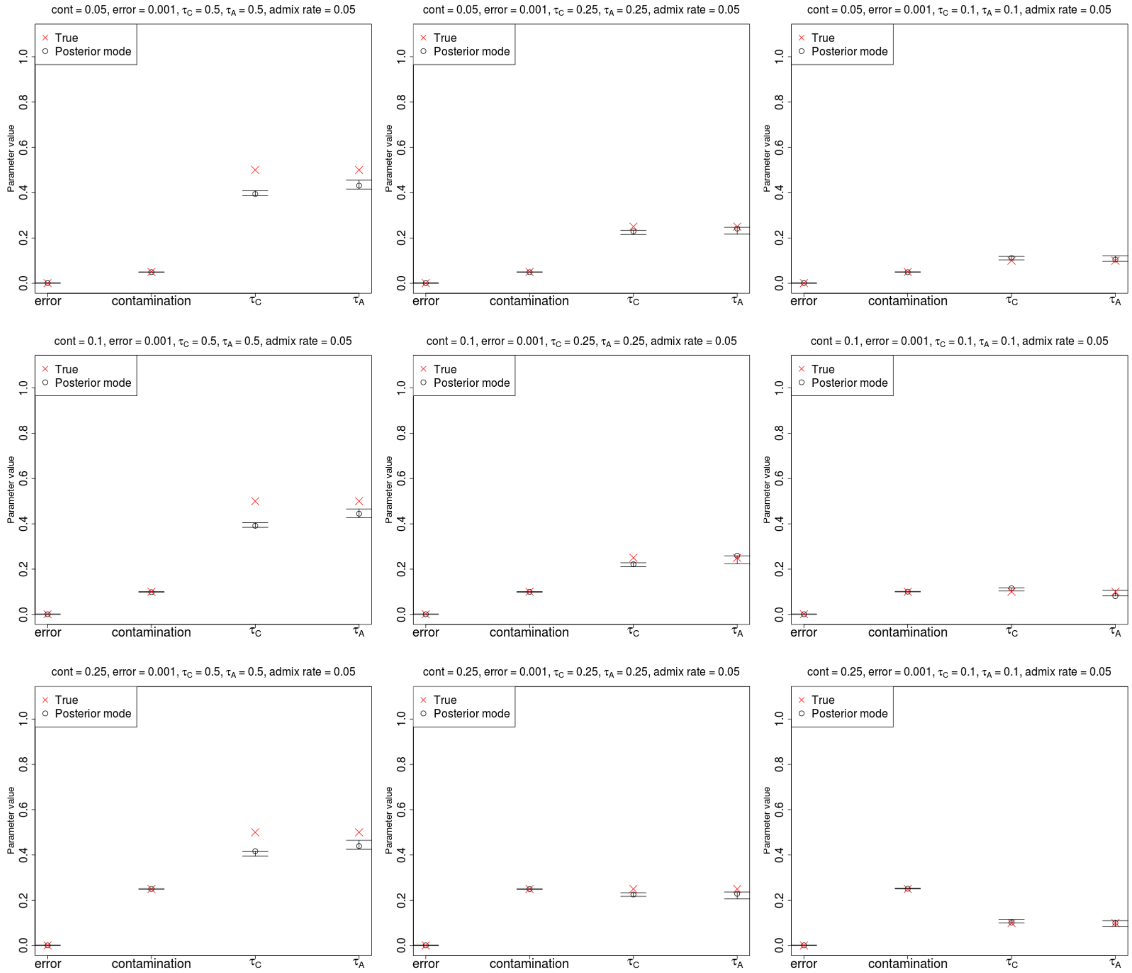


Figure 2.20. Estimation of parameters for a high-coverage ancient DNA genome (30X) with low sequencing error (0.1%), a large anchor population panel (100 haploid genomes) and admixture in the anchor population from the archaic population (5%), using the two-population inference framework, which does not model admixture. Error bars represent 95% posterior intervals.

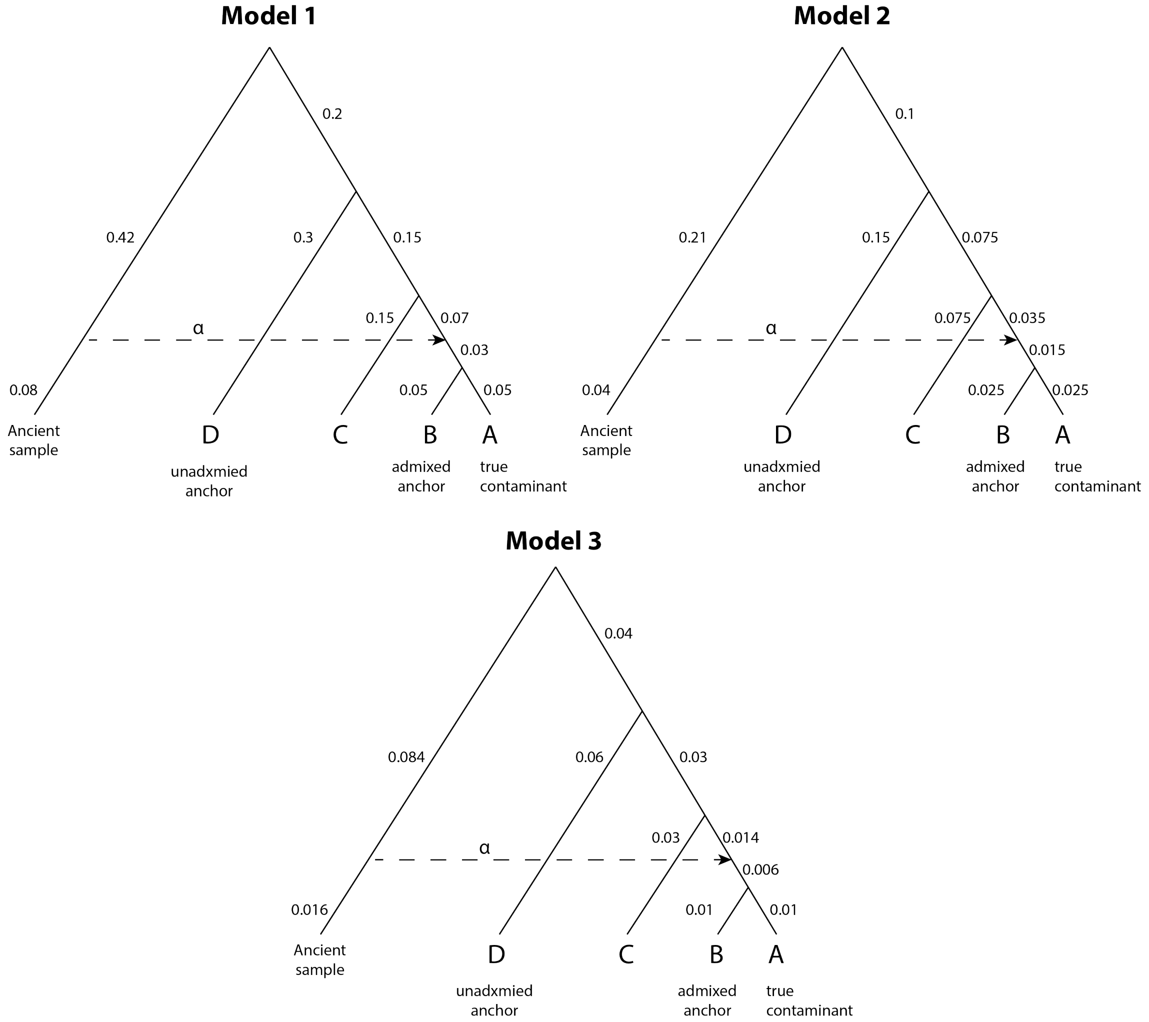
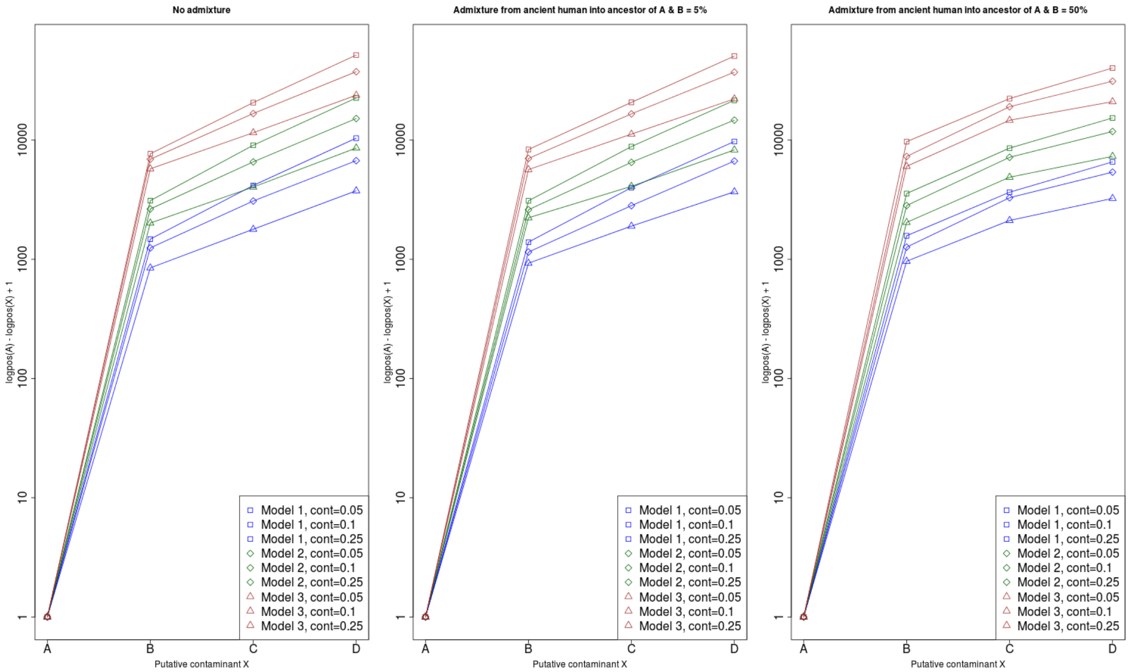


Figure 2.21. Three demographic models used to test the method when the contaminant is misspecified. When testing the two-population method, we set panel A as the true contaminant and panel D as the anchor. When testing the three-population method, we set panel A as the true contaminant, panel D as the unadmixed anchor and panel B as the admixed anchor. The numbers on the branches represent the drift parameters. The parameter α represents the admixture rate from the ancient population into the ancestor of A and B.


No admixture

Admixture from ancient human into ancestor of A & B = 5%

Admixture from ancient human into ancestor of A & B = 50%

Figure 2.22. When testing different putative contaminants, the highest mode of the posterior likelihoods from the MCMC under the two-population model corresponds to the true contaminant (panel A). The y-axis shows the difference between the log-posterior for contaminant panel A and the log-posterior for different candidate contaminant panels (A, B, C, D), so low values correspond to high posterior probabilities for each of the candidates. We added a 1 to the difference to be able to plot the difference on a logarithmic scale. The three panels contain results for three admixture scenarios (from left to right: admixture rate of 0%, 5% and 50%) and each panel shows the difference under different contamination rates and demographic models (see Figure 2.21).

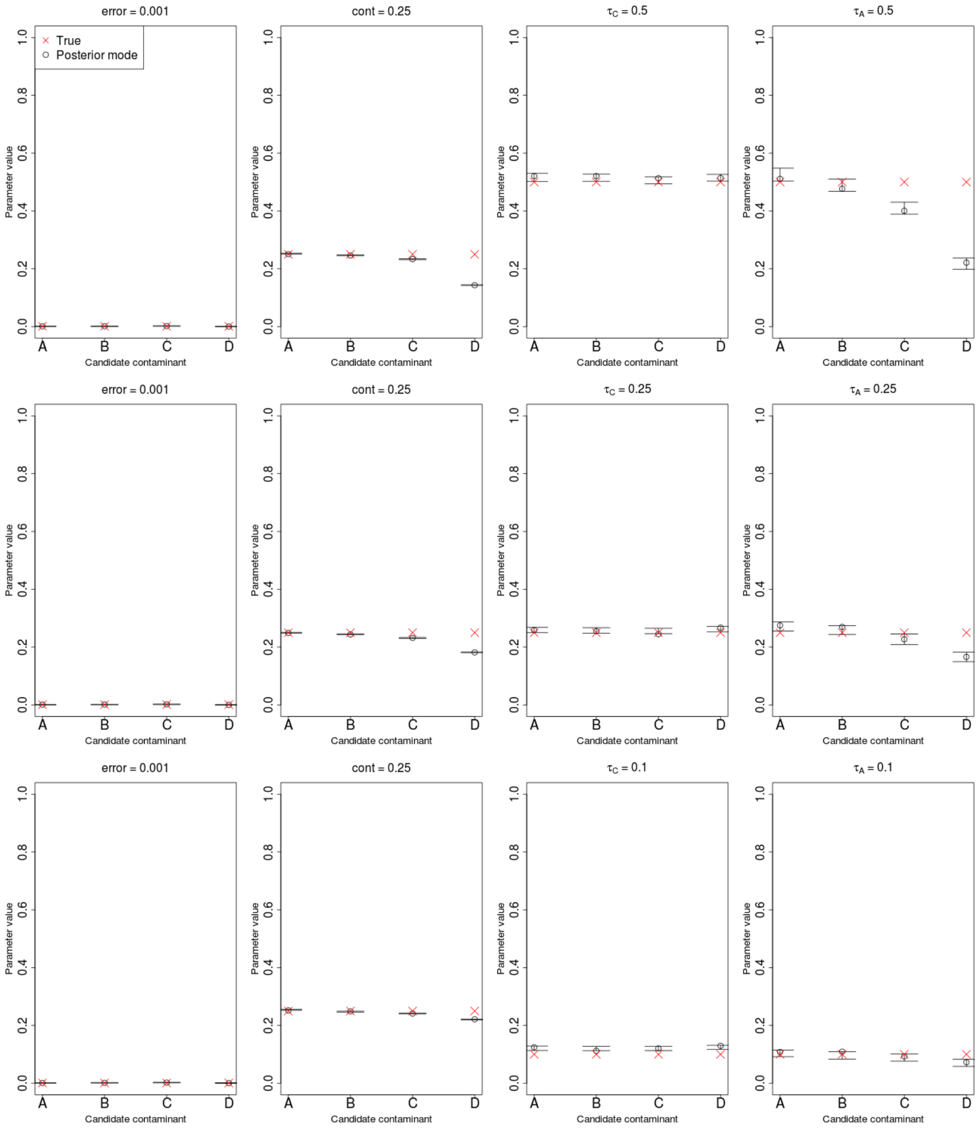


Figure 2.23. Parameters estimates under the two-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 0%. The anchor panel used was panel D (see Figure 2.21).

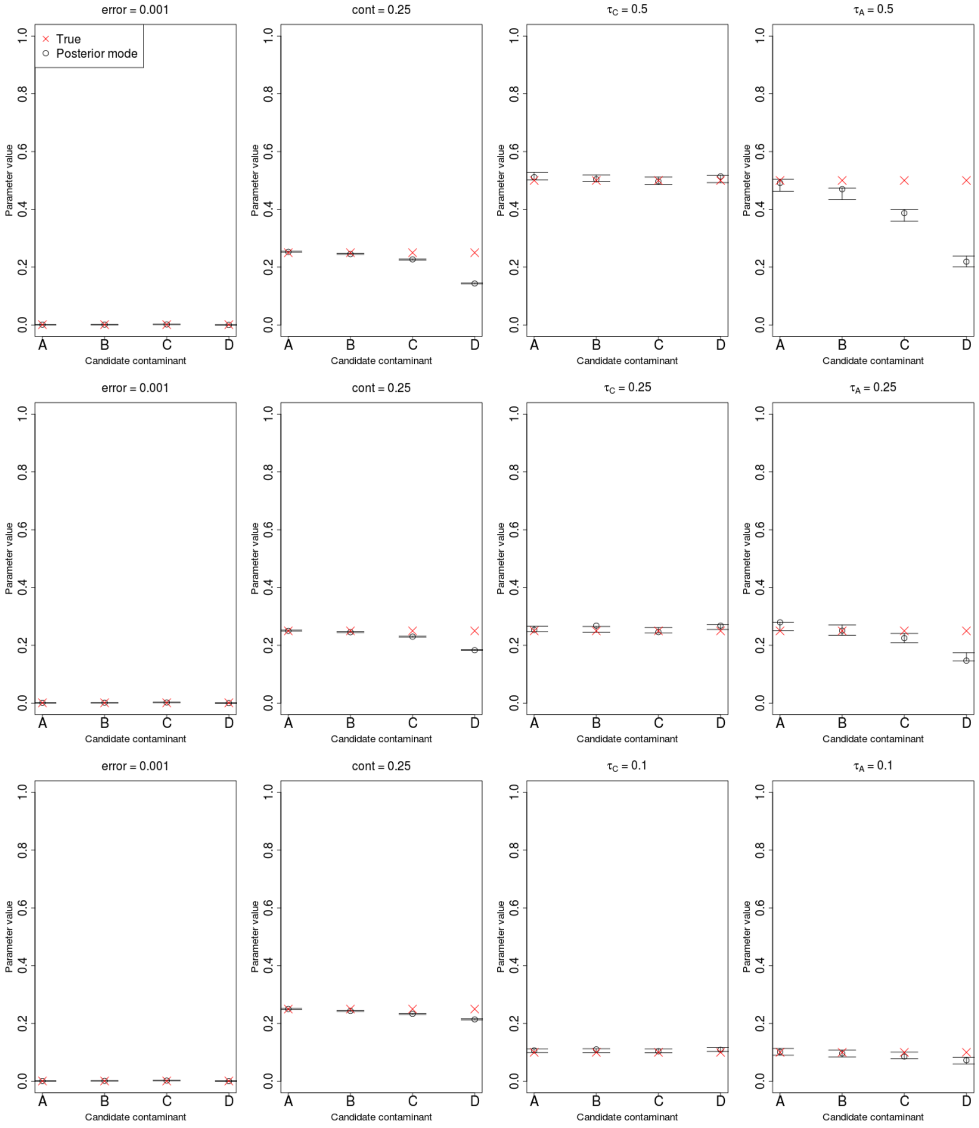


Figure 2.24. Parameters estimates under the two-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 5%. The anchor panel used was panel D (see Figure 2.21).

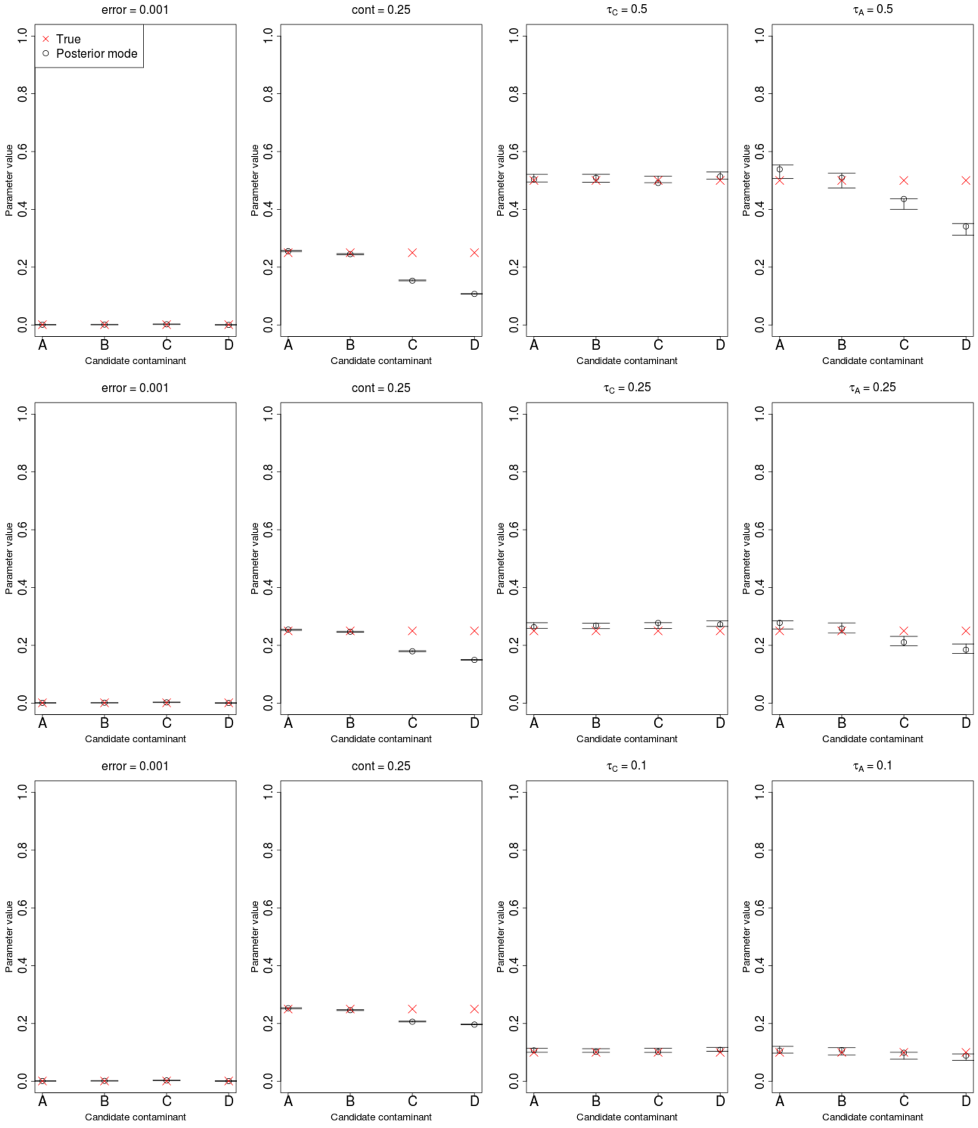


Figure 2.25. Parameters estimates under the two-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 50%. The anchor panel used was panel D (see Figure 2.21).

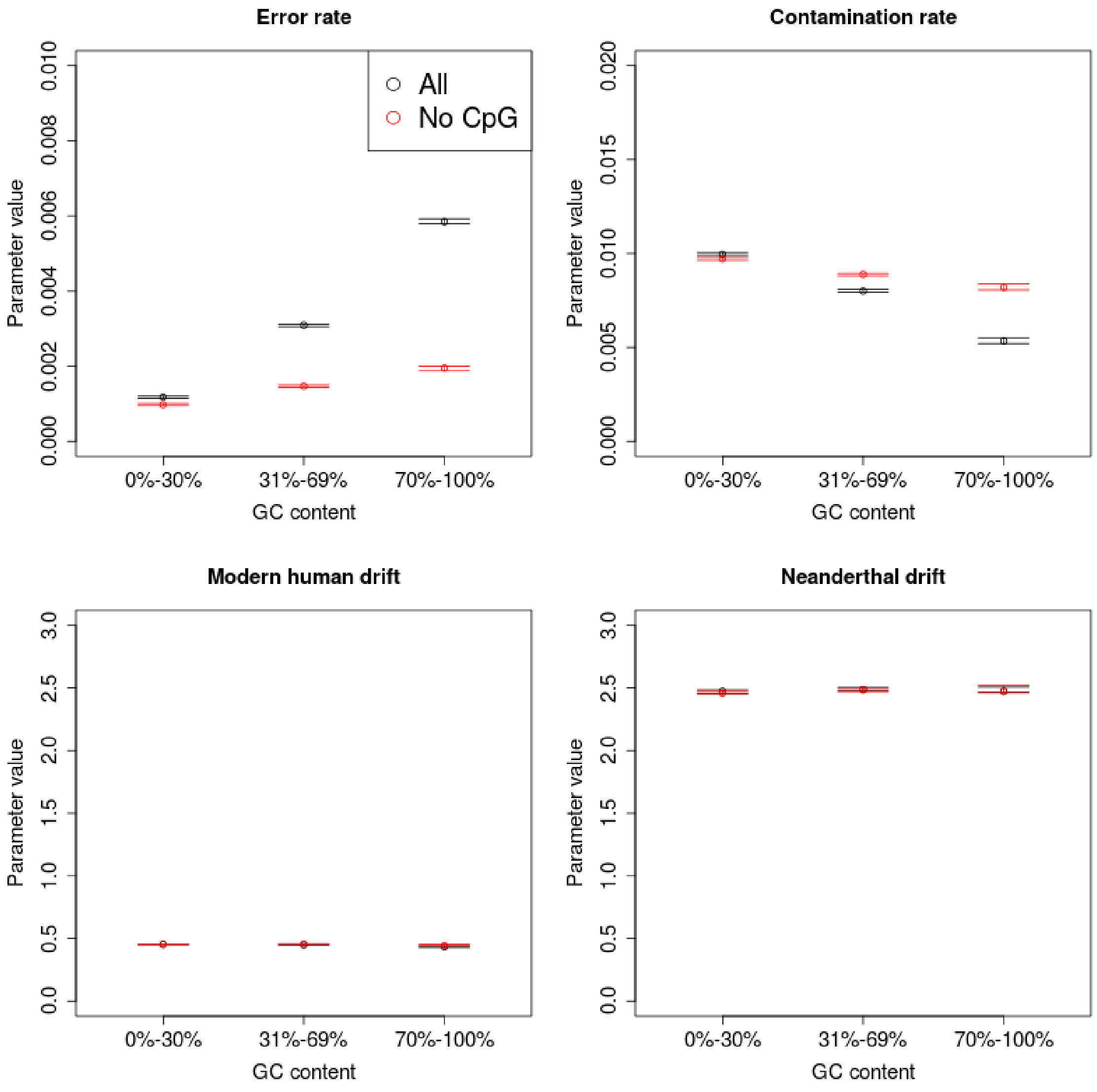


Figure 2.26. Estimation of parameters for the Altai Neanderthal genome across different GC levels using the two-population model, while keeping (black) or removing (red) CpG sites from the input dataset. Error bars represent 95% posterior intervals.

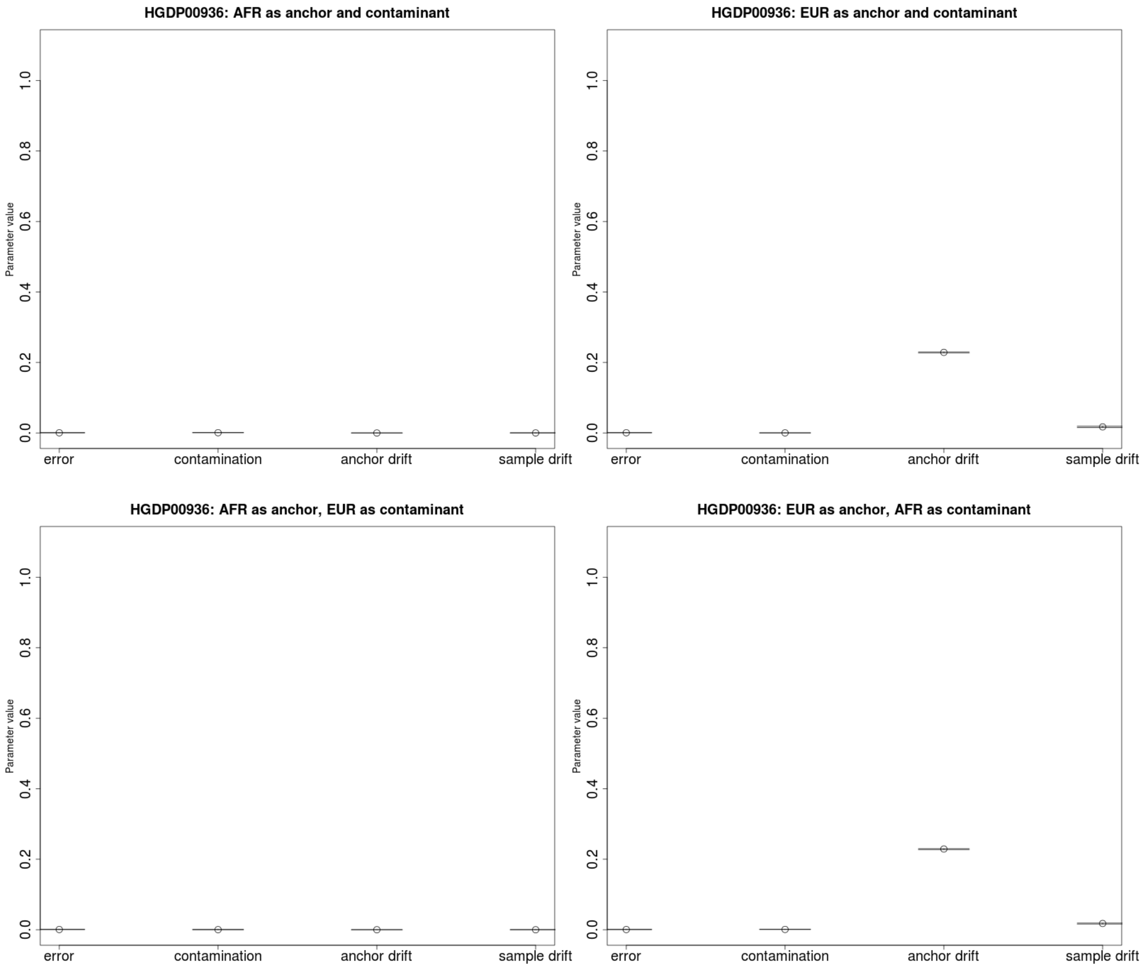


Figure 2.27. We tested one of the Yoruba genomes from [16] and obtain an estimate of 0% contamination, regardless of whether we use Europeans or Africans as the candidate contaminant. The anchor drift time is close to 0 when using Africans as the anchor population, as the sample belongs to that same population, while it is non-zero (= 0.22) when using Europeans. Error bars represent 95% posterior intervals.

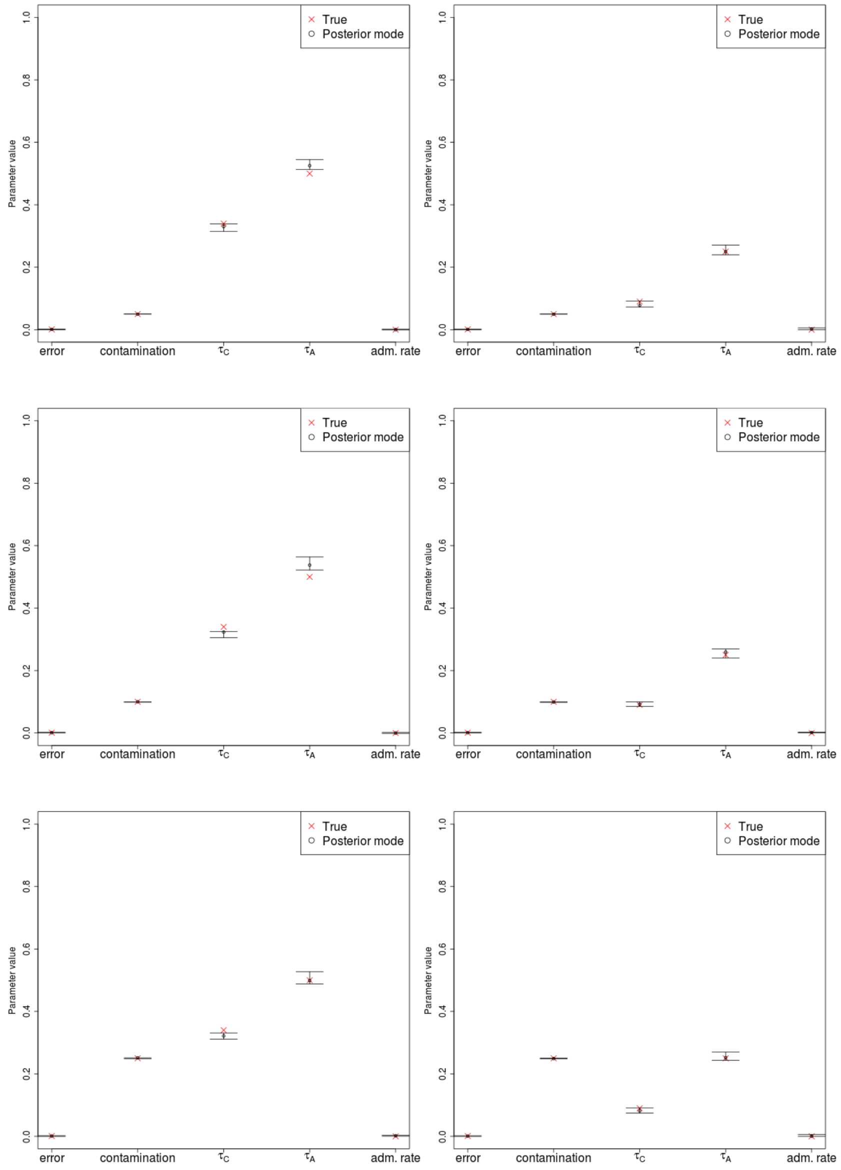


Figure 2.28. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 0%. The prior used for the admixture time was uniform over $[0.06, 0.1]$. Error bars represent 95% posterior intervals.

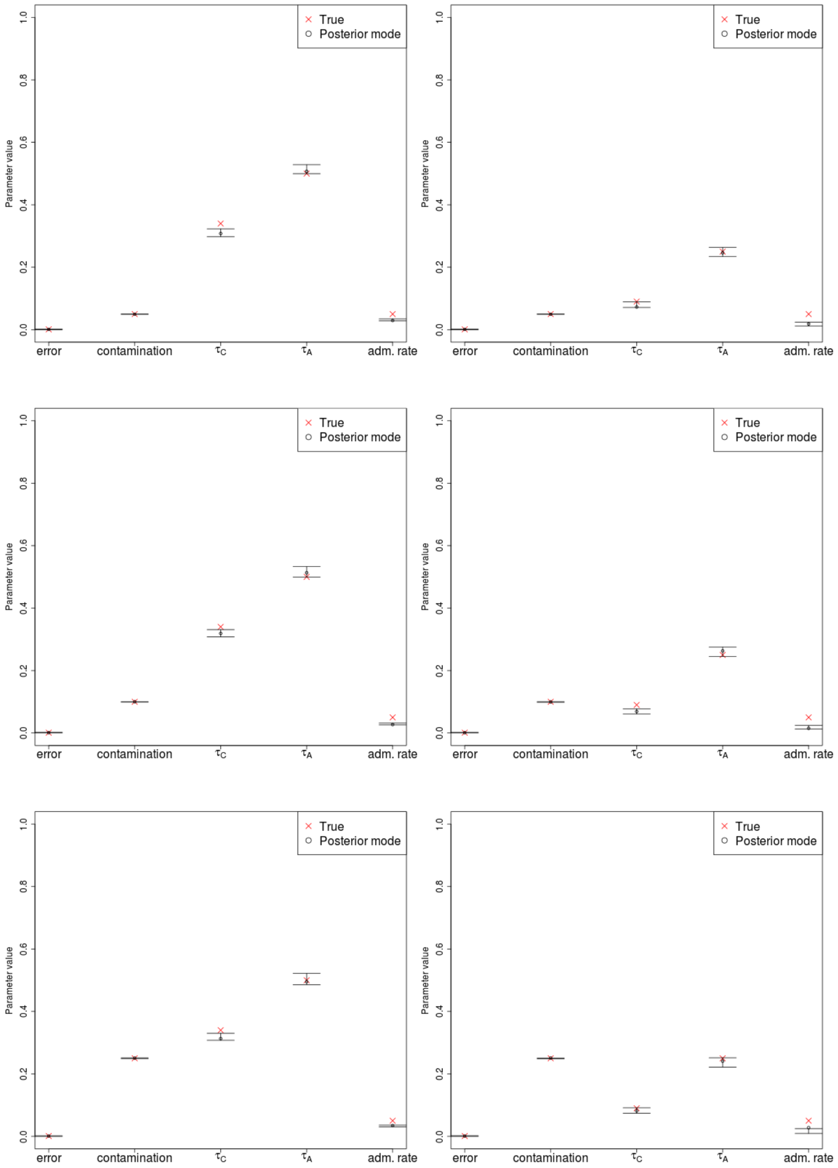


Figure 2.29. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 5% and the admixture time is ancient (0.08 drift units ago). The prior used for the admixture time was uniform over $[0.06, 0.1]$. Error bars represent 95% posterior intervals.

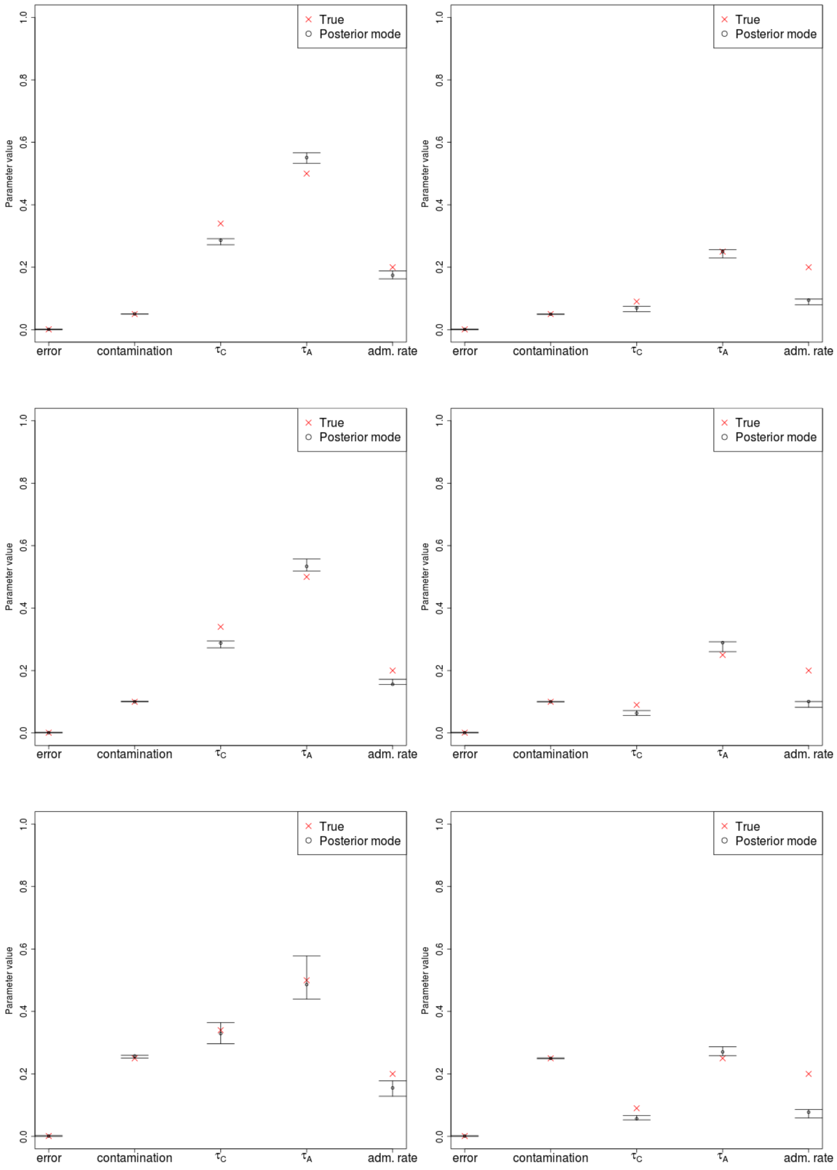


Figure 2.30. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 20% and the admixture time is ancient (0.08 drift units ago). The prior used for the admixture time was uniform over $[0.06, 0.1]$. Error bars represent 95% posterior intervals.

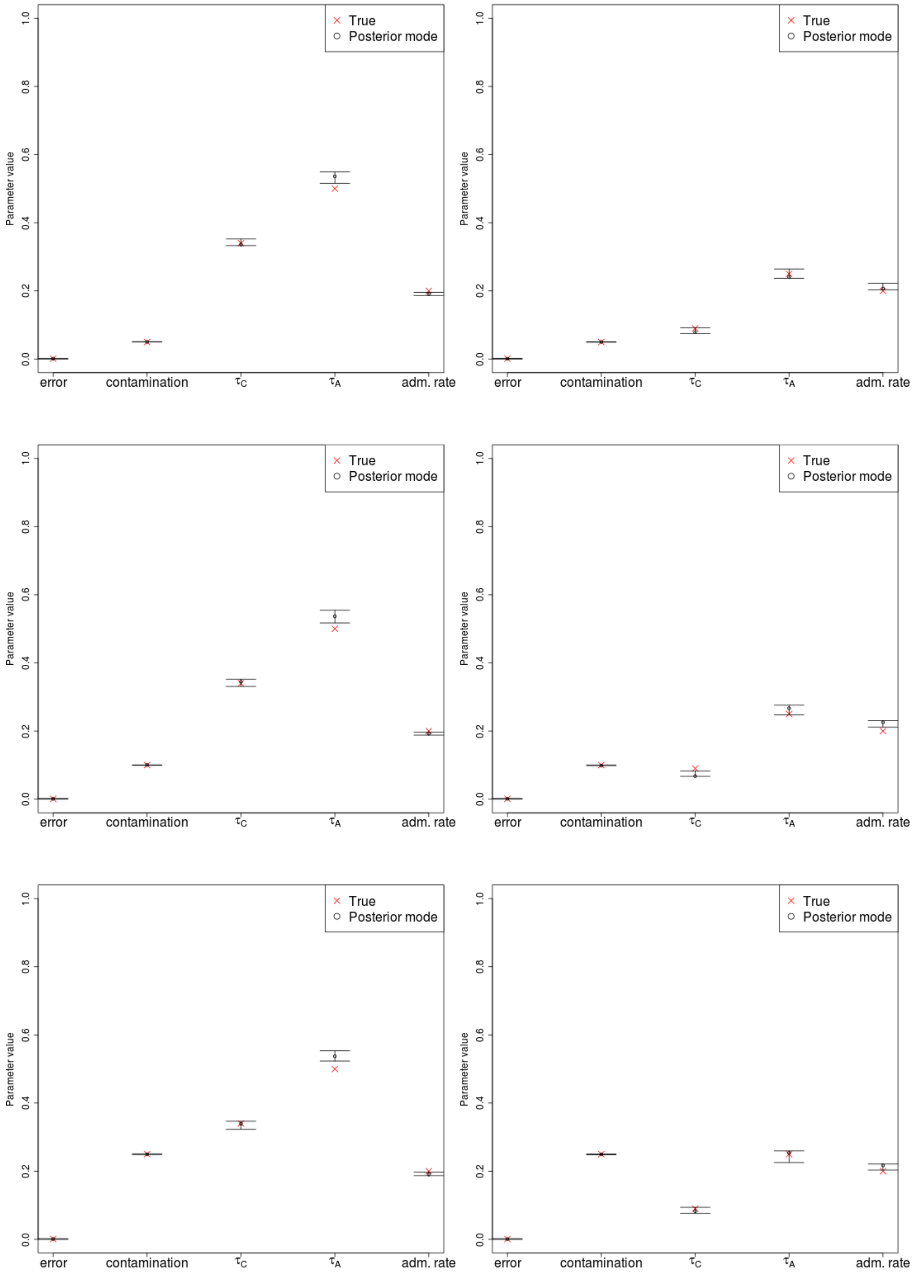


Figure 2.31. Estimation of error, contamination and demographic parameters in various three-population demographic scenarios, where the admixture rate is 20% and the admixture time was recent (0.005 drift units ago). The prior used for the admixture time was uniform over $[0, 0.01]$. Error bars represent 95% posterior intervals.

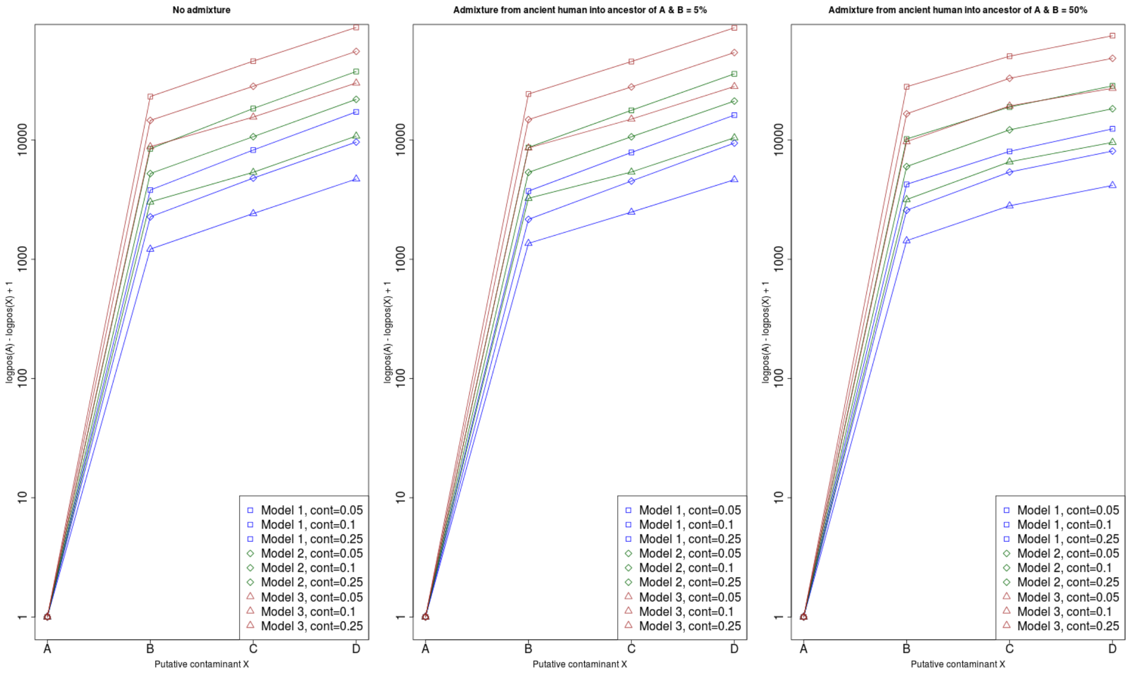


Figure 2.32. When testing different putative contaminants, the highest mode of the posterior likelihoods from the MCMC under the three-population model corresponds to the true contaminant (panel A). The y-axis shows the difference between the log-posterior for contaminant panel A and the log-posterior for different candidate contaminant panels (A, B, C, D), so low values correspond to high posterior probabilities for each of the candidates. We added a 1 to the difference to be able to plot the difference on a logarithmic scale. The three panels contain results for three admixture scenarios (from left to right: admixture rate of 0%, 5% and 50%) and each panel shows the difference under different contamination rates and demographic models (see Figure 2.21).

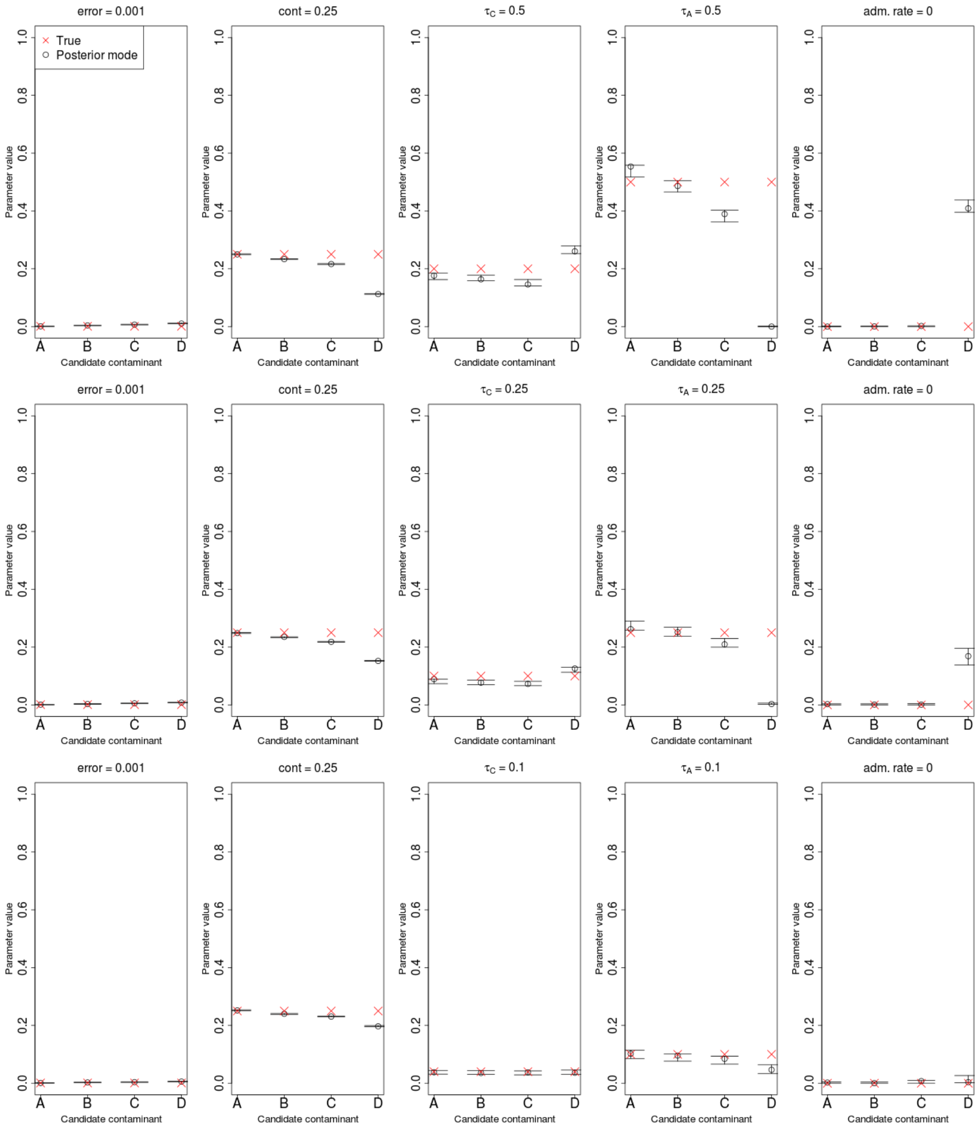


Figure 2.33. Parameters estimates under the three-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 0%. The unadmixed anchor panel used was panel D and the admixed anchor panel was panel B (see Figure 2.21).

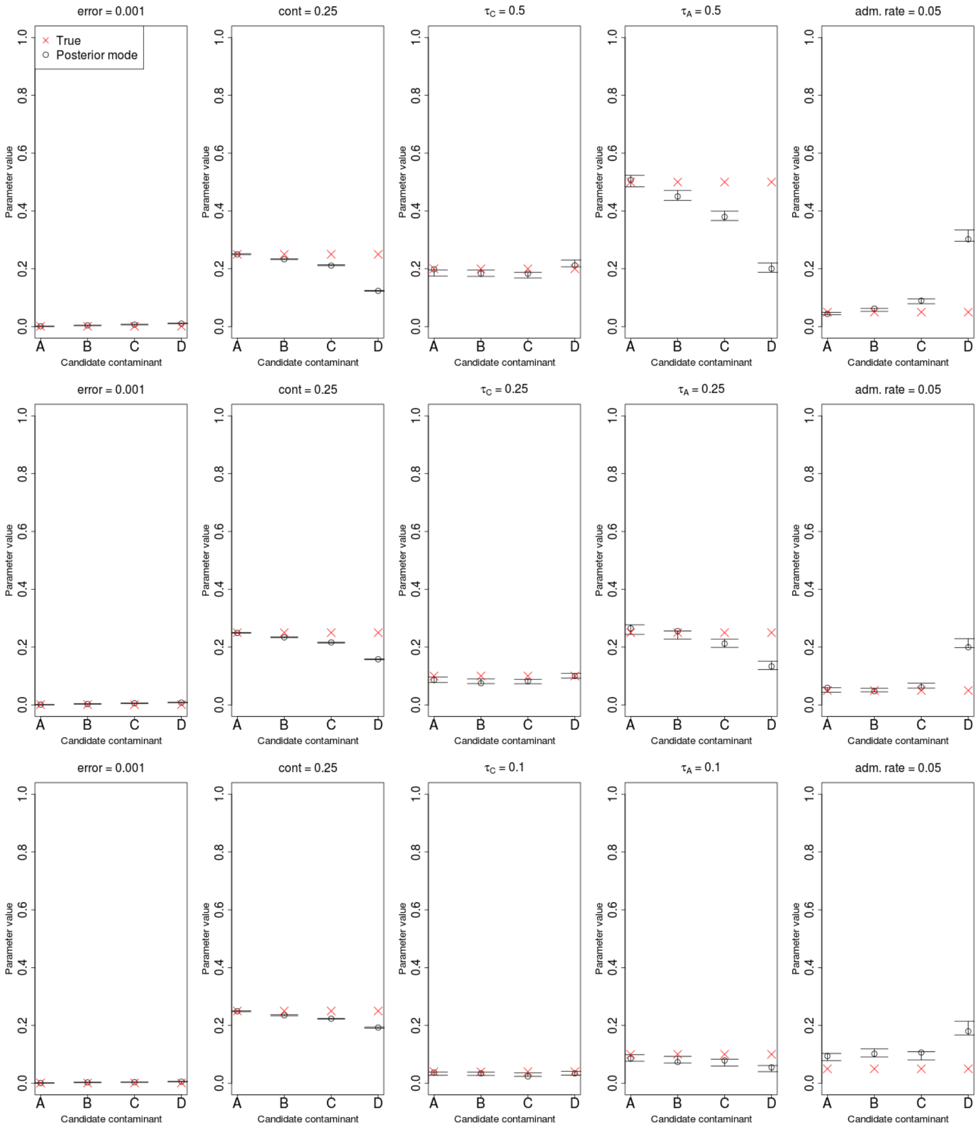


Figure 2.34. Parameters estimates under the three-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 5%. The unadmixed anchor panel used was panel D and the admixed anchor panel was panel B (see Figure 2.21).

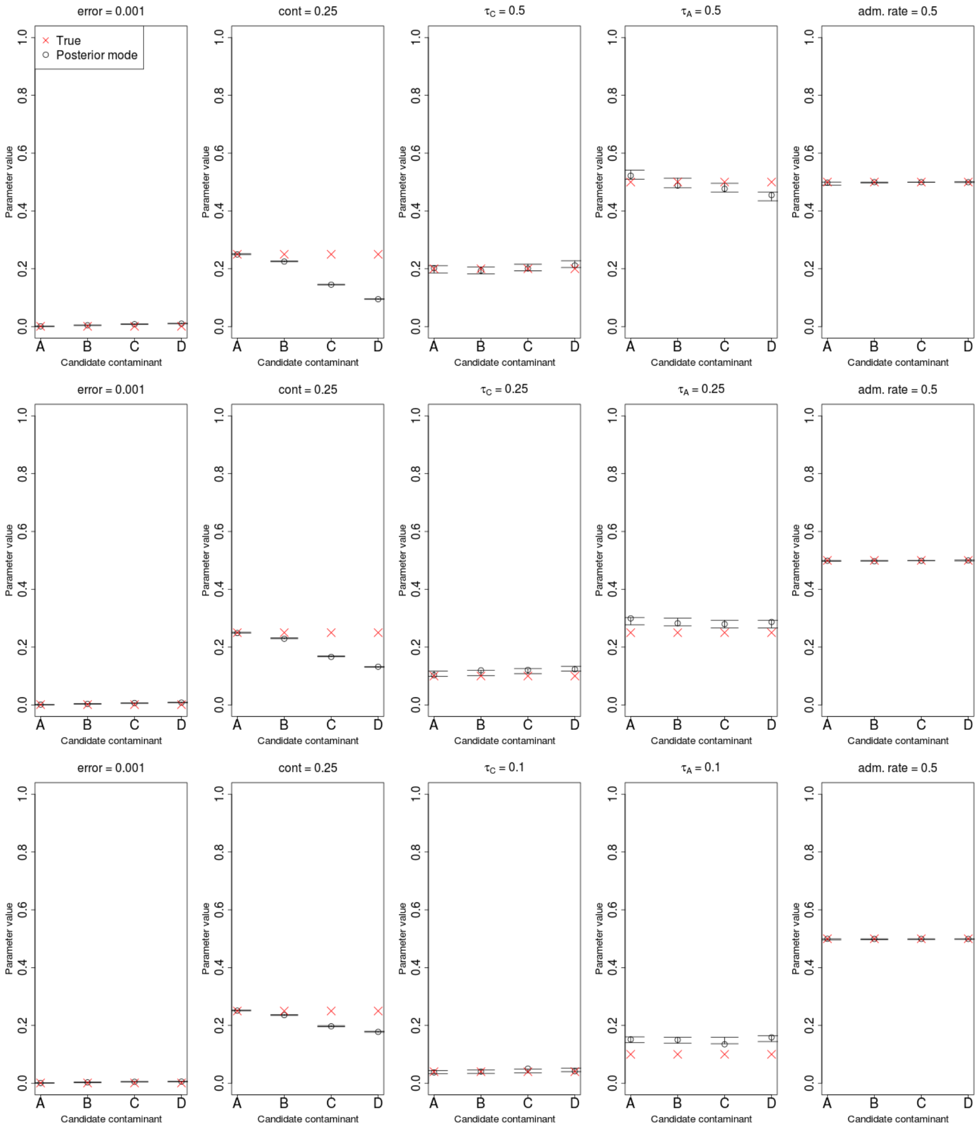


Figure 2.35. Parameters estimates under the three-population model using different putative contaminants, when the true contaminant is panel A. Each row of panels represents a different set of drift parameters, keeping the contamination rate fixed at 25% and the error rate at 0.1%. In this case, the admixture rate from the ancient population to the ancestor of A and B was kept at 50%. The unadmixed anchor panel used was panel D and the admixed anchor panel was panel B (see Figure 2.21).

3. Signatures of archaic adaptive introgression in present-day human populations

Fernando Racimo, Davide Marnetto, Emilia Huerta-Sánchez

3.1 Introduction

There is a growing body of evidence supporting the idea that certain modern human populations admixed with archaic groups of humans after expanding out of Africa. In particular, non-African populations have $1-2\%$ Neanderthal ancestry [8, 16], while Melanesians and East Asians have 3% and 0.2% ancestry, respectively, from Denisovans [26, 2, 16].

Recently, it has become possible to identify the fragments of the human genome that were introgressed and survive in present-day individuals [20, 21]. Researches have also detected which of these introgressed regions are present at high frequencies in some present-day non-Africans but not others. These regions are likely to have undergone positive selection in those populations after they were introgressed, a phenomenon known as adaptive introgression (AI). One particularly striking example of AI is the gene *EPAS1* [67] which confers a selective advantage in Tibetans by making them less prone to hypoxia at high altitudes [32]. The selected Tibetan haplotype is known to have been introduced in the human gene pool by Denisovans or a population closely related to them [33, 34].

In this study, we use simulations to assess the power to detect AI using different summary statistics that do not require the introgressed fragments to be identified *a priori*. Some of these are inspired by the signatures observed in *EPAS1*, which contains an elevated number of sites with alleles uniquely shared between the Denisovan genome and Tibetans. We then apply these statistics to real human genomic data from phase 3 of the 1000 Genomes Project [45], to detect AI in human populations, and find candidate genes. While these statistics are sensitive to adaptive introgression, they may also be sensitive to other phenomena that generate genomic patterns similar to those generated by AI, like ancestral population structure and incomplete lineage sorting. These processes, however, should not generate long regions of the genome where haplotypes from the source and the recipient population are highly similar. To assess whether the candidates we found are truly generated by AI, we explored the haplotype structure of some of the most promising candidates, and used a probabilistic method [23] that infers introgressed segments along the genome by looking at the spatial arrangement of SNPs that are consistent with introgression. This allows us to verify that the candidate regions contain introgressed haplotypes at high frequencies: a hallmark of AI.

3.2 Methods

Summary statistics sensitive to adaptive introgression

Several statistics have been previously deployed to detect AI events (reviewed in Racimo et al. [31]). We briefly describe these below, as well as three new statistics tailored specifically to find this signal (Table 3.1). One of the simplest approaches consists of applying the D statistic [8, 68] locally over windows of the genome. The D statistic was originally applied to compare a single human genome against another human genome, so as to detect excess shared ancestry between one of the genomes and a genome from an outgroup population. Application of this statistic comparing non-Africans and Africans served as one of the pieces of evidence in support of Neanderthal admixture into non-Africans. However, it can also be computed from large panels of multiple individuals instead of single genomes. This form of the D statistic has been applied locally over windows of the genome, to detect regions of excess shared ancestry between an admixed population and a source population [69, 70].

The D statistic, however, can be confounded by local patterns of diversity, as regions of low diversity may artificially inflate the statistic even when a region was not adaptively introgressed. To correct for this, Martin et al. [71] developed a similar statistic called f_D which is less sensitive to differences in diversity along the genome. Both of these patterns exploit the excess relatedness between the admixed and the source population.

AI is also expected to increase linkage disequilibrium (LD), as an introgressed fragment that rises in frequency in the population will have several closely linked loci that together will be segregating at different frequencies than they were in the recipient population before admixture. Thus, two well-known statistics that are informative about the amount of LD in a region - D' and r^2 - could also be informative about adaptive introgression. To apply them over regions of the genome, we can take the average of each of the two statistics over all SNP pairs in a window. In the section below, we calculate these statistics in two ways: a) using the introgressed population only ($D'[intro]$ and $r^2[intro]$), and b) using the combination of the introgressed and the non-introgressed populations ($D'[comb]$ and $r^2[comb]$).

We also introduce three new statistics that one would expect, *a priori*, to be particularly effective at identifying windows of the genome that are likely to have undergone adaptive introgression. First, in a region under adaptive introgression, one would expect the divergence between an individual from the source population and an admixed individual to be smaller than the divergence between an individual from the source population and a non-admixed individual. Thus, one could take the ratio of these two divergences over windows of the genome. One can then take the average of this ratio over all individuals in the admixed and non-admixed panels. This average should be larger if the introgressed haplotype is present in a large number of individuals of the admixed population. We call this statistic R_D.

Second, for a window of arbitrary size, let $U_{A,B,C}(w, x, y)$ be defined as the number of sites where a sample C from an archaic source population (which could be as small as a single diploid individual) has a particular allele at frequency y, and that allele is at a frequency smaller than w in a sample A of a population but larger than x in a sample B of another population (Figure 3.1). In other words, we are looking for sites that contain alleles shared between an archaic human genome and a test population, but absent or at very low frequencies in an outgroup (usually non-admixed) population. Below, we denote panels A, B and C as the “outgroup", “target" and “bait" panels, respectively. For example, suppose we are looking for Neanderthal adaptive introgression in the Han Chinese (CHB). In that case, we can consider CHB as our target panel, and use Africans as the outgroup panel and a single Neanderthal genome as the bait. If $U_{AFR,CHB,Nea}(1\%, 20\%, 100\%) = 4$ in a window of the genome, that means there are 4 sites in that window where the Neanderthal genome is homozygous for a particular allele and that allele is present at a frequency smaller than 1% in Africans but larger than 20% in Han Chinese. In other words, there are 4 sites that are uniquely shared at more than 20% frequency between Han Chinese and Neanderthal, but not with Africans.

This statistic can be further generalized if we have samples from two different archaic populations (for example, a Neanderthal genome and a Denisova genome). In that case, we can define $U_{A,B,C,D}(w, x, y, z)$ as the number of sites where the archaic sample C has a particular allele at frequency y and the archaic sample D has that allele at frequency z, while the same allele is at a frequency smaller than w in an outgroup panel A and larger than x in a target panel B (Figure 3.2). For example, if we were interested in looking for Neanderthal-specific AI, we could set $y = 100\%$ and $z = 0\%$, to find alleles uniquely shared with Neanderthal, but not Denisova. If we were interested in archaic alleles shared with both Neanderthal and Denisova, we could set $y = 100\%$ and $z = 100\%$.

Another statistic that we found to be useful for finding AI events is $Q95_{A,B,C}(w, y)$, and is here defined as the 95^{th} percentile of derived frequencies in an admixed sample B of all SNPs that have a derived allele frequency y in the archaic sample C, but where the derived allele is at a frequency smaller than w in a sample A of a non-admixed population (Figure 3.1). For example, $Q95_{AFR,CHB,Nea}(1\%, 100\%) = 0.65$ means that if one computes the 95% quantile of all the Han Chinese derived allele frequencies of SNPs where the Neanderthal genome is homozygous derived and the derived allele has frequency smaller than 1% in Africans, that quantile will be equal to 0.65. As before, we can generalize this statistic if we have a sample D from a second archaic population. Then, $Q95_{A,B,C,D}(w, y, z)$ is the 95^{th} percentile of derived frequencies in the sample B of all SNPs that have a derived allele frequency y in the archaic sample C and derived allele frequency z in the archaic sample D, but where the derived allele is at a frequency smaller than w in the sample A (Figure 3.2).

In the section below, we evaluate the sensitivity and specificity of all these statistics using simulations. We also evaluate the effect of adaptive introgression on a common statistic that is indicative of population variation - expected heterozygosity (Het), as this statistic was previously found to be affected by archaic introgression in a serial founder model of human history [72]. We measured Het as the average of 2*p*(1-p) over all sites in a window, where p is the sample derived allele frequency in the introgressed population.

Simulations

None of these statistics have been explicitly vetted under scenarios of AI so far, though the performance of D and f_D has been previously evaluated for detecting local introgression [71]. Therefore, we aimed to test how each of the statistics described above performed in detecting AI. We began by simulating a three population tree in Slim [73] with constant $N_e = 10,000$, mutation rate equal to $1.5 * 10^{-8}$ per bp per generation, recombination rate equal to 10^{-8} per bp per generation, and split times emulating the African-Eurasian and Neanderthal-modern human split times (4,000 and 16,000 generations ago, respectively). We allowed for admixture between the most distantly diverged population and one of the closely related sister populations, at different rates: 2%, 10% and 25% (Figure 3.3.A). This is meant to represent Neanderthal admixture into Eurasians, with Africans as the non-admixed population. Under each of the three admixture rate scenarios, we simulated regions that were evolving neutrally, regions where the central SNP was under weak positive additive selection ($s = 0.01$) and regions with a central SNP under strong selection ($s = 0.1$).

We also tested how the statistics perform at detecting adaptive introgression when the alternative model is not a neutral introgression model, but a neutral model with ancestral structure (Figure 3.3.G). We followed a model described in Huerta-Sanchez et al (2014) and simulated a population in which an African population splits from archaic humans before Eurasians, but is allowed to exchange migrants with them. Afterwards, we split Eurasians and archaic humans. At that point, we stop the previous migration and only allow for migration between the Eurasian and African populations until the present, at double the previous rate. This is meant to generate loci where Eurasians and archaic humans share a more recent common ancestor with each other than with Africans, but because of ancient shared ancestry, not recent introgression. We simulated 3 scenarios, in which we set the per-generation ancient(recent) migration rate to be 0.01(0.02), 0.001(0.002) and 0.0001(0.0002). We call these the strong-, medium-, and weak-migration scenarios, respectively. The stronger the migration, the weaker the ancestral structure, as archaic-shared segments in Eurasians will tend to be removed by migration with Africans.

Plotting haplotype structure

The *Haplostrips* software (Marnetto et al. in prep.) was used to produce plots of haplotypes at candidate regions for AI. This software displays each SNP within a predefined region as a column, while each row represents a phased haplotype. Each haplotype is labeled with a color that corresponds to the 1000 Genomes panel of its carrier individual. The haplotypes are ordered by decreasing similarity to a reference that contains all derived alleles found in the archaic genome (Altai Neanderthal or Denisova), so that haplotypes with more derived alleles shared with the archaic population are at the top of the plot. Derived alleles are represented as black spots and ancestral alleles are represented as white spots. Variant positions were filtered out when the site in the archaic genome had mapping quality less than 30 or genotype quality less than 40, or if the minor allele had a population frequency smaller than 5% in each of the present-day human populations included in the plot.

Hidden Markov Model

As haplotypes could look archaic simply because of ancestral structure or incomplete lineage sorting, we used a Hidden Markov Model (HMM) [23] to verify that our candidate regions truly had archaic introgressed segments. This procedure also allowed us to confirm which of the archaic genomes was closest to the original source of introgression, as using a distant archaic source as input (for example, the Denisova genome when the true source is closest to the Neanderthal genome) produced shorter or less frequent inferred segments in the HMM output than when using the closer source genome.

The HMM we used requires us to specify a prior for the admixture rate. We tried two priors: 2% and 50%. The first was chosen because it is consistent with the genome-wide admixture rate for Neanderthals into Eurasians. The second, larger, value was chosen because each candidate region should *a priori* have a larger probability of being admixed, as they were found using statistics that are indicative of admixture in the first place. We observe almost no differences in the number of haplotypes inferred using either value. However, the larger prior leads to longer and less fragmented introgressed chunks, as the HMM is less likely to transition into a non-introgressed state between two introgressed states, so all figures we show below were obtained using a 50% admixture prior. The admixture time was set to 1,900 generations ago and the recombination rate was set to $2.3 * 10^{-8}$ per bp per generation. A tract was called as introgressed if the posterior probability for introgression was higher than 90%.

Testing for enrichment in genic regions

To test for whether uniquely shared archaic alleles at high frequencies were enriched in genic regions of the genome, we looked at archaic alleles at high frequency in any of the Non-African panels that were also at low frequency ($< 1\%$) in Africans. As background, we used all archaic alleles that were at any frequency larger than 0 in the same Non-African populations, and that were also at low frequency in Africans. We then tested whether the high-frequency archaic alleles tended to occur in genic regions more often than expected.

SNPs in introgressed blocks will tend to cluster together and have similar allele frequencies, which could cause a spurious enrichment signal. To correct for the fact that SNPs at similar allele frequencies will cluster together (as they will tend to co-occur in the same haplotypes), we performed linkage disequilibrium (LD) pruning using two methods. In one (called “LD-1"), we downloaded the approximately independent European LD blocks published in ref. [74]. For each set of high frequency derived sites, we randomly sampled one

SNP from each block. In a different approach (called “LD-2"), for each set of high frequency derived sites, we subsampled SNPs such that each SNP was at least 200 kb apart from each other. We then tested these two types of LD-pruned SNP sets against 1000 SNP sets of equal length that were also LD-pruned and that were obtained randomizing frequencies and collecting SNPs in the same ways as described above.

3.3 Results

Simulations

Statistics based on shared allele configurations

We tested the performance of the statistics described above under scenarios of adaptive introgression. Figures 3.4, 3.5 and 3.6 show the distribution of statistics that rely on patterns of shared allele configurations between source and introgressed populations, for different choices of the selection coefficeint s, and under 2%, 10% and 25% admixture rates, respectively. For $Q95_{A,B,C}(w, 100\%)$ and $U_{A,B,C}(w, x, 100\%)$, we tested different choices of w (1%, 10%) and x (0%, 20%, 50% and 80%). Some statistics, like f_D and $Q95_{A,B,C}(1\%, 100\%)$ show strong separation between the selection regimes, while others, like $U_{A,B,C}(1\%, 0\%, 100\%)$, are not as effective.

LD-based statistics

In turn, Figure 3.7 shows the distribution of LD-based statistics under different selection and admixture rate regimes. Note that while $D'[intro]$, $D'[comb]$ and $r^2[comb]$ are generally increased by adaptive introgression, this is not the case with $r^2[intro]$ under strong selection and admixture regimes. This is because r^2 will tend to decrease if the minor allele frequency is very small, which will occur if this frequency is only measured in the population undergoing adaptive introgression. In general, these statistics do not seem to be as powerful for detecting AI as allele configuration statistics like U or $Q95$.

Receiving operator curves

In Figures 3.8 and 3.9, we plot receiving operator curves (ROC) of all these statistics, for various selection and admixture regimes. In general, $Q_{A,B,C}(1\%, 100\%)$, $Q_{A,B,C}(10\%, 100\%)$ and f_D are very powerful statistics for detecting AI. The number of uniquely shared sites $U_{A,B,C}(x, y, z)$ is also powerful, so long as y is large. Additionally, for different choices of y, using $w = 1\%$ yields a more powerful statistic than using $w = 10\%$.

Joint distributions

We were also interested in the joint distribution of pairs of these statistics. Figure 3.10 shows the joint distribution of $Q95_{A,B,C}(1\%, 100\%)$ in the y-axis and four other statistics

(R_D, Het, D and f_D) in the x-axis, under different admixture and selection regimes. One can observe, for example, that while $Q_{A,B,C}(1\%, 100\%)$ increases with increasing selection intensity and admixture rates, Het increases with increasing admixture rates, but decreases with increasing selection intensity. Thus, under AI the two forces cancel each other out, and we obtain a similar value of Het as under neutrality. Furthermore, the joint distributions of $Q95_{A,B,C}(1\%, 100\%)$ and f_D or R_D show particularly good separation among the different AI scenarios.

Another joint distribution that is especially good at separating different AI regimes is the combination of $Q95_{A,B,C}(w, 100\%)$ and $U_{A,B,C}(w, x, 100\%)$. In Figure 3.11, we show this joint distribution, for different choices of w (1%, 10%) and x (20%, 50%). Here, with increasing intensity of selection and admixture, the number of uniquely shared sites and the quantile statistic increase, but the quantile statistic tends to only reach high values when selection is strong, even if admixture rates are low.

Alternative demographic scenarios

We evaluated the performance of our statistics under various alternative demographic scenarios. First, we simulated a 5X bottleneck occurring in population B 1,600 generations before the admixture event, and lasting 200 generations, to observe its effects on the power of the statistics for detecting AI (Figure 3.3.B). Though we observe a reduction in power - most evident in the heterozygosity statistics - none of the statistics are very strongly affected by this event (Figure 3.12). We also simulated a bottleneck of equal size but occurring after the admixture event - starting 1,400 generations ago, and lasting 200 generations (Figure 3.3.C). In this case, the sensitivity of all the statistics is strongly reduced when the admixture rate is low (Figure 3.13). For example, when looking at the raw values of the $U_{A,B,C}$ and $Q95_{A,B,C}$ statistics, we observe that for low admixture rates the distribution under selection has more overlap with the distribution under neutrality, which explains the low power (Figures 3.14, 3.15). Additionally, $U_{A,B,C}$ seems to display more elevated values under neutrality than in the constant population size model. However, the relative performance of each statistic with respect to all the others does not appear to change much (Figure 3.13).

Then, we set the introgressed haplotype to not be immediately adaptive in the Eurasian population, but to instead undergo an intermediate period of neutral drift, before it becomes advantageous (Figure 3.3.D). In such a situation, our power to detect AI is reduced, for all statistics (Figure 3.16). This is particularly an issue when the admixture rate is low, as in those cases the starting frequency of the selected allele in the Eurasian population is low, so it is more likely to drift to extinction during the neutral period, before it can become advantageous.

We also evaluated the performance of our statistics under selective scenarios that did not involve adaptive introgression, to check which of them were sensitive to these models and which were not. Under a model of selection from de novo mutation (SDN, Figure 3.3.E), in which a single mutation appears in the receiving population after the admixture event, the heterozygosity and linkage disequilbrium statistics ($r^2[intro]$ and $D'[intro]$) are the most sensitive ones (Figure 3.17). This is expected, given that classical selective sweeps are known to strongly affect patterns of heterozygosity and linkage disequilibrium in the neighborhood of the selected site [75, 76, 77]. We also simulated a model of selection from standing variation (Figure 3.3.F), by randomly selecting 20% of haplotypes within the introgressing population to be advantageous, after the introgression event had already occurred. In this case, all statistics perform poorly, especially when admixture is low. Interestingly, when admixture is high (Figure 3.18), $Q95_{A,B,C}(1\%, 100\%)$ and $U_{A,B,C}(1\%, 0\%, 100\%)$ are the best performing statistics. This is likely because some of the haplotypes that are randomly chosen to be selected also happen to be ancestrally polymorphic and present in the archaic humans.

When we set ancestral structure to be our null model, we observe different behaviors depending on the strength of the migration rates. When the migration rates are strong (Figure 3.19), we have excellent power to detect AI with several statistics, including $Q95_{A,B,C}(1\%, 100\%)$, D, f_D, R_D and $U_{A,B,C}(1\%, 50\%, 100\%)$. When the rates are of medium strength (Figure 3.20), the power is slightly reduced, but the same statistics are the ones that perform best. When the migration rates are weak - meaning ancestral structure is very strong - $Q95_{A,B,C}(1\%, 100\%)$ loses power, and the best-performing statistics are R_D, D and f_D (Figure 3.21). We note, though, that the genome-wide D observed under this last ancestral structure model ($D = 0.24$) is much more extreme than the genome-wide D observed empirically between any Eurasian population and Neanderthals or Denisovans, suggesting that if there was ancestral structure between archaic and modern humans, it was likely not of this magnitude.

Global features of uniquely shared archaic alleles

Before identifying candidate genes for adaptive introgression, we investigated the frequency and number of uniquely shared sites at the genome-wide level. Specifically, we wanted to know whether human populations varied in the number of sites with uniquely shared archaic alleles, and whether they also varied in the frequency distribution of these alleles. Therefore, we computed $U_{A,B,Nea,Den}$(1%,x,y,z) and $Q95_{A,B,Nea,Den}$(1%,y,z) for different choices of x, y and z. We used each of the non-African panels in the 1000 Genomes Project phase 3 data [45] as the "test" panel (B), and chose the outgroup panel (A) to be the combination of all African populations (YRI, LWK, GWD, MSL, ESN), excluding admixed African-Americans. When setting $x = 0\%$ (i.e. not imposing a frequency cutoff in the target panel B), South Asians as a target population show the largest number of archaic alleles (Figure 3.22). However, East Asians have a larger number of high-frequency uniquely shared archaic alleles than Europeans and South Asians, for both $x = 20\%$ and $x = 50\%$ (Figure 3.22). Population-specific D-statistics (using YRI as the non-admixed population) also follow this trend (Figure 3.23) and we observe this pattern when looking only at the X chromosome as well (Figure 3.24). These results hold in comparisons with both archaic human genomes, but we observe a stronger signal when looking at Neanderthal-specific shared alleles. We observe a similar pattern when calculating $Q95_{A,B,Nea,Den}(1\%, y, z)$ genome-wide (Figure 3.22) with the exception of Denisova-specific shared alleles. The elevation in $U_{A,B,Nea,Den}$

and $Q95_{A,B,Nea,Den}$ in East Asians may result from higher levels of archaic ancestry in East Asians than in Europeans [18], and agrees with studies indicating that more than one pulse of admixture likely occurred in East Asians [19, 78].

Surprisingly, the Peruvians (PEL) harbor the largest amount of high frequency mutations of archaic origin than any other single population, especially when using Neanderthals as bait (Figures 3.22,3.24). It is unclear whether this signal is due to increased drift or selection in this population. Skoglund et al. [79] argue that drift coupled with an ascertainment scheme that excludes low frequency alleles should artificially increase a signal of archaic ancestry, which could explain this pattern. PEL has a history of low effective population sizes relative to other Non-Africans [45], and our $U_{AFR,PEL,Nea,Den}(w, x, y, z)$ statistic is a form of high-frequency ascertainment (in the sense that we only count mutations that have more than x frequency in PEL). This could also explain why the effect is not seen when x = 0% (Figure 3.22), or when computing D-statistics (Figure 3.23), both of which effectively have no high-frequency ascertainment.

Additionally, we plotted the values of $U_{AFR,X,Nea,Den}(w, 1\%, y, z)$ and $Q95_{AFR,X,Nea,Den}(1\%, y, z)$ jointly for each population X, under different frequency cutoffs w. When $w = 0\%$, there is a generally inversely proportional relationship between the two scores (Figure 3.25), but this becomes a directly proportional relationship when $w = 20\%$ (Figure 3.26) or $w = 50\%$ (Figure 3.27). Here, we also clearly observe that PEL is an extreme region with respect to both the number and frequency of archaic shared derived alleles, and that East Asian and American populations have more high-frequency archaic shared alleles than Europeans.

We checked via simulations if the observed excess of high frequency archaic derived mutations in Americans and especially Peruvians could be caused by genetic drift, as a consequence of the bottleneck that occurred in the ancestors of Native Americans as they crossed Beringia. We observe that if the introgressed population B undergoes a bottleneck (5X for 200 generations, starting 200 generations after the admixture event) this can lead to a larger number of $U_{A,B,C}(w, x, y, z)$ for large values of x (Figure 3.14,3.15,3.28). Indeed, population structure analyses of the 1000 Genomes samples suggest that Peruvians have the largest amount of Native American ancestry [45] and show a lack of recent population growth, which could explain this pattern.

Candidate regions for adaptive introgression

To identify adaptively introgressed regions of the genome, we computed $U_{A,B,C,D}(w, x, y, z)$ and $Q95_{A,B,C,D}(w, y, z)$ in 40kb non-overlapping windows along the genome, using the low-coverage sequencing data from phase 3 of the 1000 Genomes Project [45]. We used this window size because the mean length of introgressed haplotypes found in ref. [16] was 44,078 bp (Supplementary Information 13). Our motivation was to find regions under AI in a particular panel B, using panel A as a non-introgressed out-group (generally Africans, unless otherwise stated). We used the high-coverage Altai Neanderthal genome [16] as bait panel C and the high-coverage Denisova genome [2] as bait panel D. We deployed these statistics in three ways: a) to look for Neanderthal-specific AI, we set $y = 100\%$ and $z = 0\%$;

b) to look for Denisova-specific AI, we set $y = 0\%$ and $z = 100\%$; c) to look for AI matching both of the archaic genomes, we set $y = 100\%$ and $z = 100\%$ (Figure 3.2, Table 3.4). To try to determine the adaptive pressure behind the putative AI event, we obtained all the CCDS-verified genes located inside each window [80].

For guidance as to how high a value of U and $Q95$ we would expect under neutrality, we used the simulations from Figure 3.3 to obtain 95% empirical quantiles of the distribution of these scores under neutrality. Tables 3.2 and 3.3 show the 95% quantiles for these two statistics under various models of adaptive introgression and ancestral structure, for different choices of parameter values (see Methods Section). When examining our candidates for AI below, we focused on windows whose values for $U_{A,B,Nea,Den}(w, y, z)$ and $Q95_{A,B,Nea,Den}(w, x, y, z)$ were both in the 99.9% quantile of their respective genome-wide distributions, and also verified that these values would be statistically significant at the 5% level under a simple model of neutral admixture.

We also calculated D and f_D along the same windows (using Africans as the non-admixed population), and saw good agreement with the new statistics presented here (Table 3.4). Finally, we validated the regions most likely to have been adaptively introgressed by searching for archaic tracts of introgression within them that were at high frequency, using a Hidden Markov Model (see below).

Continental populations

When focusing on adaptive introgression in continental populations, we first looked for uniquely shared archaic alleles specific to Europeans that were absent or almost absent ($< 1\%$ frequency) in Africans and East Asians. Conversely, we also looked for uniquely shared archaic alleles in East Asians, which were absent or almost absent in Africans and Europeans. In this continental survey, we ignored Latin American populations as they have high amounts of European and African ancestry, which could confound our analyses. Figure 3.29 shows the number of sites with uniquely shared alleles for increasing frequency cutoffs in the introgressed population, and for different types of archaic alleles (Neanderthal-specific, Denisova-specific or common to both archaic humans). In other words, we calculated $U_{AFR,EUR,Nea,Den}(1\%, x, y, z)$ and $U_{AFR,EAS,Nea,Den}(1\%, x, y, z)$ for different values of x (0%, 20%, 50% and 80%) and different choices of y and z, depending on which type of archaic alleles we were looking for. We observe that the regions in the extreme of the distributions for $x = 50\%$ corresponded very well to genes that had been previously found to be candidates for adaptive introgression from archaic humans in these populations, using more complex probabilistic methods [21, 20] or gene-centric approaches [36]. These include *BNC2* (involved in skin pigmentation [81, 82]), *POU2F3* (involved in skin keratinocyte differentiation [83, 84]), *HYAL2* (involved in the response to UV radiation on human keratinocytes [85]), *SIPA1L2* (involved in neuronal signaling [86]) and *CHMP1A* (a regulator of cerebellar development [87]). To be more rigorous in our search for adaptive introgression, we looked at the joint distribution of the U statistic and the $Q95$ statistic for the same choices of w, y and z, and then selected the regions that were in the 99.9% quantiles of the distributions of both statistics (Figures 3.30, 3.31, 3.32). We find that the strongest candidates here are *BNC2*, *POU2F3*, *SIPA1L2* and the *HYAL2* region.

We also scanned for regions of the genome where South Asians (SAS) had uniquely shared archaic alleles at high frequency, which were absent or almost absent in Europeans, East Asians and Africans. In this case, we focused on $x = 20\%$ because we found that $x = 50\%$ left us with no candidate regions. Among the candidate regions sharing a large number of high-frequency Neanderthal alleles in South Asians, we find genes *ASTN2*, *SFMBT1*, *MUSTN1* and *MAML2* (Figure 3.33). *ASTN2* is involved in neuronal migration [88] and is associated with schizophrenia [89, 90]. *SFMBT1* is involved in myogenesis [91] and is associated with hydrocephalus [92]. *MUSTN1* plays a role in the regeneration of the muscoskeletal system [93]. Finally, *MAML2* codes for a signaling protein [94, 95], and is associated with cutaneous carcinoma [96] and lacrimal gland cancer [97].

Eurasia

We then looked for AI in all Eurasians (EUA=EUR+SAS+EAS, ignoring American populations) using Africans as the non-admixed population (AFR, ignoring admixed African-Americans). Figure 3.30 shows the extreme outlier regions that are in the 99.9% quantiles for both $U_{EUA,AFR,Nea,Den}(1\%, 20\%, y, z)$ and $Q95_{EUA,AFR,Nea,Den}(1\%, y, z)$, while Figure 3.34 shows the entire distribution. We focused on $x = 20\%$ because we found that $x = 50\%$ left us with almost no candidate regions. In this case, the region with by far the largest number of uniquely shared archaic alleles is the one containing genes *OAS1* and *OAS3*, involved in immunity [98, 99, 100, 101]. This region was previously identified as a candidate for AI from Neanderthals in non-Africans [35]. Another region that we recover and was previously identified as a candidate for AI is the one containing genes *TLR1* and *TLR6* [102, 103]. These genes are also involved in immunity and have been shown to be under positive selection in some non-African populations [104, 105].

Interestingly, we find that a very strong candidate region in Eurasia contains genes *TBX15* and *WARS2*. This region has been associated with a variety of traits, including adipose tissue differentiation [106], body fat distribution [107, 108, 109, 110], hair pigmentation [111], facial morphology [112, 113], ear morphology [114], stature [113] and skeletal development [115, 113]. It was previously identified as being under positive selection in Greenlanders [116], and it shows particularly striking signatures of adaptive introgression, so we devote a separate study to its analysis [117].

Population-specific signals of adaptive introgression

To identify population-specific signals of AI, we looked for archaic alleles at high frequency in a particular non-African panel X, which were also at less than 1% frequency in all other non-African and African panels, excluding X (Table 3.4). This is a very restrictive requirement, and indeed, we only find a few windows in a single panel (PEL) with archaic alleles at more than 20% frequency, at sites where the archaic alleles is at less than 1%

frequency in all other panels. One of the regions with the largest number of uniquely shared Neanderthal sites in PEL contains gene *CHD2*, which codes for a DNA helicase [118] involved in myogenesis (UniProtKB by similarity), and that is associated with epilepsy [119, 120].

Shared signals among populations

In the previous section, we focused on regions where archaic alleles were uniquely at high frequencies in particular populations, but at low frequencies in all other populations. This precludes us from detecting AI regions that are shared across more than one non-African population. To address this, we conditioned on observing the archaic allele at less than 1% frequency in a non-admixed outgroup panel composed of all the African panels (YRI, LWK, GWD, MSL, ESN), excluding African-Americans, and then looked for archaic alleles at high frequency in particular non-African populations. Unlike the previous section, we did not condition on the archaic allele being at low frequency in other non-African populations as well. The whole joint distributions of U and $Q95$ for this choice of parameters for each non-African panel are shown in figs. 3.35 to 3.53, while regions in the 99.9% quantile for both statistics are shown in Figure 3.30.

Here, we recapitulate many of the findings from our Eurasian and continental-specific analyses above, like *TLR1/TLR6*, *BNC2*, *OAS1/OAS3*, *POU2F3*, *LIPA* and *TBX15/WARS2* (Figure 3.30). For example, just as we found that *POU2F3* was an extreme region in the East Asian (EAS) continental panel, we separately find that almost all populations composing that panel (CHB, KHV, CHS, CDX, JPT) have archaic alleles in that region at disproportionately high frequency, relative to their frequency in Africans. Additionally though, we can learn things we would not have detected at the continental level. For example, the Bengali from Bangladesh (BEB) - a South Asian population - also have archaic alleles at very high frequencies in this region.

We detected several genes that appear to show signatures of AI across various populations (Figures 3.30). One of the most extreme examples is a 120 kb region containing the *LARS* gene, with 76 uniquely shared Neanderthal alleles at $< 1\%$ frequency in Africans and $> 50\%$ frequency in Peruvians, which are also at $> 20\%$ frequency in Mexicans. *LARS* codes for a leucin-tRNA synthetase [121], and is associated with liver failure syndrome [122]. Additionally, a region containing gene *ZFHX3* displays an elevated number of uniquely shared Neanderthal sites in PEL, and we also observe this when looking more broadly at East Asians (EAS) and - based on the patterns of inferred introgressed tracts (see below) - in various American (AMR) populations as well. *ZFHX3* is involved in the inhibition of estrogen receptor-mediated transcription [123] and has been associated with prostate cancer [124].

We also find several Neanderthal-specific uniquely shared sites in American panels (PEL, CLM, MXL) in a region previously identified as harboring a risk haplotype for type 2 diabetes (chr17:6880001-6960000) [125]. This is consistent with previous findings suggesting the risk haplotype was introgressed from Neanderthals and is specifically present at high frequencies in Latin Americans [125]. The region contains gene *SLC16A11*, whose expression is known to alter lipid metabolism [125]. We also find that the genes *FAP/IFIH1* have signals consistent with AI, particularly in PEL. This region has been previously associated with type 1 diabetes [126, 127].

Another interesting candidate region contains two genes involved in lipid metabolism: *LIPA* and *CH25H*. We find a 40 kb region with 11 uniquely shared Denisovan alleles that are at low ($< 1\%$) frequency in Africans and at very high ($> 50\%$) frequency in various South and East Asian populations (JPT, KHV, CHB, CHS, CDX and BEB). The Q95 and D statistics in this region are also high across all of these populations, and we also find this region to have extreme values of these statistics in our broader Eurasian scan. The *LIPA* gene codes for a lipase [128] and is associated with cholesterol ester storage disease [129] and Wolman disease [130]. In turn, the *CH25H* gene codes for a membrane hydroxylase involved in the metabolism of cholesterol [131] and associated with Alzheimer's disease [132] and antiviral activity [133].

Finally, we find a region harboring between 3 and 10 uniquely shared Neanderthal alleles (depending on the panel used) in various non-African populations. This region was identified earlier by ref. [20] and contains genes *PPDPF*, *PTK6* and *HELZ2*. *PPDPF* codes for a probable regulator of pancreas development (UniProtKB by similarity). *PTK6* codes for an epithelial signal transducer [134] and *HELZ2* codes for a helicase that works as a transcriptional coactivator for nuclear receptors [135, 136].

The X chromosome

Previous studies have observed lower levels of archaic introgression in the X chromosome relative to the autosomes [20, 21] . Here, we observe a similar trend: compared to the autosomes, the X chromosome contains a smaller number of windows with sites that are uniquely shared with archaic humans (Figure 3.29). For example, for $w = 1\%$ and $x = 20\%$, we observe that, in Europeans, 0.4% of all windows in the autosomes have at least one uniquely shared site with Neanderthals or Denisovans, while only 0.05% of all windows in the X chromosome have at least one uniquely shared site ($P = 4.985 \times 10^{-4}$, chi-squared test assuming independence between windows). The same pattern is observed in East Asians ($P = 1.852 \times 10^{-8}$).

Nevertheless, we do identify some regions in the X chromosome exhibiting high values for both $U_{A,B,C,D}(w, x, y, z)$ and $Q95_{A,B,C,D}(w, y, z)$. For example, a region containing gene DHRSX contains a uniquely shared site where a Neanderthal allele is at $< 1\%$ frequency in Africans, but at $> 50\%$ frequency in a British panel (GBR). Another region contains gene DMD and harbors two uniquely shared sites where two archaic (Denisovan/Neanderthal) alleles are also at low ($< 1\%$) frequency in Africans but at $> 50\%$ frequency in Peruvians. DHRSX codes for an oxidoreductase enzyme [137], while DMD is a well-known gene because mutations in it cause muscular dystrophy [138], and was also previously identified as having signatures of archaic introgression in non-Africans [139].

Consequences of relaxing the outgroup frequency cutoff

When using a more lenient cutoff for the outgroup panel (10% maximum frequency, rather than 1%), we find a few genes that display values of the U statistic that are suggestive of AI, and that have been previously found to be under strong positive selection in particular human populations [140, 141]. The most striking examples are *TYRP1* in EUR (using EAS+AFR as outgroup) and *OCA2* in EAS (using EUR+AFR as outgroup)(Table 3.4). Both of these genes are involved in pigmentation. We caution, however, that the reason why they carry archaic alleles at high frequency may simply be because their respective selective sweeps pushed an allele that was segregating in both archaic and modern humans to high frequency in modern humans, but not necessarily via introgression. In fact, *TYRP1* only stands out as an extreme region for the number of archaic shared alleles in EUR when using the lenient 10% cutoff, but not when using the more stringent 1% cutoff. When looking at these SNPs in more detail, we find that their allele frequency in Africans ($\sim$ 20%) is even higher than in East Asians ($\sim$ 1%), largely reflecting population differentiation across Eurasia due to positive selection [141], rather than adaptive introgression. When exploring the haplotype structure of this gene (see below), we find one haplotype that shows similarities to archaic humans but is at low frequency worldwide, and a second - more frequent - haplotype that is more distinct from archaic humans but present at high frequency in Europeans. We find that the uniquely shared sites obtained using the lenient (<10%) allele frequency outgroup cutoff are tagging both haplotypes together, giving the illusion of a strong signal of AI.

OCA2 has several sites with uniquely shared alleles in EAS (AFR+EUR as outgroup) when using the lenient 10% cutoff, but only a few (2) shared archaic sites when using the < 1% outgroup frequency cutoff. When exploring the haplotype structure of this gene (see below), we fail to find a clear-cut differentiation between putatively introgressed and non-introgressed haplotypes, so the evidence for adaptive introgression in this region is also weak. Overall, this suggests that using a lenient outgroup frequency cutoff may lead to misleading inferences. Nevertheless, the particular haplotype structure of these genes and their relationship to their archaic human counterparts warrant further investigation.

Introgressed haplotypes in candidate loci

We inspected the haplotype patterns of candidate loci. We displayed the haplotypes for selected populations at seven regions with evidence for AI: *POU2F3* (Figure 3.54.A), *BNC2* (Figure 3.54.B), *OAS1* (Figure 3.54.C), *LARS* (Figure 3.54.D), *FAP/IFIH1* (Figure 3.54.E), *LIPA* (Figure 3.54.F) and *SLC16A11* (Figure 3.55.C). We included continental populations that show a large number of uniquely shared archaic alleles, and included YRI as a representative African population. We then ordered the haplotypes by similarity to the closest archaic genome (Altai Neanderthal or Denisova) (Figure 3.54). As can be observed, all these regions tend to show sharp distinctions between the putatively introgressed haplotypes and the non-introgressed ones. This is also evident when looking at the cumulative number of differences of each haplotype to the closest archaic haplotype, where we see a sharp rise in the number of differences, indicating strong differentiation between the two sets of haplotypes. Additionally, the YRI haplotypes tend to predominantly belong to the non-introgressed group, as expected.

Moreover, we used this same approach to look at the haplotype structure of genes that, as mentioned above, showed suggestive but not necessarily strong signals of AI (Figure 3.55). First, we see that *OCA2* does not show a large number of differences between the haplotypes that are closer to the archaic humans (Figure 3.55.A). Second, the *TYRP1* region only contains a small number of haplotypes that are similar to the Neanderthal genome: in the combined YRI+EUR panel, just 6% of haplotypes have less than 73 differences to the Neanderthal genome, and this number is roughly the point of transition between the archaic-like and the non-archaic-like haplotypes (Figure 3.55.B).

Finally, we used a HMM [23] to verify that the strongest candidate regions effectively contained archaic segments of a length that would be consistent with introgression after the divergence between archaic and modern humans. For each region, we used the closest archaic genome (Altai Neanderthal or Denisova) as the putative source of introgression. We then plotted the inferred segments in non-African continental populations for genes with strong evidence for AI. Among these, genes with Neanderthal as the closest source (figs. 3.56 to 3.63) include: *POU2F3* (EAS,SAS), *BNC2* (EUR), *OAS1* (Eurasians), *LARS* (AMR), *FAP/IFIH1* (PEL), *CHD2* (PEL), *TLR1-6* (EAS) and *ZFHX3* (PEL). Genes with Denisova as the closest source (figs. 3.64 and 3.65) include: *LIPA* (EAS, SAS, AMR) and *MUSTN1* (SAS).

Testing for enrichment in genic regions

We aimed to test whether uniquely shared archaic alleles at high frequencies were enriched in genic regions of the genome. SNPs in introgressed blocks will tend to cluster together and have similar allele frequencies, which could cause a spurious enrichment signal. Therefore, we performed two types of LD pruning, which we described in the Methods section.

Regardless of which LD method we used, we find no significant enrichment in genic regions for high-frequency ($> 50\%$) Neanderthal alleles (LD-1 P=352, LD-2 P=0.161) or Denisovan alleles (LD-1 P=0.348, LD-2 P=0.192). Similarly, we find no enrichment for medium-to-high-frequency ($> 20\%$) Neanderthal alleles (LD-1 P=0.553, LD-2 P=0.874) or Denisovan alleles (LD-1 P=0.838, LD-2 P=0.44).

3.4 Discussion

Here, we have tested which statistics are most informative in the detection of AI. We find that one of the most powerful ways to detect AI is to look at both the number and allele frequency of mutations that are uniquely shared between the introgressed and the archaic populations. Such mutations should be abundant and at high-frequencies in the introgressed population if AI occurred. In particular, we identified two novel summaries of the data that capture this pattern quite well: the statistics U and $Q95$. These statistics can recover loci under AI and are easy to compute from genomic data, as they do not require phasing.

We have also studied the general landscape of archaic alleles and their frequencies in present-day human populations. While scanning the present-day human genomes from phase 3 of the 1000 Genomes Project [45] using these and other summary statistics, we were able to recapitulate previous AI findings (like the *TLR* [102, 103] and *OAS* regions [35]) as well as identify new candidate regions (like the *LIPA* gene and the *FAP/IFIH1* region). These mostly include genes involved in lipid metabolism, pigmentation and immunity, as observed in previous studies [20, 21, 142]. Phenotypic changes in these systems may have allowed archaic humans to survive in Eurasia during the Pleistocene, and may have been passed on to present-day human populations during their expansion out of Africa.

When using more lenient definitions of what we consider to be "uniquely shared archaic alleles" we find sites containing these alleles in genes that have been previously found to be under positive selection (like *OCA2* and *TYRP1*) but not necessarily under adaptive introgression. While these do not show as strong signatures of adaptive introgression as genes like *BNC2* and *POU2F3*, their curious haplotype patterns and their relationship to archaic genomes warrants further exploration.

In this study, we have mostly focused on positive selection for archaic alleles. One should remember, though, that a larger proportion of introgressed genetic material was likely maladaptive to modern humans, and therefore selected against. Indeed, two recent studies have shown that negative selection on archaic haplotypes may have reduced the initial proportion of archaic material present in modern humans immediately after the hybridization event(s) [30, 29]. Another caveat is that some regions of the genome display patterns that could be consistent with multiple introgression events, followed by positive selection on one or more distinct archaic haplotypes [102]. In this study, we have simply focused on models with a single pulse of admixture, and have not considered complex scenarios with multiple sources of introgression.

It is also worth noting that positive selection for archaic haplotypes may be due to heterosis, rather than adaptation to particular environments [30]. That is, archaic alleles may not have been intrinsically beneficial, but simply protective against deleterious recessive modern human alleles, and therefore selected after their introduction into the modern human gene pool.

Although many of the statistics we introduced in this study have their draw-backs - notably their dependence on simulations to assess significance - they highlight a characteristic signature left by AI in present-day human genomes. Future avenues of research could involve developing ways to incorporate uniquely shared sites into a robust test of selection that specifically targets regions under AI. For example, one could think about modifying statistics based on local between-population population differentiation, like PBS [32], so that they are only sensitive to allele frequency differences at sites that show signatures of archaic introgression.

Finally, while this study has largely focused on human AI, several other species also show suggestive signatures of AI [143]. Assessing the extent and prevalence of AI and uniquely shared sites in other biological systems could provide new insights into their biology and evolutionary history. This may also serve to better understand how populations of organisms respond to introgression events, and to derive general principles about the interplay between admixture and natural selection.

3.5 Acknowledgments

We are grateful to Montgomery Slatkin, Rasmus Nielsen and Amy Ko for helpful advice and comments. E.H.S. is supported by UC Merced start-up funds. D.M. is supported by a University of Torino PhD Scholarship. F.R. is supported by N.I.H. grant R01HG003229 to Rasmus Nielsen.

3.6 Tables

Table 3.2. 95% quantiles of the $U_{A,B,C}$ statistic in a 40 kb window, under different demographic scenarios and archaic allele frequency cutoffs in the outgroup (A) and target (B) population panels. The demographic scenarios correspond to scenarios A, B, C and G from Figure 3.3.

Max. outgroup freq.	**Min. target freq.**	**Scenario**	**95% quantile under neutrality**
0.01	0.8	Admixture (2%)	0
0.01	0.8	Admixture (10%)	0
0.01	0.8	Admixture (25%)	0
0.01	0.8	Ancestral Structure (strong mig.)	0
0.01	0.8	Ancestral Structure (medium mig.)	1
0.01	0.8	Ancestral Structure (weak mig.)	18
0.01	0.8	Admixture (2%), then bottleneck	0
0.01	0.8	Admixture (10%), then bottleneck	0
0.01	0.8	Admixture (25%), then bottleneck	0.05
0.01	0.8	Bottleneck, then admixture (2%)	0
0.01	0.8	Bottleneck, then admixture (10%)	0
0.01	0.8	Bottleneck, then admixture (25%)	0
0.01	0.5	Admixture (2%)	2
0.01	0.5	Admixture (10%)	2
0.01	0.5	Admixture (25%)	5
0.01	0.5	Ancestral Structure (strong mig.)	0
0.01	0.5	Ancestral Structure (medium mig.)	5
0.01	0.5	Ancestral Structure (weak mig.)	22
0.01	0.5	Admixture (2%), then bottleneck	2
0.01	0.5	Admixture (10%), then bottleneck	2
0.01	0.5	Admixture (25%), then bottleneck	8
0.01	0.5	Bottleneck, then admixture (2%)	2
0.01	0.5	Bottleneck, then admixture (10%)	2
0.01	0.5	Bottleneck, then admixture (25%)	6
0.01	0.2	Admixture (2%)	6
0.01	0.2	Admixture (10%)	13
0.01	0.2	Admixture (25%)	29.05
0.01	0.2	Ancestral Structure (strong mig.)	0
0.01	0.2	Ancestral Structure (medium mig.)	9.05
0.01	0.2	Ancestral Structure (weak mig.)	25

0.01	0.2	Admixture (2%), then bottleneck	6
0.01	0.2	Admixture (10%), then bottleneck	17
0.01	0.2	Admixture (25%), then bottleneck	30
0.01	0.2	Bottleneck, then admixture (2%)	8
0.01	0.2	Bottleneck, then admixture (10%)	13.05
0.01	0.2	Bottleneck, then admixture (25%)	29
0.01	0	Admixture (2%)	24
0.01	0	Admixture (10%)	37
0.01	0	Admixture (25%)	39
0.01	0	Ancestral Structure (strong mig.)	3
0.01	0	Ancestral Structure (medium mig.)	12.05
0.01	0	Ancestral Structure (weak mig.)	27
0.01	0	Admixture (2%), then bottleneck	21
0.01	0	Admixture (10%), then bottleneck	34
0.01	0	Admixture (25%), then bottleneck	38
0.01	0	Bottleneck, then admixture (2%)	28
0.01	0	Bottleneck, then admixture (10%)	34.05
0.01	0	Bottleneck, then admixture (25%)	37.05
0.1	0.8	Admixture (2%)	0
0.1	0.8	Admixture (10%)	2
0.1	0.8	Admixture (25%)	2
0.1	0.8	Ancestral Structure (strong mig.)	0
0.1	0.8	Ancestral Structure (medium mig.)	11
0.1	0.8	Ancestral Structure (weak mig.)	23.05
0.1	0.8	Admixture (2%), then bottleneck	0
0.1	0.8	Admixture (10%), then bottleneck	2
0.1	0.8	Admixture (25%), then bottleneck	2
0.1	0.8	Bottleneck, then admixture (2%)	1
0.1	0.8	Bottleneck, then admixture (10%)	2
0.1	0.8	Bottleneck, then admixture (25%)	2
0.1	0.5	Admixture (2%)	5
0.1	0.5	Admixture (10%)	6
0.1	0.5	Admixture (25%)	12
0.1	0.5	Ancestral Structure (strong mig.)	0
0.1	0.5	Ancestral Structure (medium mig.)	17
0.1	0.5	Ancestral Structure (weak mig.)	29
0.1	0.5	Admixture (2%), then bottleneck	6
0.1	0.5	Admixture (10%), then bottleneck	7
0.1	0.5	Admixture (25%), then bottleneck	12
0.1	0.5	Bottleneck, then admixture (2%)	6
0.1	0.5	Bottleneck, then admixture (10%)	6.05
0.1	0.5	Bottleneck, then admixture (25%)	12
0.1	0.2	Admixture (2%)	12
0.1	0.2	Admixture (10%)	18.05
0.1	0.2	Admixture (25%)	35
0.1	0.2	Ancestral Structure (strong mig.)	4
0.1	0.2	Ancestral Structure (medium mig.)	21
0.1	0.2	Ancestral Structure (weak mig.)	32.05
0.1	0.2	Admixture (2%), then bottleneck	14
0.1	0.2	Admixture (10%), then bottleneck	22
0.1	0.2	Admixture (25%), then bottleneck	37
0.1	0.2	Bottleneck, then admixture (2%)	14
0.1	0.2	Bottleneck, then admixture (10%)	20
0.1	0.2	Bottleneck, then admixture (25%)	37
0.1	0	Admixture (2%)	29
0.1	0	Admixture (10%)	44
0.1	0	Admixture (25%)	45
0.1	0	Ancestral Structure (strong mig.)	11
0.1	0	Ancestral Structure (medium mig.)	25
0.1	0	Ancestral Structure (weak mig.)	34

0.1	0	Admixture (2%), then bottleneck	28
0.1	0	Admixture (10%), then bottleneck	40
0.1	0	Admixture (25%), then bottleneck	44
0.1	0	Bottleneck, then admixture (2%)	35
0.1	0	Bottleneck, then admixture (10%)	41
0.1	0	Bottleneck, then admixture (25%)	45

Table 3.1. Summary statistics mentioned in the main text.

Statistic	Explanation	Reference
D	D-statistic: measures excess allele sharing between a test population and an outgroup using a sister population that is more closely related to the test than the ougroup	[8][68]
f_D	Similar to the D-statistic, but serves to better control for local variation in diversity patterns if one is interested in finding loci with excess ancestry from an admixing population.	[71]
R_D	Expected ratio of the divergence between an individual from the source population and an individual from the admixed population, and the divergence between an individual from the source population and an individual from the non-admixed individual. This is computed by taking the average over all pairs of admixed and non-admixed individuals.	This study
$U_{A,B,C}(w,x,y)$	Number of sites in which any allele is at a frequency lower than w in panel B, higher than x in panel B and equal to y in panel C.	This study
$U_{A,B,C,D}(w,x,y,z)$	Number of sites in which any allele is at a frequency lower than w in panel A, higher than x in panel B, equal to$y$ in panel C and equal to z in panel D.	This study
$Q95_{A,B,C}(w,y)$	95% quantile of the distribution of derived allele frequencies in panel B, for sites where the derived allele is at a frequency lower than w in panel A and equal to y in panel C.	This study
$Q95_{A,B,C,D}(w,y,z)$	95% quantile of the distribution of derived allele frequencies in panel B, for sites where the derived allele is at a frequency lower than w in panel A, equal to y in panel C and equal to z in panel D.	This study
Het	Expected heterozygosity, measured as the average of $2p(1-p)$ over all sites in a window, where p is the frequency of an arbitrarily chosen allele.	[144]
$D'[intro]$	A measure of linkage disequilibrium. Computed as D/D_{max} where $D = p_{XY} - p_X p_Y$, p_{XY} is the frequency of haplotype XY, p_X is the frequency of allele X, p_Y is the frequency of allele Y, and D_{max} is the maximum theoretical value that D can take. $D'[intro]$ is computed only using frequencies from the introgressed panel.	[145]
$D'[comb]$	D' computed using haplotype and allele frequencies from the combination of the introgressed and non-introgressed panels.	[145]
$r^2[intro]$	A measure of linkage disequilibrium. Computed as $D^2/(p_X(1-p_X)p_Y(1-p_Y))$. $r^2[intro]$ is computed only using frequencies from the introgressed panel.	[146]
$r^2[comb]$	r^2 computed using haplotype and allele frequencies from the combination of the introgressed and non-introgressed panels.	[146]

Table 3.3. 95% quantiles of the $Q95_{A,B,C}$ statistic in a 40 kb window, under different demographic scenarios and archaic allele frequency cutoffs in the outgroup (A) population panel. The demographic scenarios correspond to scenarios A, B, C and G from Figure 3.3.

Max. outgroup freq.	**Scenario**	**95% quantile under neutrality**
0.01	Admixture (2%)	0.28
0.01	Admixture (10%)	0.37
0.01	Admixture (25%)	0.54
0.01	Ancestral Structure (strong mig.)	0.04
0.01	Ancestral Structure (medium mig.)	0.67
0.01	Ancestral Structure (weak mig.)	1
0.01	Admixture (2%), then bottleneck	0.31
0.01	Admixture (10%), then bottleneck	0.44
0.01	Admixture (25%), then bottleneck	0.6
0.01	Bottleneck, then admixture (2%)	0.28
0.01	Bottleneck, then admixture (10%)	0.42
0.01	Bottleneck, then admixture (25%)	0.55
0.1	Admixture (2%)	0.47
0.1	Admixture (10%)	0.51
0.1	Admixture (25%)	0.63
0.1	Ancestral Structure (strong mig.)	0.25
0.1	Ancestral Structure (medium mig.)	0.91
0.1	Ancestral Structure (weak mig.)	1
0.1	Admixture (2%), then bottleneck	0.53
0.1	Admixture (10%), then bottleneck	0.58
0.1	Admixture (25%), then bottleneck	0.67
0.1	Bottleneck, then admixture (2%)	0.47
0.1	Bottleneck, then admixture (10%)	0.53
0.1	Bottleneck, then admixture (25%)	0.66

Table 3.4. 40 kb windows that lie in the highest 99.9% quantile of both $U_{A,B,Nea,Den}$ and $Q95_{A,B,Nea,Den}$ for various outgroup panels A and target panels B, using an outgroup maximum frequency cutoff of 1%, and using different target allele frequency cutoffs (20%, 50%). For each region, we also show other statistics indicative of AI for reference. We partitioned the 1000 Genomes panels into outgroup panel A and target panel B in different ways (column "Mode"), depending on the signals we were looking for. These modes of partitioning are as follows. "Populations" = outgroup panel was the combination of all the populations that were not the target panel. "PopulationsB" = outgroup panel was the combination of all African panels (excluding admixed African-Americans), while target panel was one of the non-African panels. "Continents" = target panel was either the EUR continental panel (in which case the outgroup was AFR+EAS) or the EAS continental panel (in which case the outgroup was AFR+EUR). "ContinentsB" = target panel was the EUR continental panel (in which case the outgroup was AFR+EAS+SAS) or the EAS continental panel (in which case the outgroup was AFR+EUR+SAS) or the SAS continental panel (in which case the outgroup was AFR+EUR+EAS). "Eurasia" = target panel was EUR+EAS, while outgroup panel was AFR.

`https://www.dropbox.com/s/p9k94i2c50rincq/Extreme_gene_table.xlsx?dl=0`

3.7 Figures

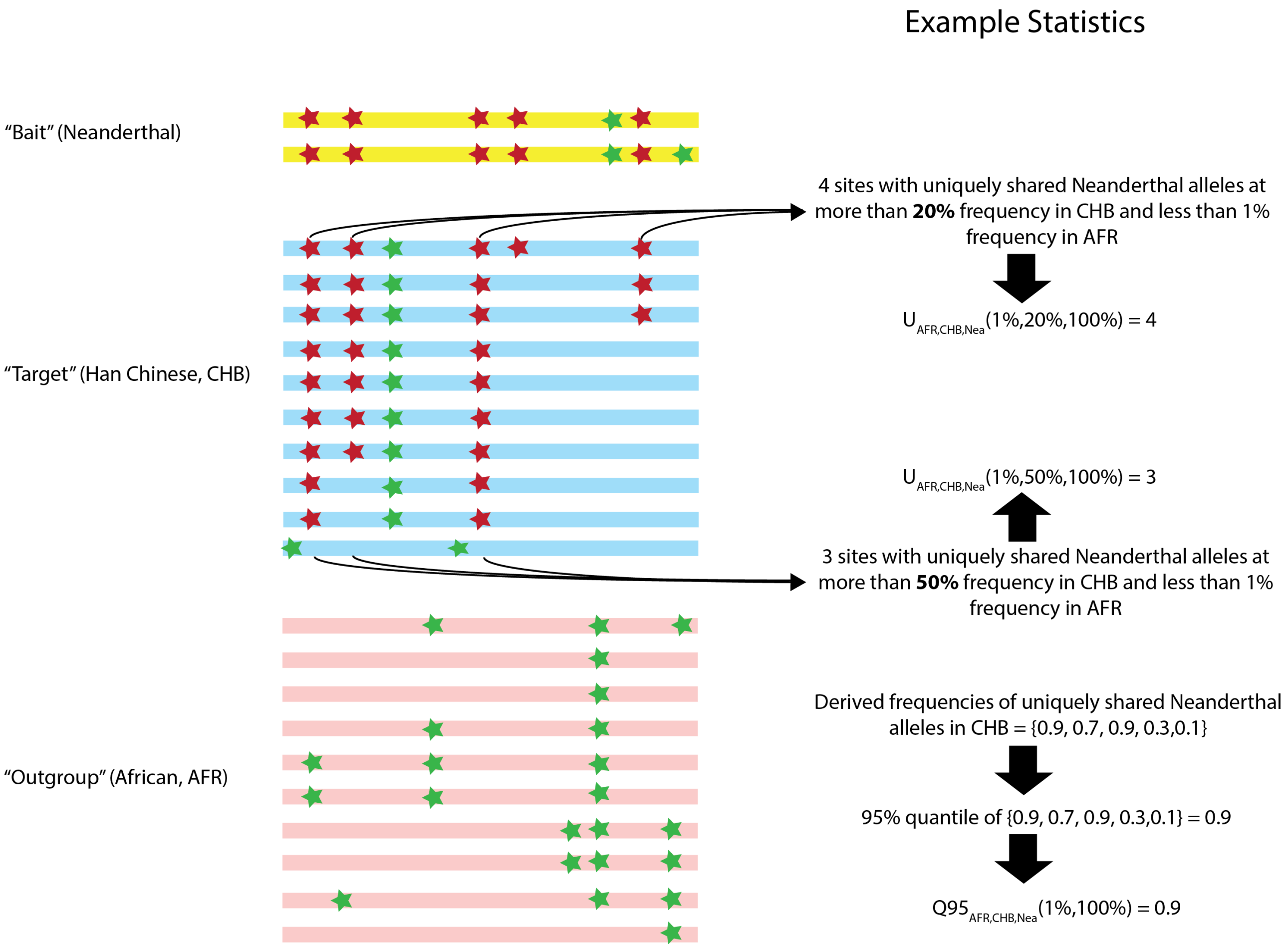

Example Statistics
"Bait" (Neanderthal)
"Target" (Han Chinese, CHB)
"Outgroup" (African, AFR)
4 sites with uniquely shared Neanderthal alleles at more than **20%** frequency in CHB and less than 1% frequency in AFR
$U_{AFR,CHB,Nea}(1\%,20\%,100\%) = 4$
$U_{AFR,CHB,Nea}(1\%,50\%,100\%) = 3$
3 sites with uniquely shared Neanderthal alleles at more than **50%** frequency in CHB and less than 1% frequency in AFR
Derived frequencies of uniquely shared Neanderthal alleles in CHB = {0.9, 0.7, 0.9, 0.3,0.1}
95% quantile of {0.9, 0.7, 0.9, 0.3,0.1} = 0.9
$Q95_{AFR,CHB,Nea}(1\%,100\%) = 0.9$

Figure 3.1. Schematic illustration of the way the $U_{A,B,C}$ and $Q95_{A,B,C}$ statistics are calculated.

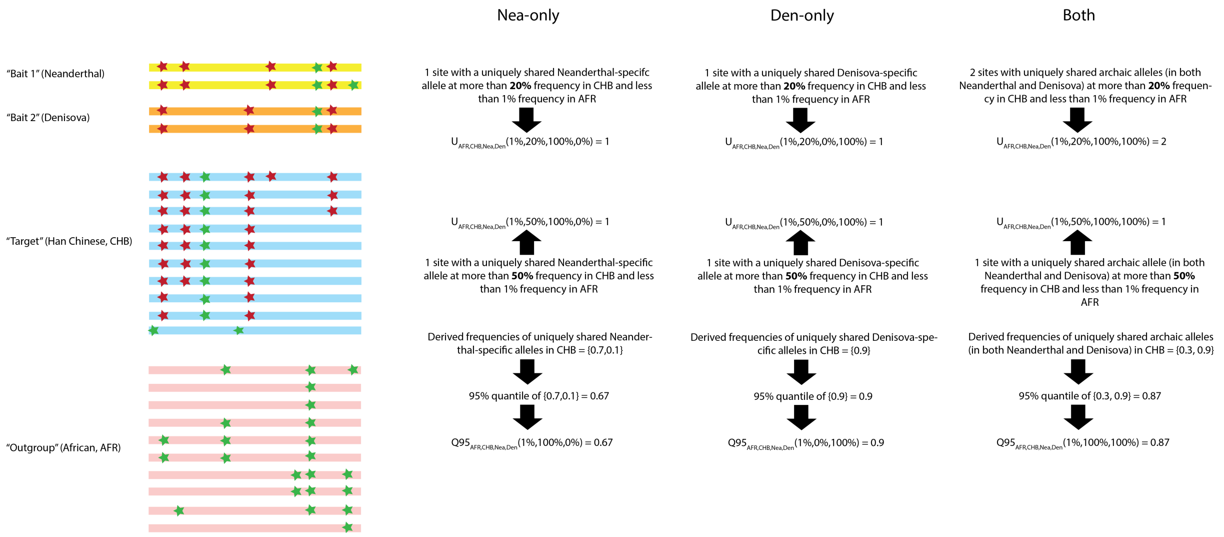

Nea-only
Den-only
Both
"Bait 1" (Neanderthal)
"Bait 2" (Denisova)
"Target" (Han Chinese, CHB)
"Outgroup" (African, AFR)
1 site with a uniquely shared Neanderthal-specifc allele at more than **20%** frequency in CHB and less than 1% frequency in AFR
1 site with a uniquely shared Denisova-specific allele at more than **20%** frequency in CHB and less than 1% frequency in AFR
2 sites with uniquely shared archaic alleles (in both Neanderthal and Denisova) at more than **20%** frequency in CHB and less than 1% frequency in AFR
$U_{AFR,CHB,Nea,Den}(1\%,20\%,100\%,0\%) = 1$
$U_{AFR,CHB,Nea,Den}(1\%,20\%,0\%,100\%) = 1$
$U_{AFR,CHB,Nea,Den}(1\%,20\%,100\%,100\%) = 2$
$U_{AFR,CHB,Nea,Den}(1\%,50\%,100\%,0\%) = 1$
$U_{AFR,CHB,Nea,Den}(1\%,50\%,0\%,100\%) = 1$
$U_{AFR,CHB,Nea,Den}(1\%,50\%,100\%,100\%) = 1$
1 site with a uniquely shared Neanderthal-specific allele at more than **50%** frequency in CHB and less than 1% frequency in AFR
1 site with a uniquely shared Denisova-specific allele at more than **50%** frequency in CHB and less than 1% frequency in AFR
1 site with a uniquely shared archaic allele (in both Neanderthal and Denisova) at more than **50%** frequency in CHB and less than 1% frequency in AFR
Derived frequencies of uniquely shared Neanderthal-specific alleles in CHB = {0.7,0.1}
Derived frequencies of uniquely shared Denisova-specific alleles in CHB = {0.9}
Derived frequencies of uniquely shared archaic alleles (in both Neanderthal and Denisova) in CHB = {0.3, 0.9}
95% quantile of {0.7,0.1} = 0.67
95% quantile of {0.9} = 0.9
95% quantile of {0.3, 0.9} = 0.87
$Q95_{AFR,CHB,Nea,Den}(1\%,100\%,0\%) = 0.67$
$Q95_{AFR,CHB,Nea,Den}(1\%,0\%,100\%) = 0.9$
$Q95_{AFR,CHB,Nea,Den}(1\%,100\%,100\%) = 0.87$

Figure 3.2. Schematic illustration of the way the $U_{A,B,C,D}$ and $Q95_{A,B,C,D}$ statistics are calculated.

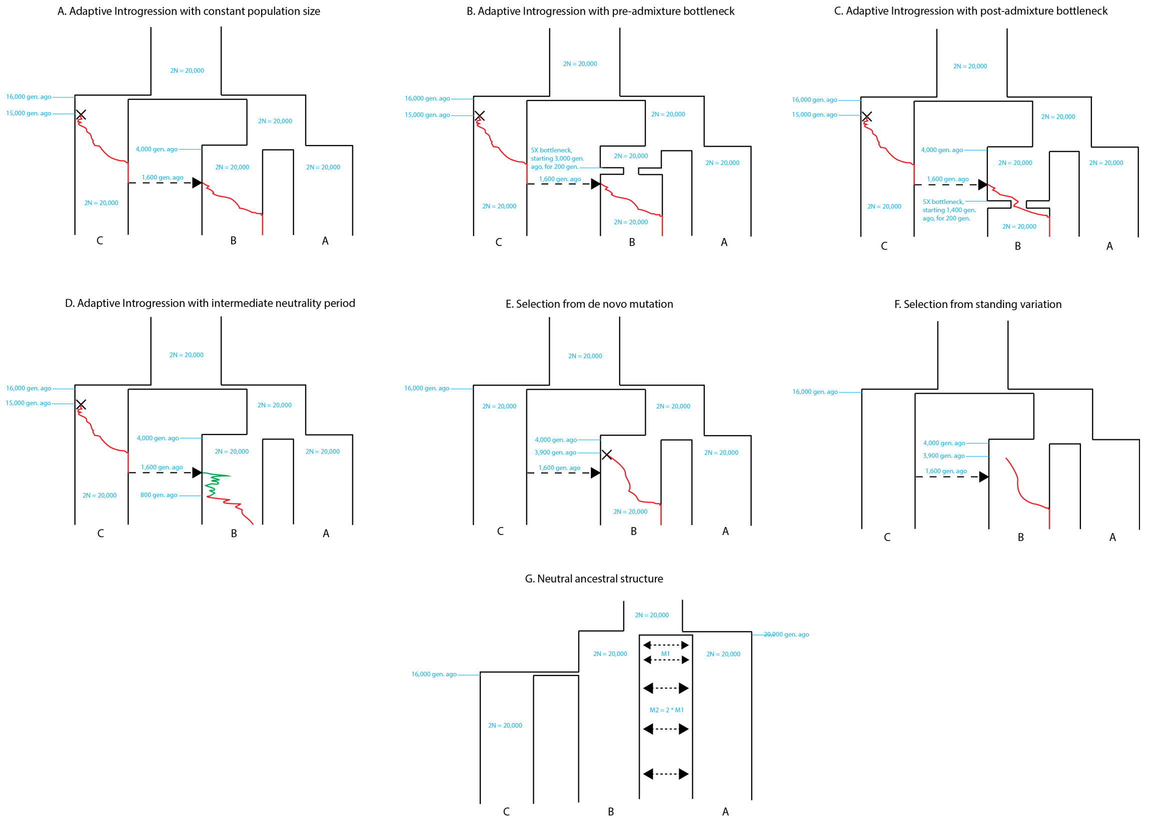

A. Adaptive Introgression with constant population size
B. Adaptive Introgression with pre-admixture bottleneck
C. Adaptive Introgression with post-admixture bottleneck
D. Adaptive Introgression with intermediate neutrality period
E. Selection from de novo mutation
F. Selection from standing variation
G. Neutral ancestral structure
2N = 20,000
16,000 gen. ago
15,000 gen. ago
20,000 gen. ago
4,000 gen. ago
3,900 gen. ago
1,600 gen. ago
800 gen. ago
5X bottleneck, starting 3,000 gen. ago, for 200 gen.
5X bottleneck, starting 1,400 gen. ago, for 200 gen.
M1
M2 = 2 * M1
C
B
A

Figure 3.3. Demographic models described in the main text.

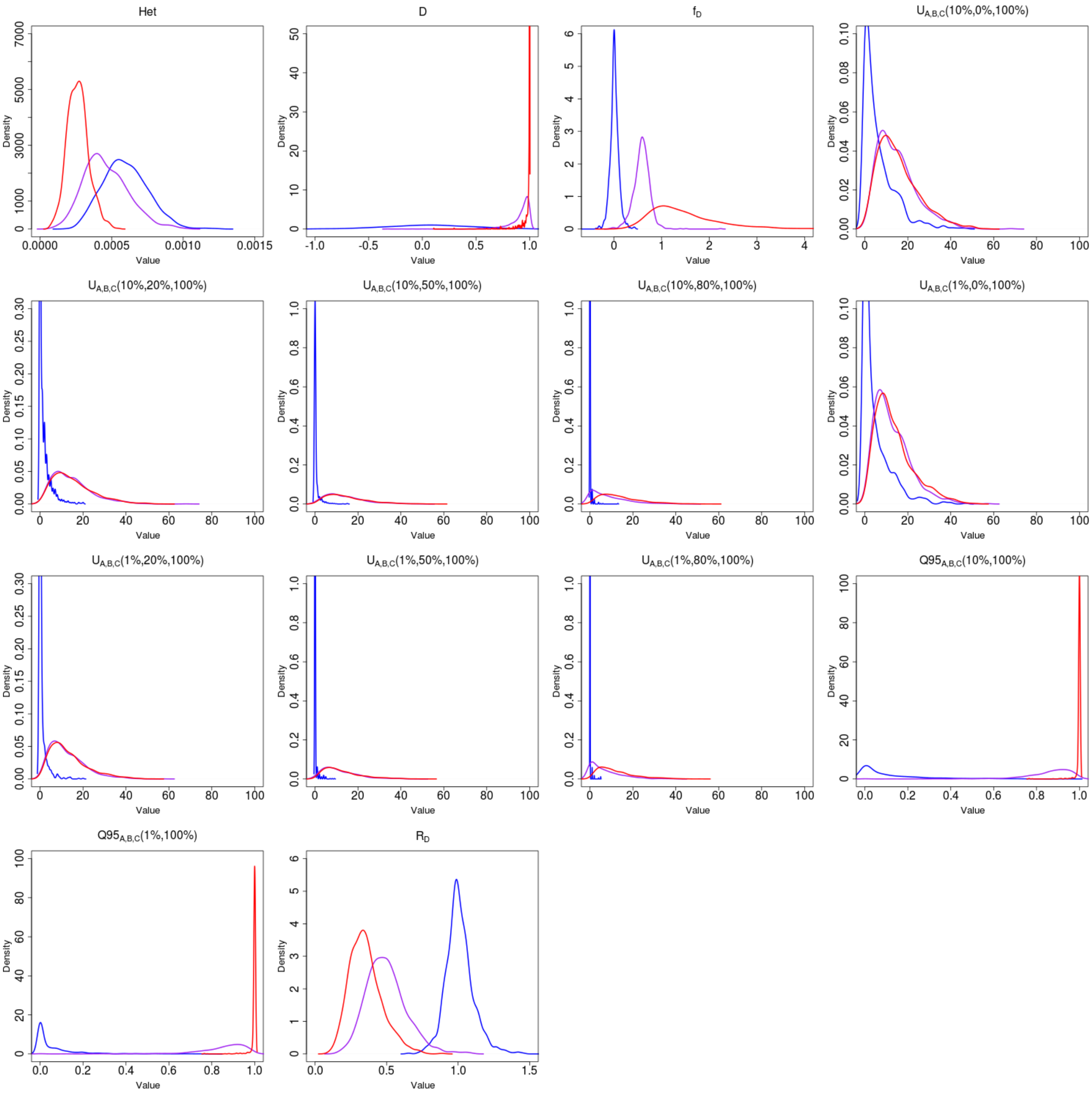


Figure 3.4. Density of various statistics meant to detect genetic patterns left by adaptive introgression, for three scenarios: neutrality (s=0) in blue, weak adaptive introgression (s=0.01) in purple and strong adaptive introgression (s=0.1) in red. The demography was the same as in Figure 3.8 and the admixture rate was set at 2%. See Table 3.1 for a definition of the statistics shown.

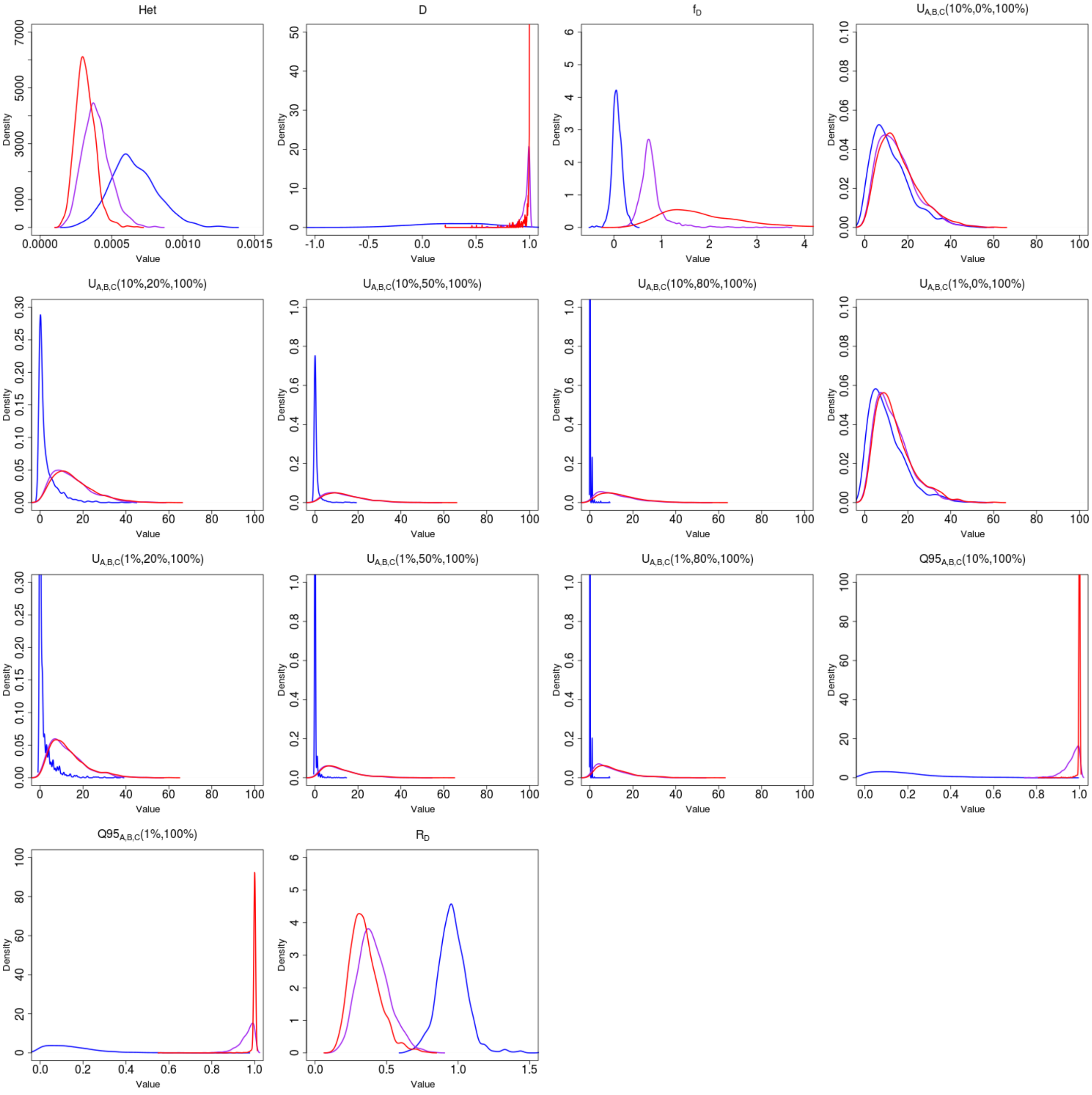


Figure 3.5. Density of various statistics meant to detect genetic patterns left by adaptive introgression, for three scenarios: neutrality (s=0) in blue, weak adaptive introgression (s=0.01) in purple and strong adaptive introgression (s=0.1) in red. The demography was the same as in Figure 3.8 and the admixture rate was set at 10%. See Table 3.1 for a definition of the statistics shown.

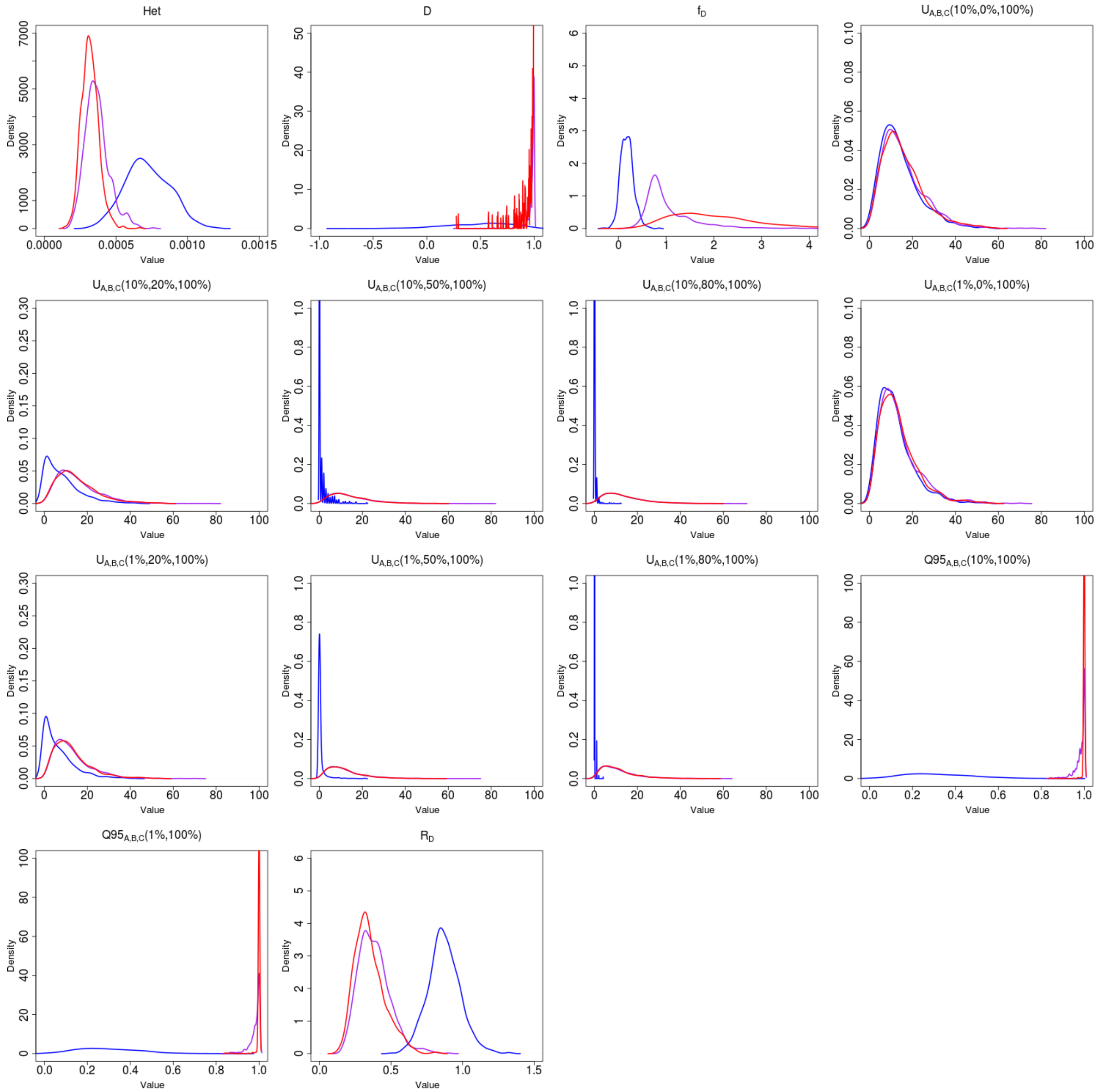


Figure 3.6. Density of various statistics meant to detect genetic patterns left by adaptive introgression, for three scenarios: neutrality (s=0) in blue, weak adaptive introgression (s=0.01) in purple and strong adaptive introgression (s=0.1) in red. The demography was the same as in Figure 3.8 and the admixture rate was set at 25%. See Table 3.1 for a definition of the statistics shown.

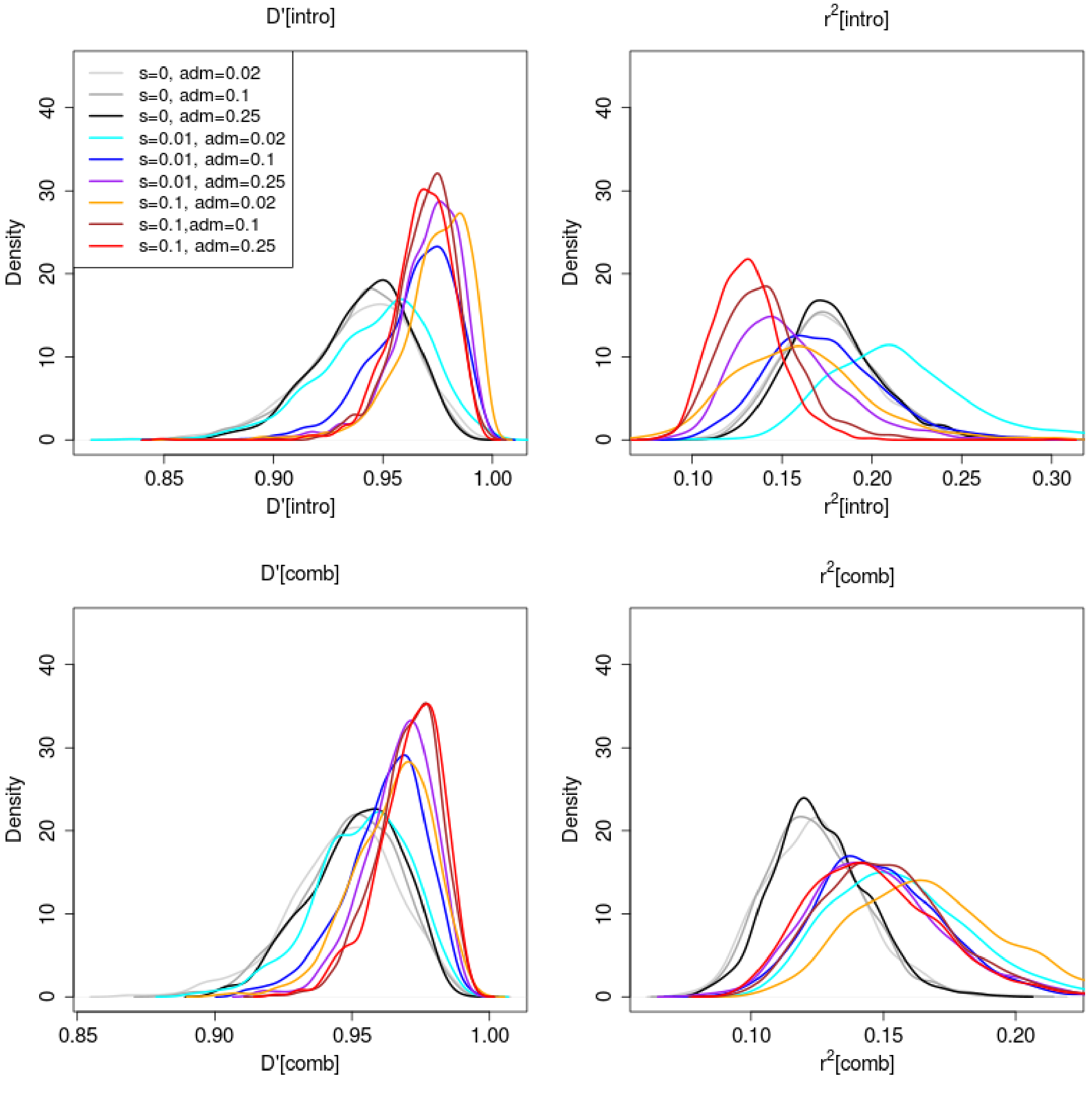


Figure 3.7. Density of statistics that detect patterns of linkage disequilibrium for various neutral and adaptive introgression scenarios. See Table 3.1 for a definition of the statistics shown.

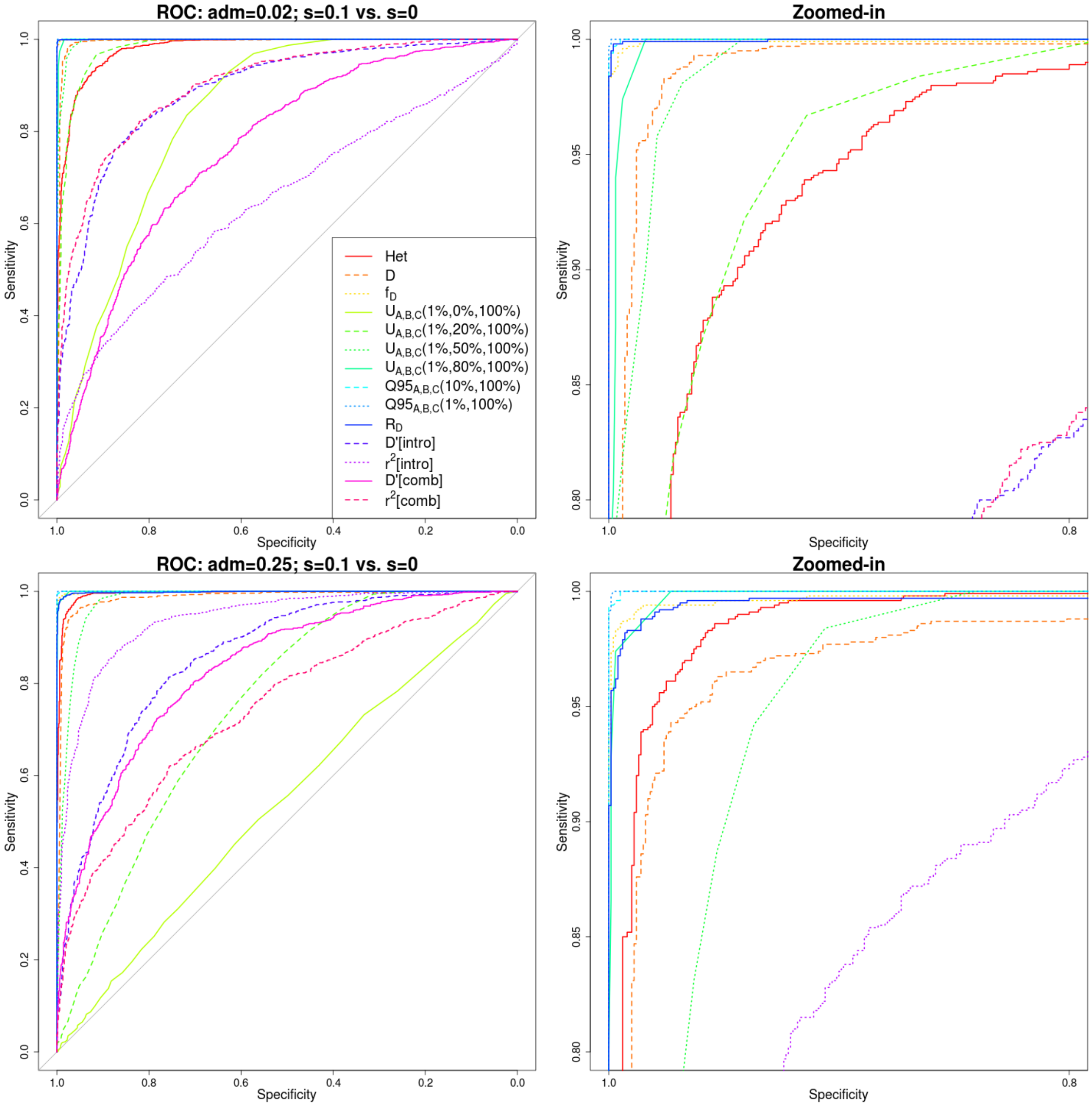


Figure 3.8. Receiver operating characteristic curves for a scenario of adaptive introgression (s=0.1) compared to a scenario of neutrality (s=0), using 1,000 simulations for each case. Populations A and B split from each other 4,000 generations ago, and their ancestral population split from population C 16,000 generations ago. Population sizes were constant and set at $2N = 20,000$. The admixture event occurred 1,600 generations ago from population C to population B, at rate 2% (top panels) or 25% (bottom panels). The right panels are zoomed-in versions of the left panels.

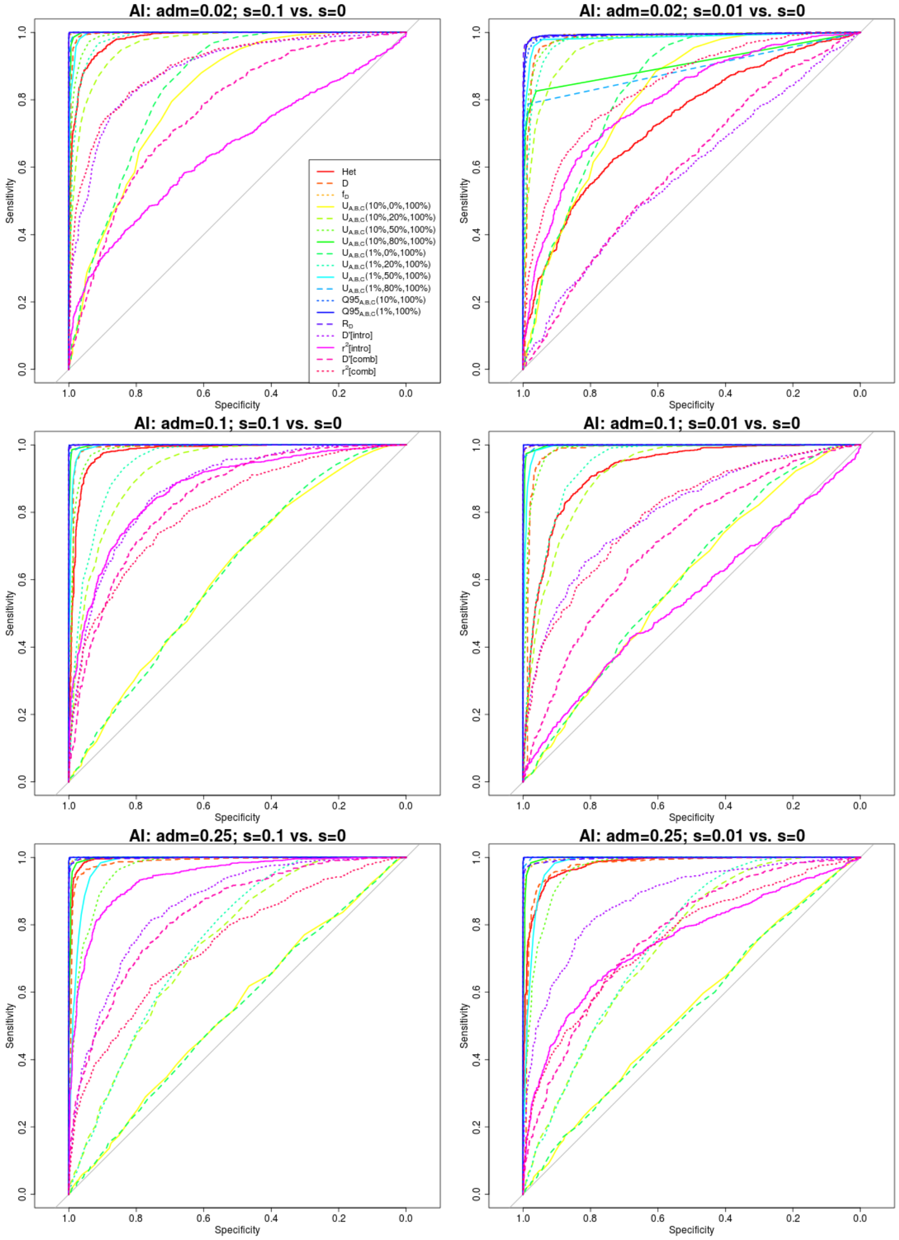


Figure 3.9. Receiver operating characteristic curves for adaptive introgression with constant population size, using 1,000 simulations of adaptive introgression, under various selection (s=0.1, s=0.01) and admixture rate (2%, 10%, 25%) regimes. Populations A and B split from each other 4,000 generations ago, and their ancestral population split from population C 16,000 generations ago. Population sizes were set at $2N = 20,000$. The admixture event occurred 1,600 generations ago from population C into population B,

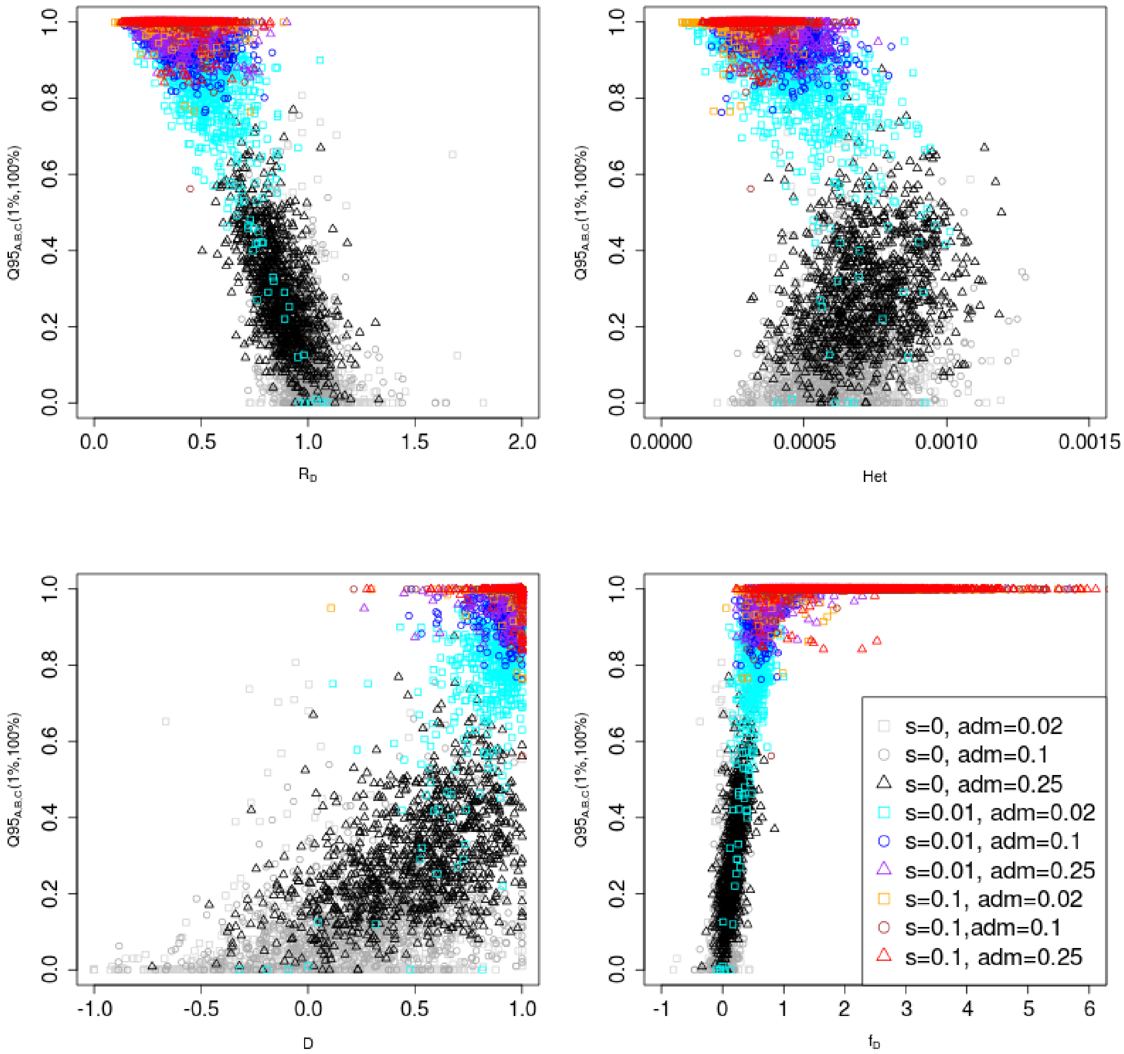


Figure 3.10. Joint distribution of $Q95_{A,B,C}$(1%,100%) and other statistics (R_D, Het, D and f_D). 100 individuals were sampled from panel A, 100 from panel B and 2 from panel C. The demographic parameters were the same as in Figure 3.8.

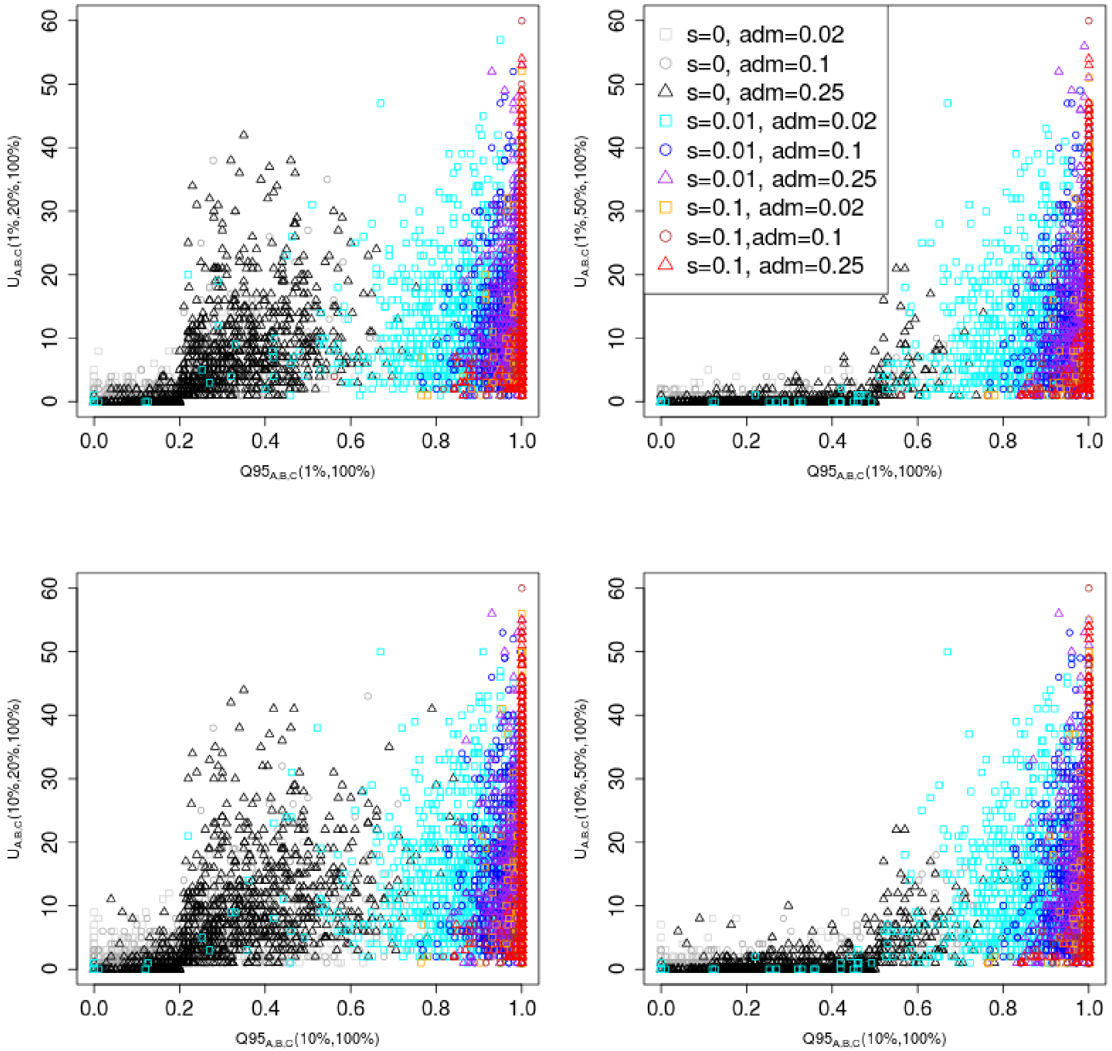


Figure 3.11. Joint distribution of $Q95_{A,B,C}(w, y)$ and $U_{A,B,C}(w, x, y)$ for different choices of w (1%, 10%) and x (20%, 50%). We set y to 100% in all cases. 100 individuals were sampled from panel A, 100 from panel B and 2 from panel C. The demographic parameters were the same as in Figure 3.8.

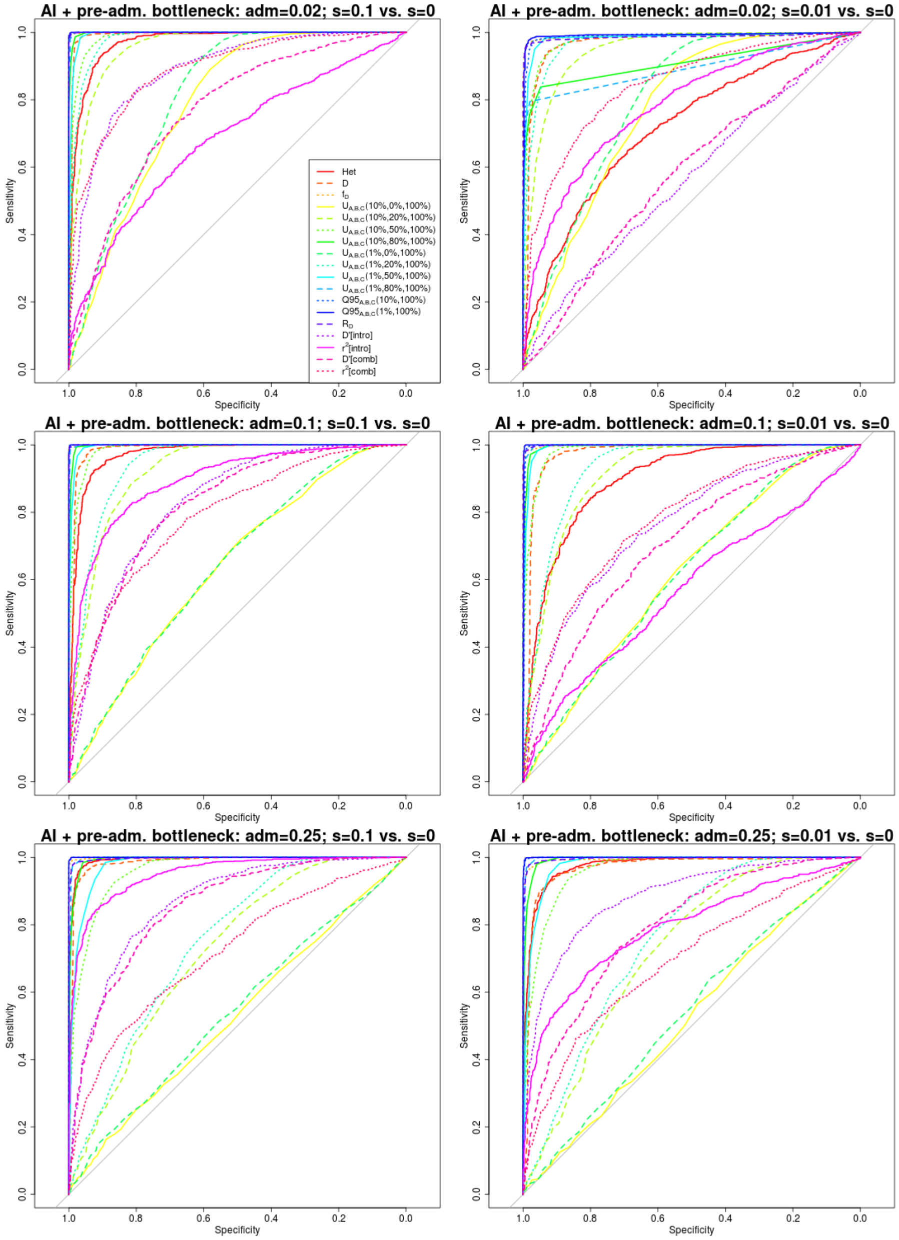


Figure 3.12. Receiver operating characteristic curves for adaptive introgression with a pre-admixture bottleneck, using 1,000 simulations under adaptive introgression . We simulated the same demography as in Figure 3.8, but also included a 5X bottleneck in population B after the introgression event, starting 3,000 generations ago and finishing 2,800 generations ago.

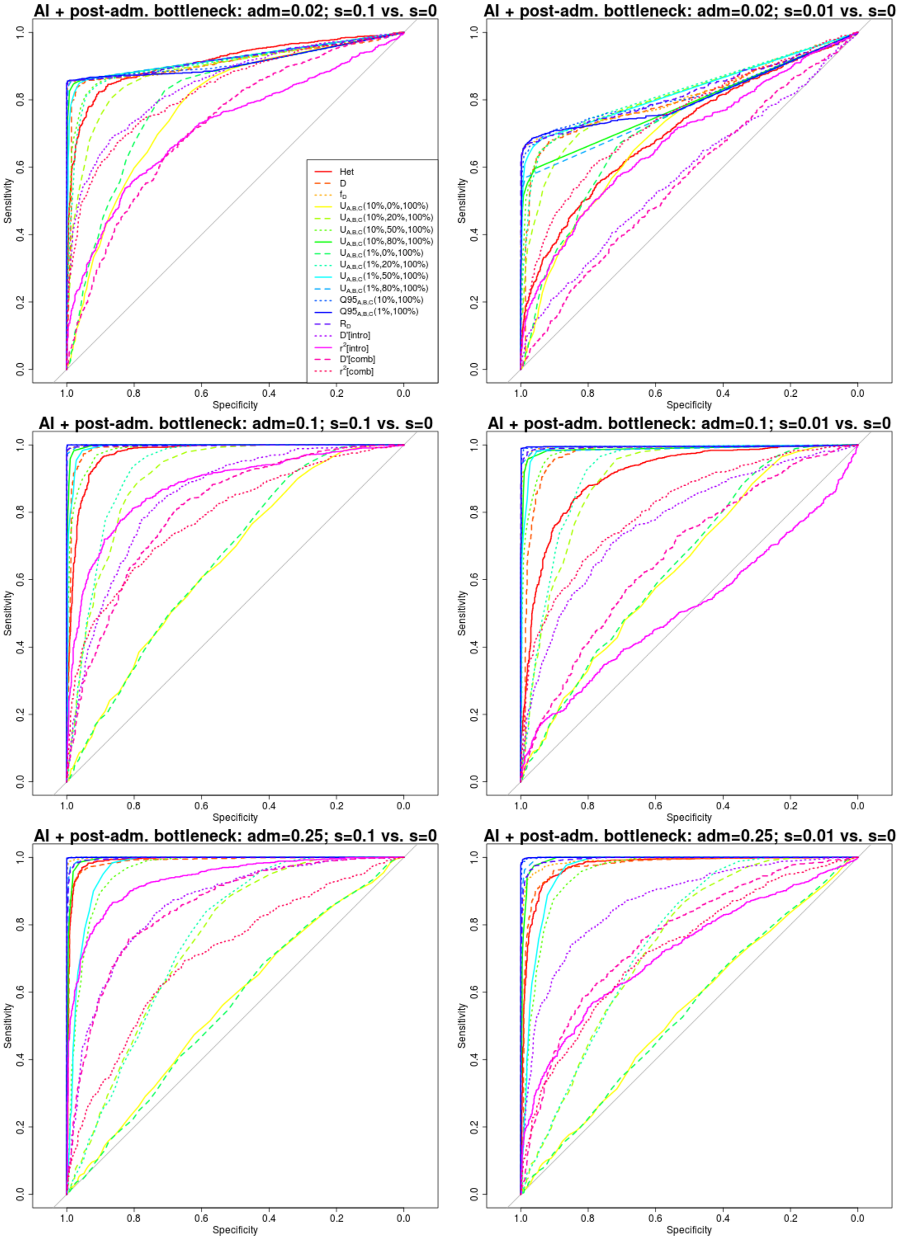


Figure 3.13. Receiver operating characteristic curves for adaptive introgression with a post-admixture bottleneck, using 1,000 simulations under adaptive introgression . We simulated the same demography as in Figure 3.8, but also included a 5X bottleneck in population B after the introgression event, starting 1,400 generations ago and finishing 1,200 generations ago.

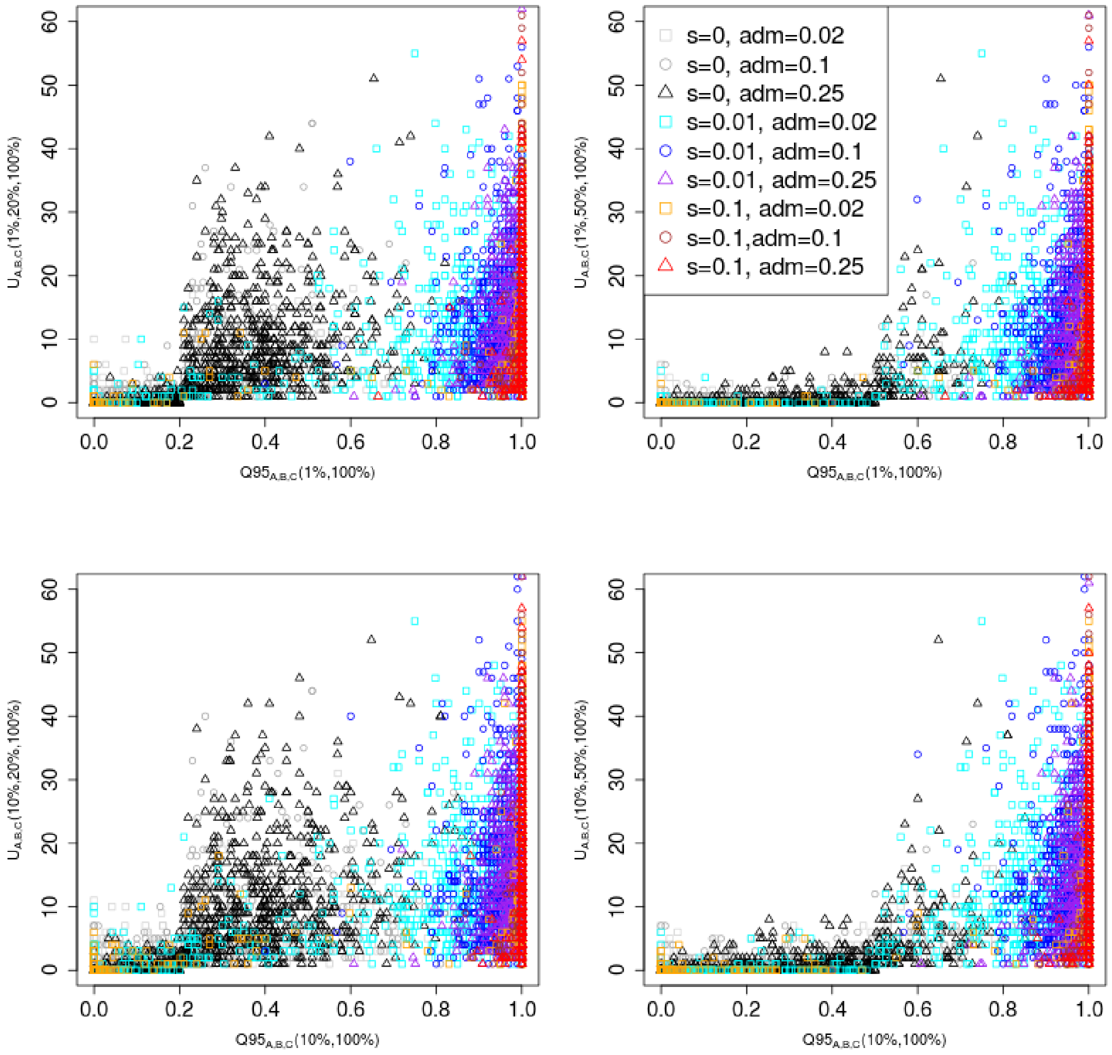


Figure 3.14. Joint distribution of $Q95_{A,B,C}(w, y)$ and $U_{A,B,C}(w, x, y)$ for different choices of w (1%, 10%) and x (20%, 50%). We set y to 100% in all cases. 100 individuals were sampled from panel A, 100 from panel B and 2 from panel C. In this case, we included a 5X bottleneck in population B after the introgression event, starting 1,400 generations ago and finishing 1,200 generations ago.

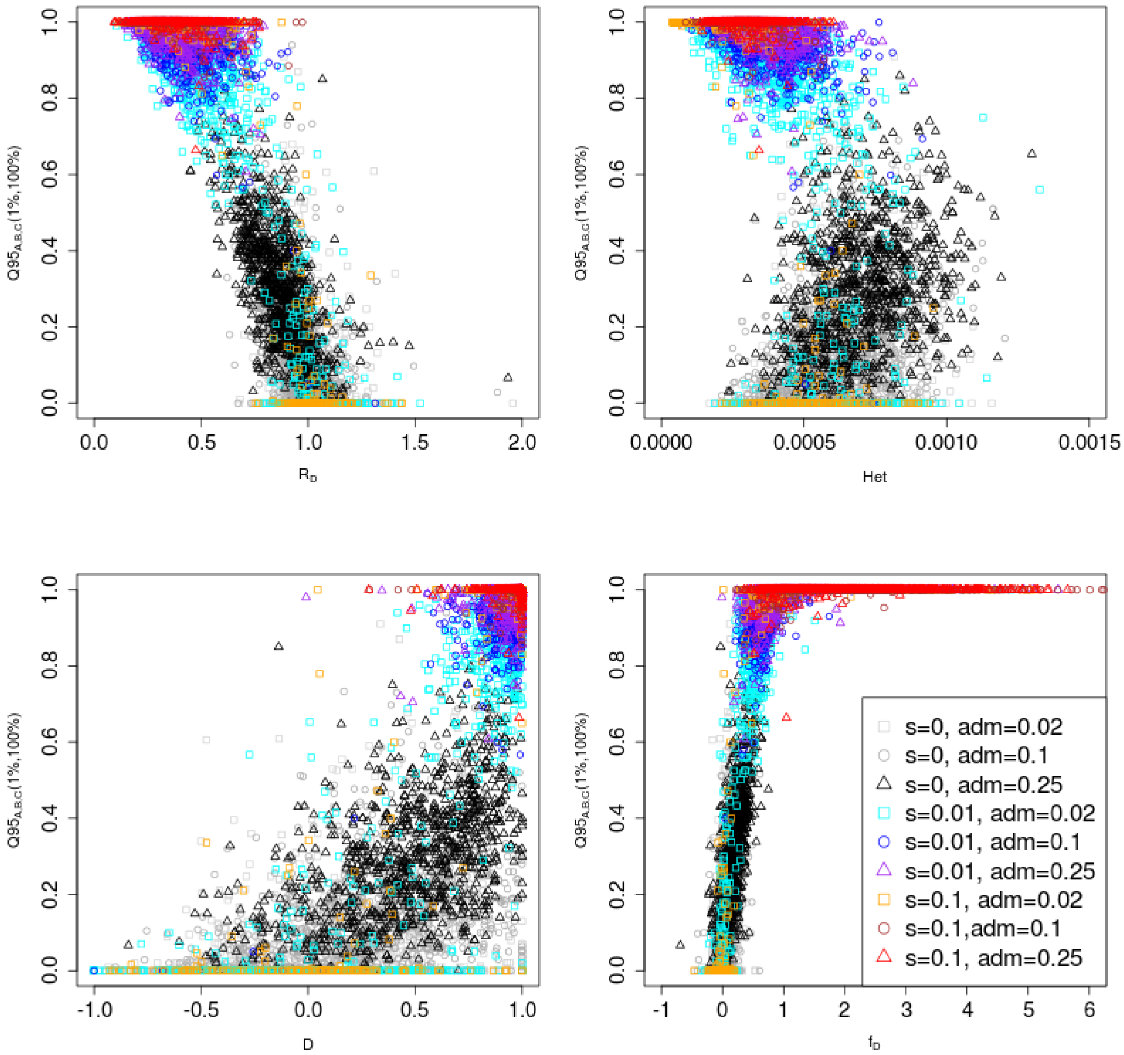


Figure 3.15. Joint distribution of $Q95_{A,B,C}$(1%,100%) and other statistics (R_D, Het, D and f_D). 100 individuals were sampled from panel A, 100 from panel B and 2 from panel C. In this case, we included a 5X bottleneck in population B after the introgression event, starting 1,400 generations ago and finishing 1,200 generations ago.

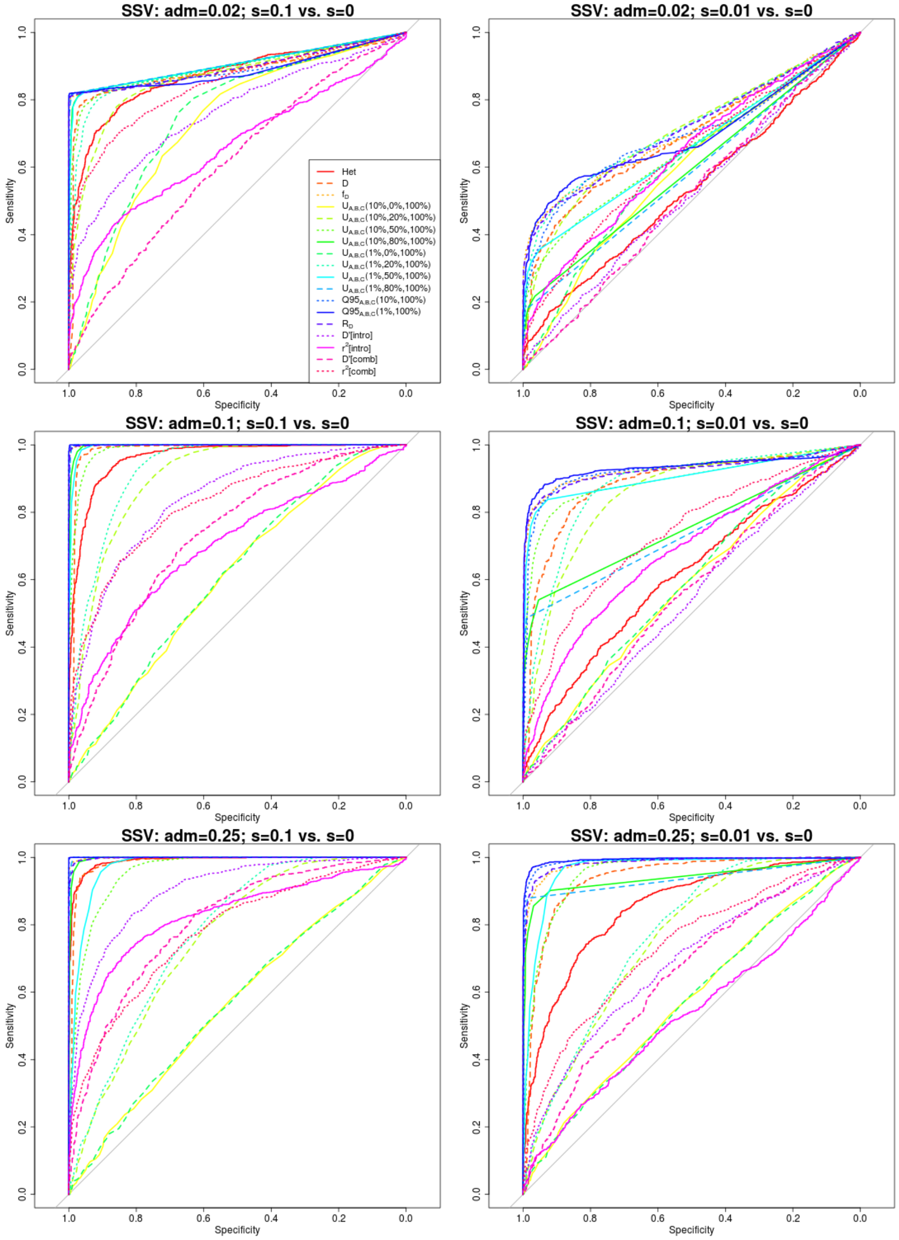


Figure 3.16. Receiver operating characteristic curves for adaptive introgression with an intermediate neutrality period. We simulated the same demography as in Figure 3.8, but changed the selection coefficient of the beneficial variant to be 0 right after the introgression event (1,600 generations ago). If still present in population B, the variant regained its original coefficient 800 generations ago.

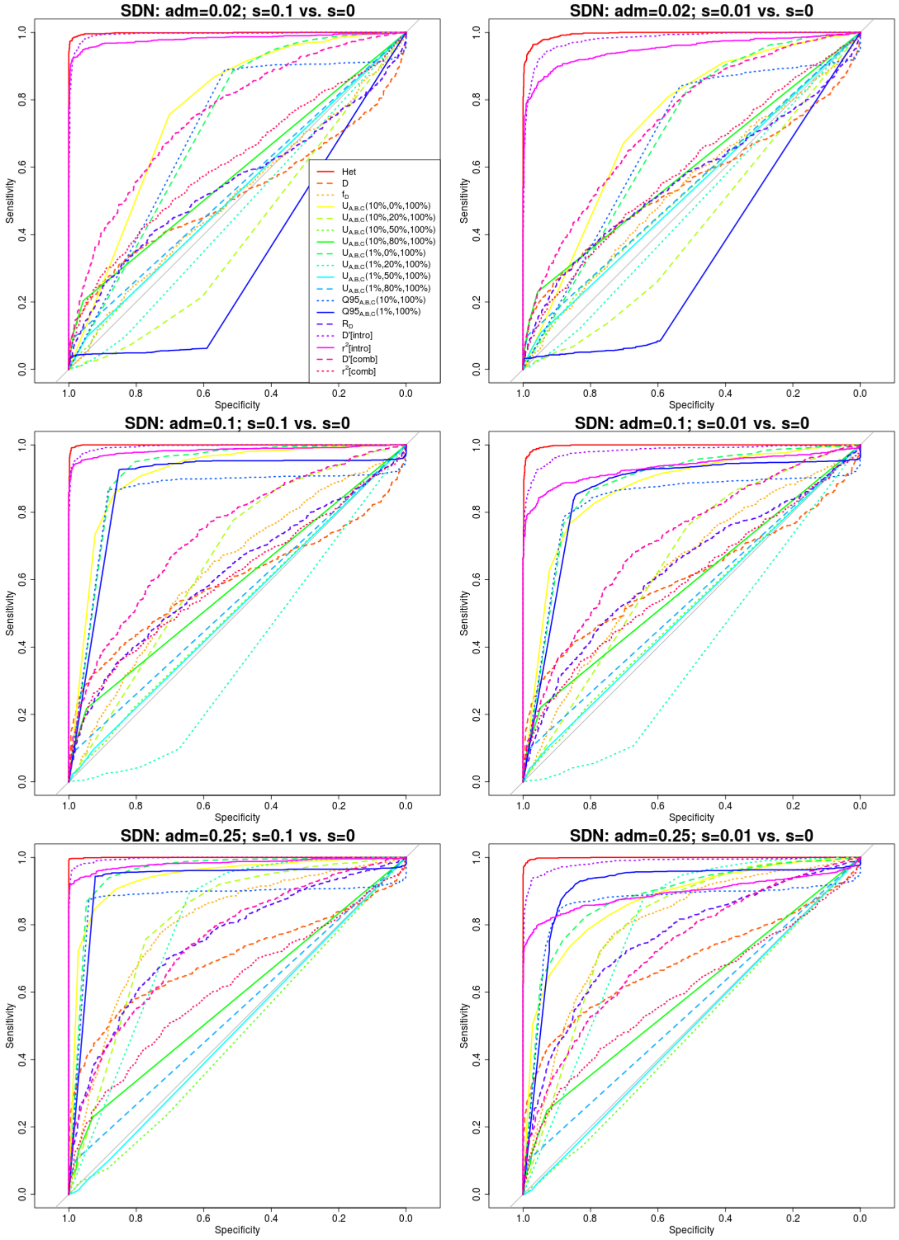


Figure 3.17. Receiver operating characteristic curves for a selective sweep from de novo mutation. We simulated the same demography as in Figure 3.8, but rather than introducing the beneficial variant in the introgressed population via admixture from an archaic population, we introduced it by mutation in the introgressed population (B) 3,900 generations ago.

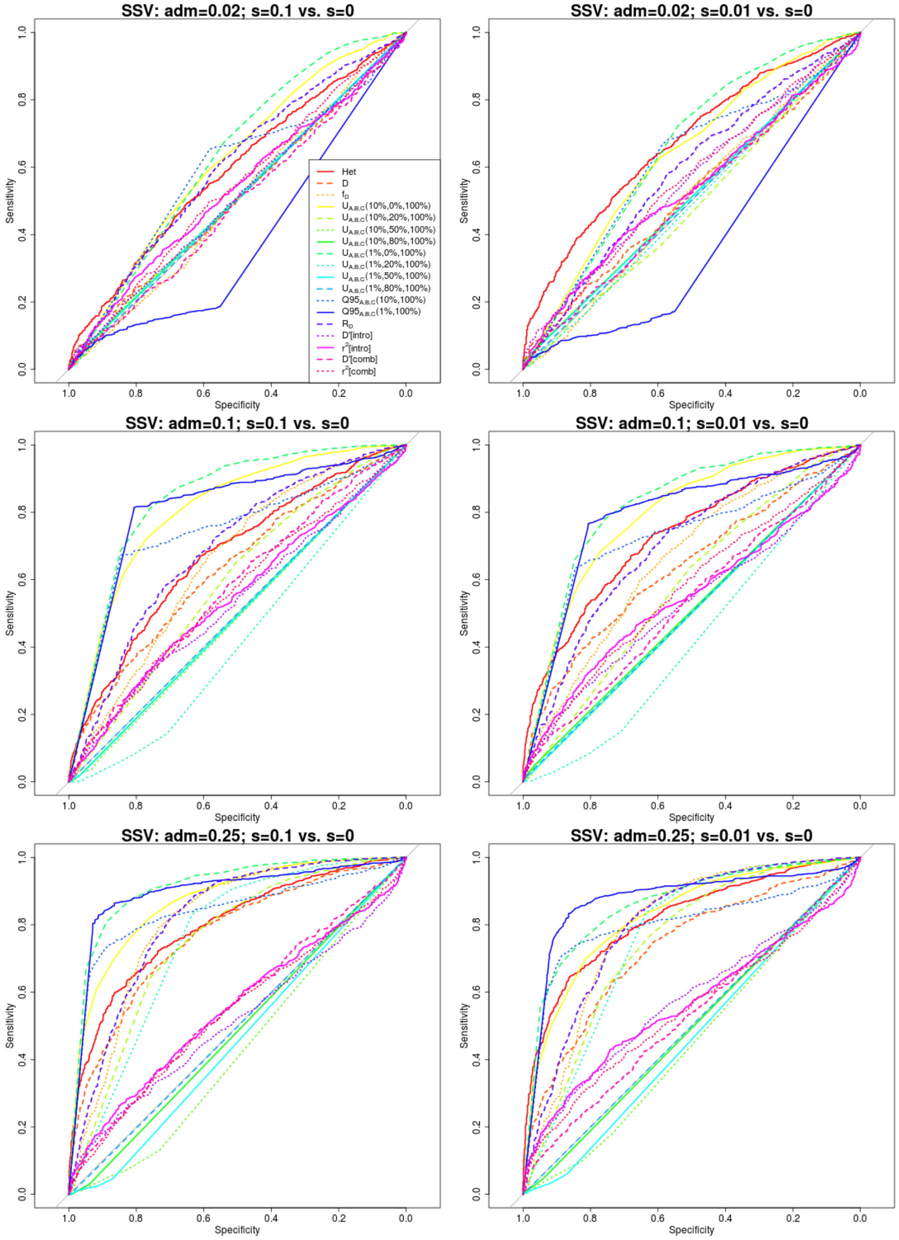


Figure 3.18. Receiver operating characteristic curves for selection from standing variation. We simulated the same demography as in Figure 3.8, but rather than introducing the beneficial variant in the introgressed population via admixture from an archaic population, we introduced it with a starting frequency of 20% in the introgressed population (B) 3,900 generations ago.

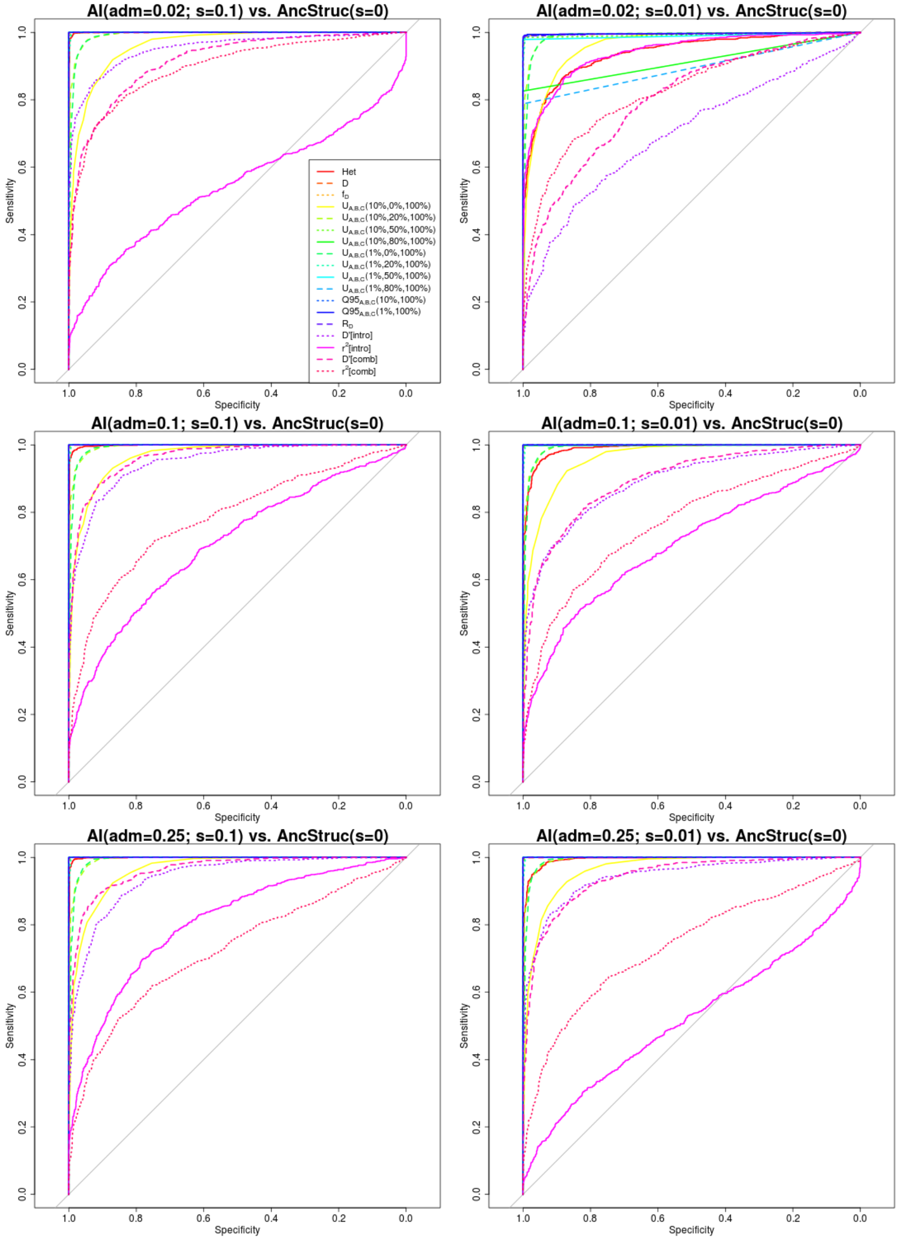


Figure 3.19. Receiving operating characteristic curves for adaptive introgression against a neutral ancestral structure model with strong migration rates. The demographic scenario for adaptive introgression was the same as in Figure 3.8. For a description of the ancestral structure model, see main text.

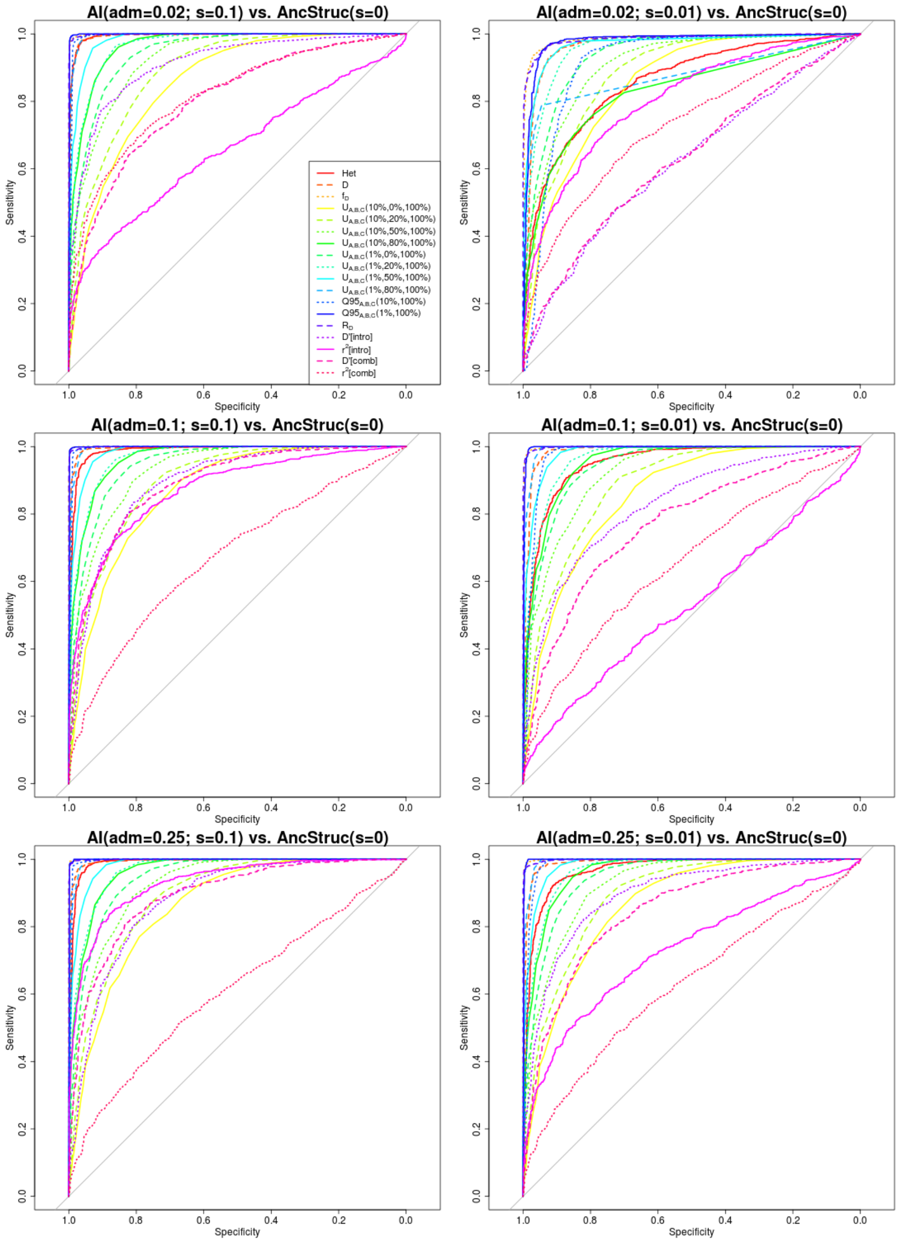


Figure 3.20. Receiving operating characteristic curves for adaptive introgression against a neutral ancestral structure model with intermediate migration rates. The demographic scenario for adaptive introgression was the same as in Figure 3.8. For a description of the ancestral structure model, see main text.

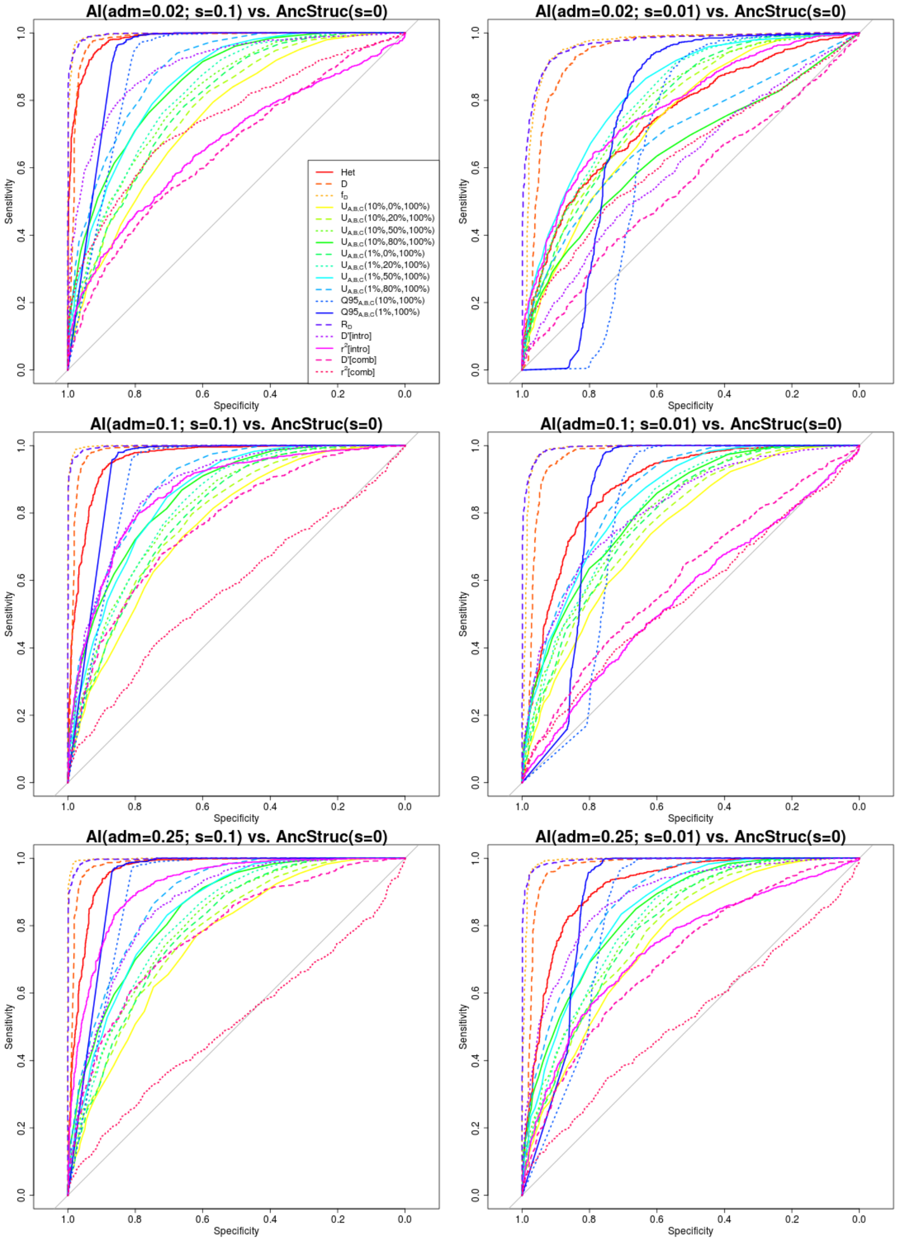


Figure 3.21. Receiving operating characteristic curves for adaptive introgression against a neutral ancestral structure model with weak migration rates. The demographic scenario for adaptive introgression was the same as in Figure 3.8. For a description of the ancestral structure model, see main text.

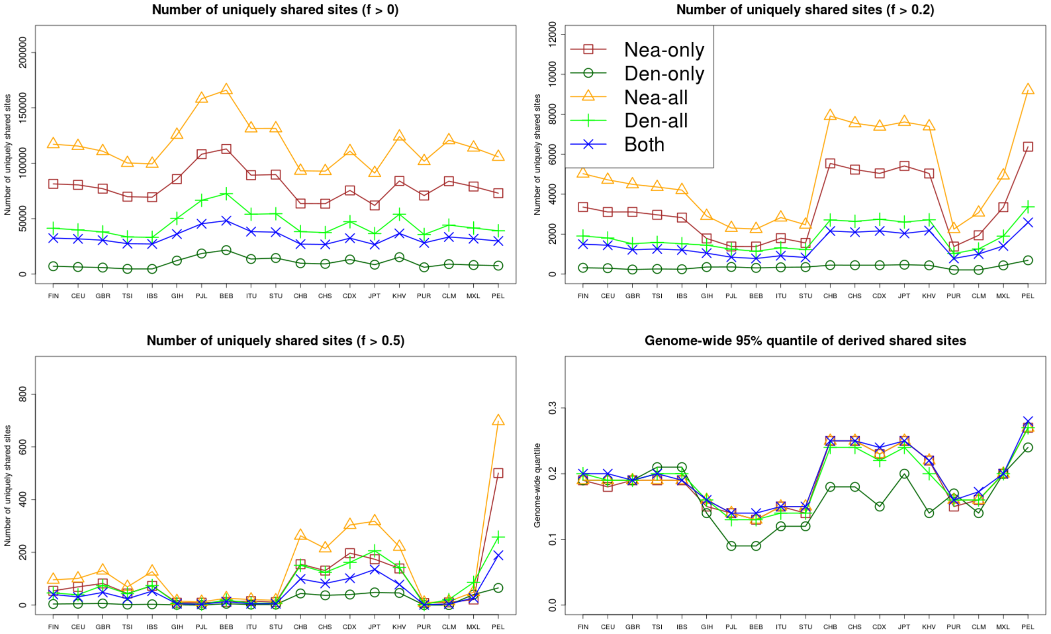


Figure 3.22. We computed the number of uniquely shared sites in the autosomes and the X chromosome between particular archaic humans and different choices of present-day non-African panels X (x-axis) from phase 3 of the 1000 Genomes Project. We used a shared frequency cutoff of 0% (top-left panel), 20% (top-right panel) and 50% (bottom-left panel). Nea-only = $U_{Afr,X,Nea,Den}(1\%, 20\%, 100\%, 0\%)$. Den-only = $U_{Afr,X,Nea,Den}(1\%, 20\%, 0\%, 100\%)$. Nea-all = $U_{Afr,X,Nea}(1\%, 20\%, 100\%)$. Den-all = $U_{Afr,X,Den}(1\%, 20\%, 100\%)$. Both = $U_{Afr,X,Nea,Den}(1\%, 20\%, 100\%, 100\%)$. We also computed the quantile statistics $Q95$ for different choices of present-day non-African human panels (x-axis) from phase 3 of the 1000 Genomes Project (bottom-right panel). Nea-only = $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 0\%)$. Den-only = $Q95_{Afr,X,Nea,Den}(1\%, 0\%, 100\%)$. Nea-all = $Q95_{Afr,X,Nea}(1\%, 100\%)$. Den-all = $Q95_{Afr,X,Den}(1\%, 100\%)$. Both = $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 100\%)$.

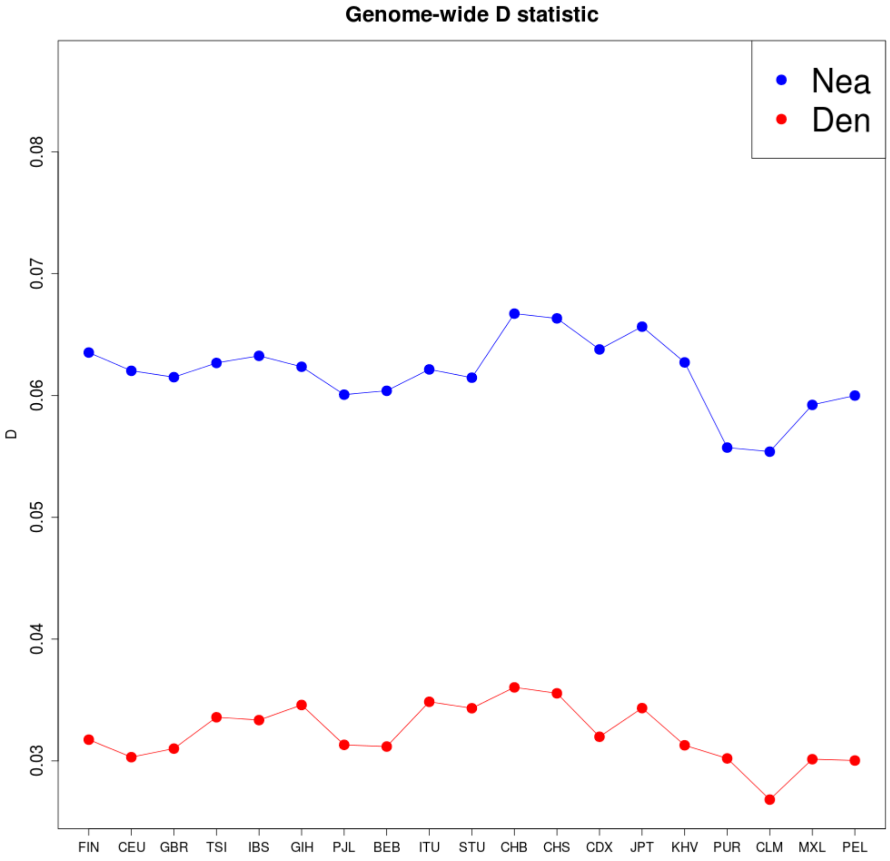


Figure 3.23. We computed D(X,YRI,Y,Chimpanzee) for different choices of present-day human panels X (x-axis) from phase 3 of the 1000 Genomes Project, and for two high-coverage archaic human genomes Y: Altai Neanderthal (blue) and Denisova (red). The low value of the right-most panel is due to that panel being composed of African-Americans, which have a higher proportion of African ancestry than the other panels.

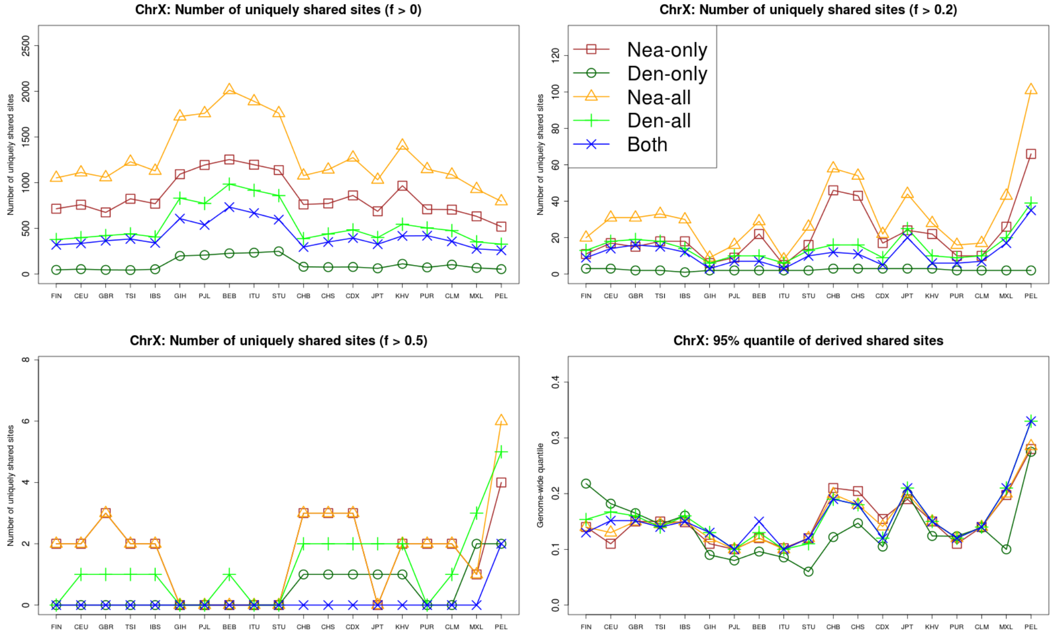


Figure 3.24. We computed the number of uniquely shared sites in the X chromosome between particular archaic humans genomes and different choices of present-day non-African human panels X (x-axis) from phase 3 of the 1000 Genomes Project, using a shared frequency cutoff of 0% (top-left panel), 20% (top-right panel) and 50% (bottom-left panel). Nea-only = $U_{Afr,X,Nea,Den}(1\%, 20\%, 100\%, 0\%)$. Den-only = $U_{Afr,X,Nea,Den}(1\%, 20\%, 0\%, 100\%)$. Nea-all = $U_{Afr,X,Nea}(1\%, 20\%, 100\%)$. Den-all = $U_{Afr,X,Den}(1\%, 20\%, 100\%)$. Both = $U_{Afr,X,Nea,Den}(1\%, 20\%, 100\%, 100\%)$. We also computed the quantile statistics Q95 for different choices of present-day non-African human panels (x-axis) from phase 3 of the 1000 Genomes Project (bottom-right panel). Nea-only = $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 0\%)$. Den-only = $Q95_{Afr,X,Nea,Den}(1\%, 0\%, 100\%)$. Nea-all = $Q95_{Afr,X,Nea}(1\%, 100\%)$. Den-all = $Q95_{Afr,X,Den}(1\%, 50\%, 100\%)$. Both = $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 100\%)$.

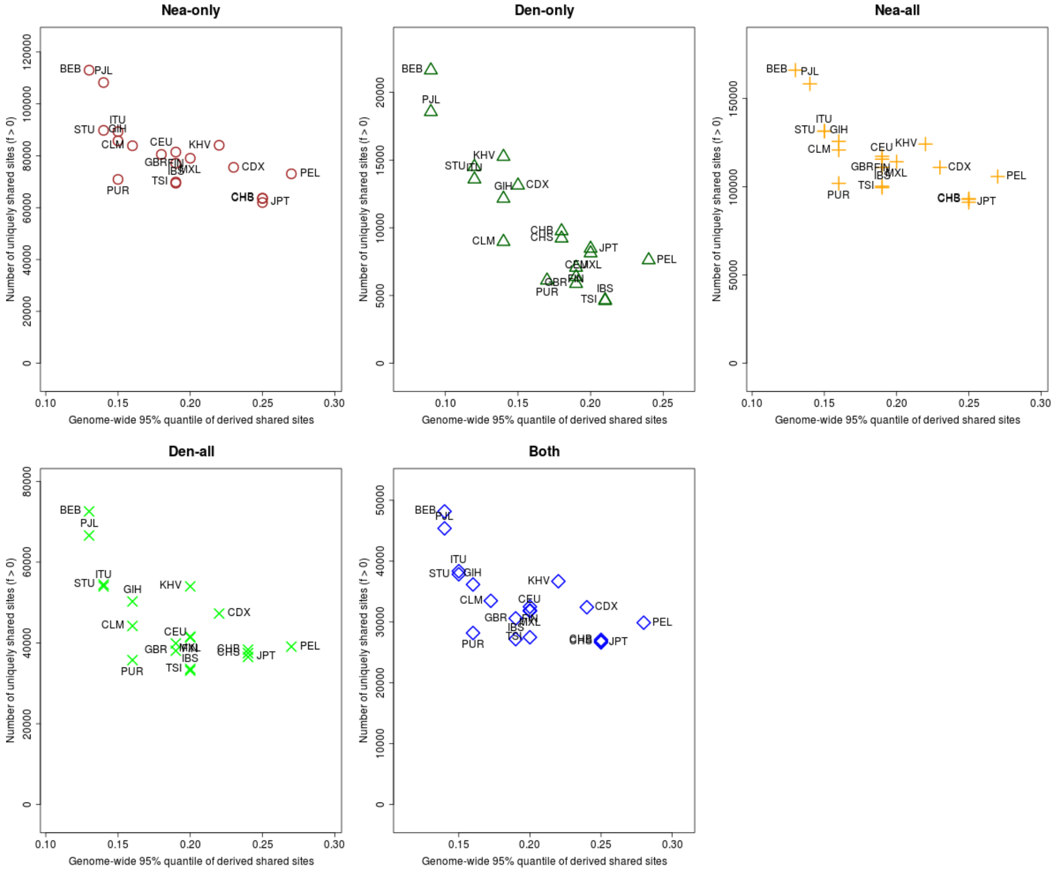


Figure 3.25. For each population panel from the 1000 Genomes Project, we jointly plotted the U and $Q95$ statistics with an archaic frequency cutoff of $> 0\%$ within each population. Nea-only $= U_{Afr,X,Nea,Den}(1\%, 0\%, 100\%, 0\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 0\%)$. Den-only $= U_{Afr,X,Nea,Den}(1\%, 0\%, 0\%, 100\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 0\%, 100\%)$. Nea-all $= U_{Afr,X,Nea}(1\%, 0\%, 100\%)$ and $Q95_{Afr,X,Nea}(1\%, 100\%)$. Den-all $= U_{Afr,X,Den}(1\%, 0\%, 100\%)$ and $Q95_{Afr,X,Den}(1\%, 100\%)$. Both $= U_{Afr,X,Nea,Den}(1\%, 0\%, 100\%, 100\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 100\%)$.

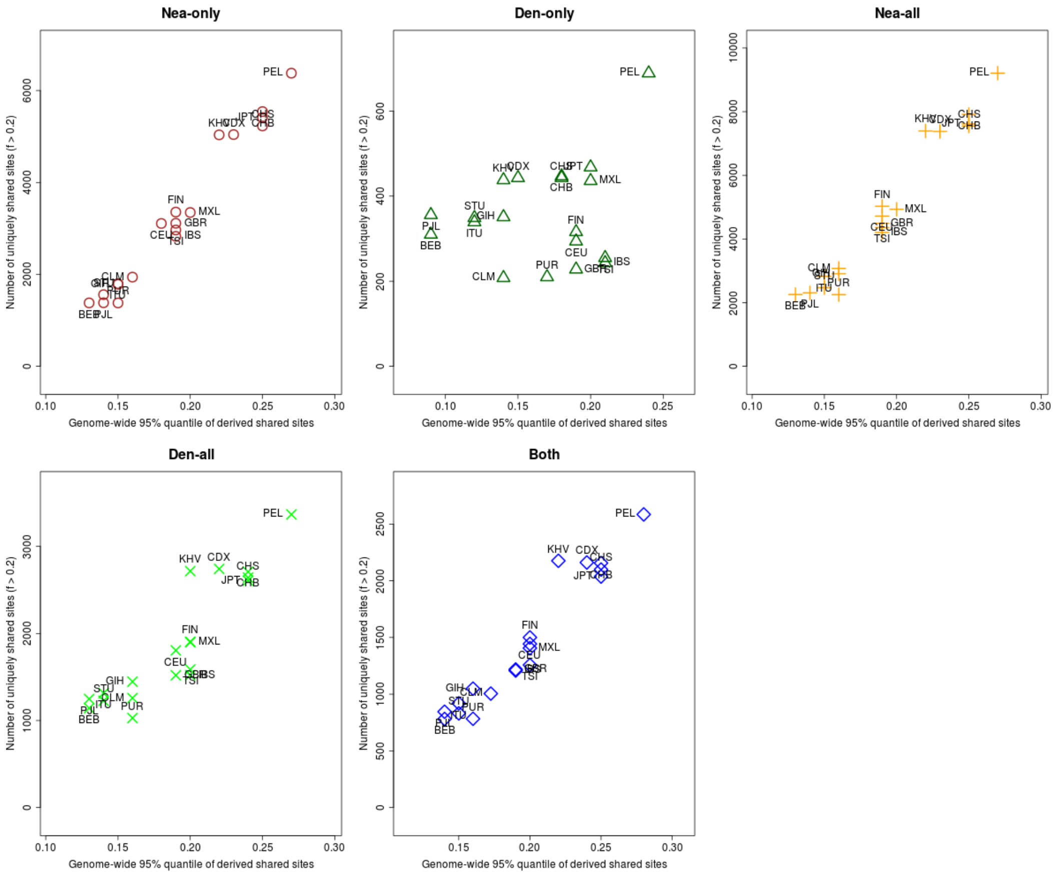


Figure 3.26. For each population panel from the 1000 Genomes Project, we jointly plotted the U and $Q95$ statistics with an archaic frequency cutoff of $> 20\%$ within each population. Nea-only $= U_{Afr,X,Nea,Den}(1\%, 20\%, 100\%, 0\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 0\%)$. Den-only $= U_{Afr,X,Nea,Den}(1\%, 20\%, 0\%, 100\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 0\%, 100\%)$. Nea-all $= U_{Afr,X,Nea}(1\%, 20\%, 100\%)$ and $Q95_{Afr,X,Nea}(1\%, 100\%)$. Den-all $= U_{Afr,X,Den}(1\%, 20\%, 100\%)$ and $Q95_{Afr,X,Den}(1\%, 100\%)$. Both $= U_{Afr,X,Nea,Den}(1\%, 20\%, 100\%, 100\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 100\%)$.

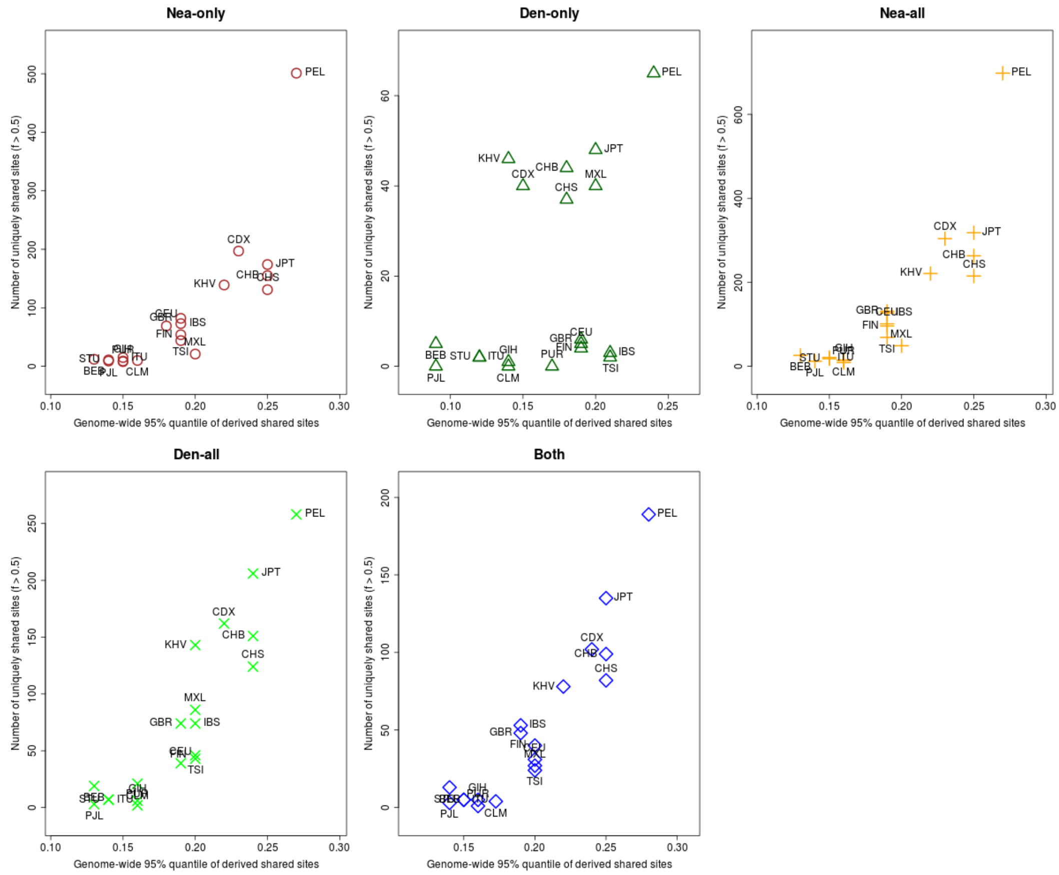


Figure 3.27. For each population panel from the 1000 Genomes Project, we jointly plotted the U and $Q95$ statistics with an archaic frequency cutoff of $> 50\%$ within each population. Nea-only $= U_{Afr,X,Nea,Den}(1\%, 50\%, 100\%, 0\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 0\%)$. Den-only $= U_{Afr,X,Nea,Den}(1\%, 50\%, 0\%, 100\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 0\%, 100\%)$. Nea-all $= U_{Afr,X,Nea}(1\%, 50\%, 100\%)$ and $Q95_{Afr,X,Nea}(1\%, 100\%)$. Den-all $= U_{Afr,X,Den}(1\%, 50\%, 100\%)$ and $Q95_{Afr,X,Den}(1\%, 100\%)$. Both $= U_{Afr,X,Nea,Den}(1\%, 50\%, 100\%, 100\%)$ and $Q95_{Afr,X,Nea,Den}(1\%, 100\%, 100\%)$.

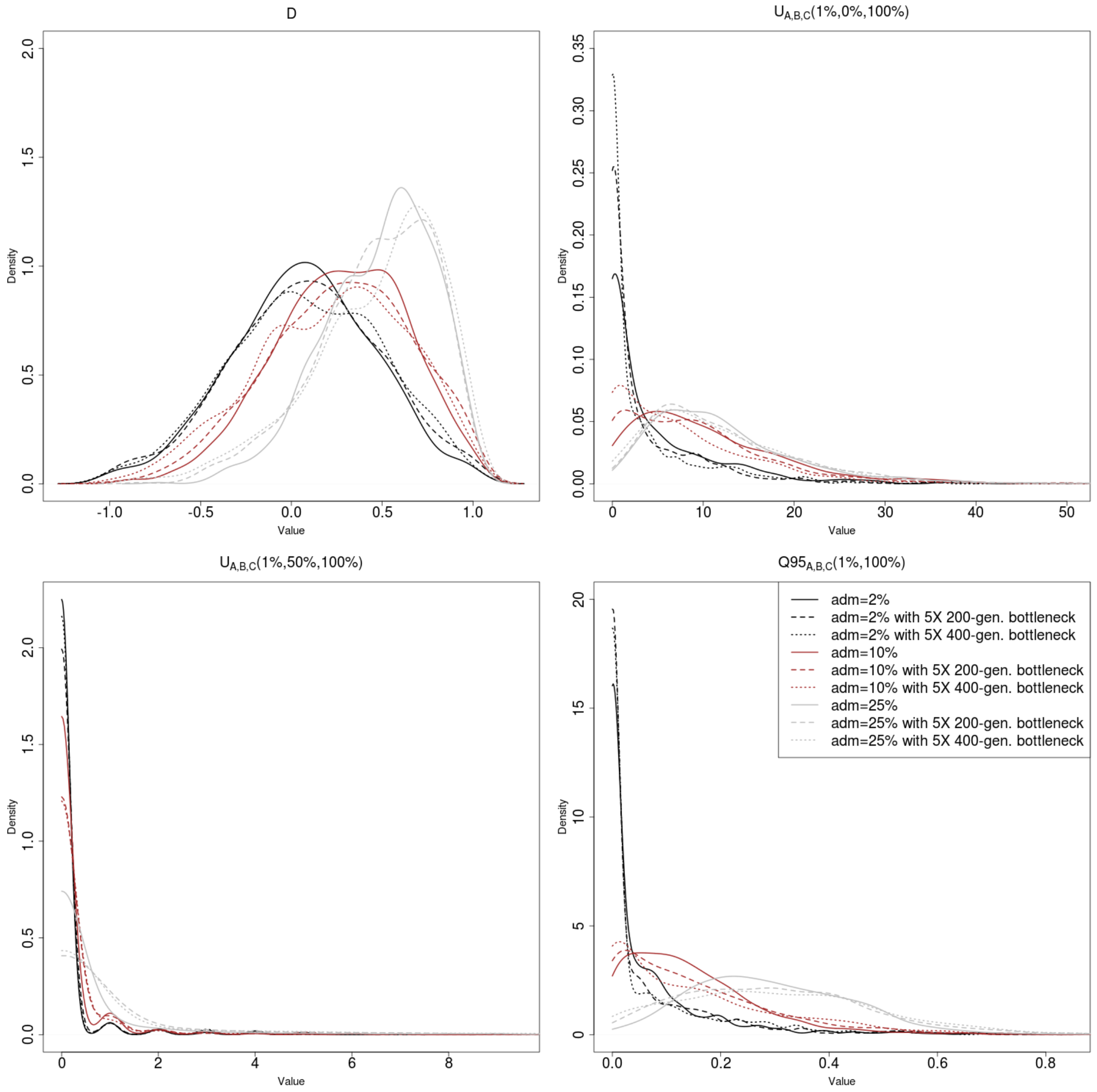


Figure 3.28. Effect of bottlenecks on the distribution of various statistics under introgression and neutrality.

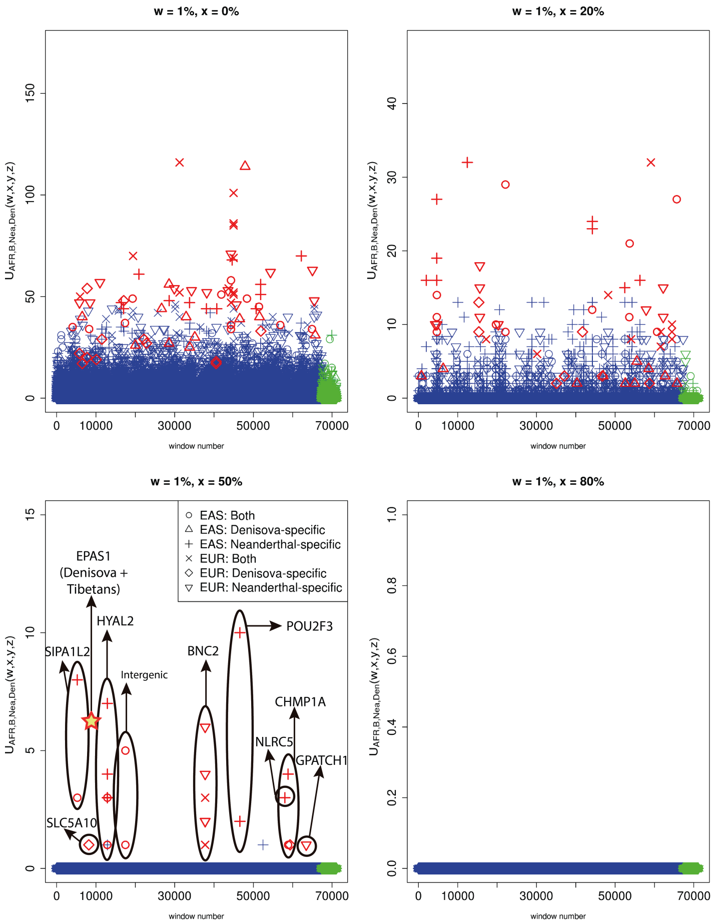


Figure 3.29. We partitioned the genome into non-overlapping windows of 40kb. Within each window, we computed $U_{EUR,Out,Nea,Den}(1\%, x, y, z)$ and $U_{EAS,Out,Nea,Den}(1\%, x, y, z)$, where Out=EAS+AFR for EUR as the target introgressed population, and Out=EUR+AFR for EAS as the target introgressed population. We searched for Neanderthal-specific alleles ($y = 100\%, z = 0\%$), Denisovan-specific alleles ($y = 0\%, z = 100\%$) and alleles present in both archaic genomes ($y = 100\%, z = 100\%$) that were uniquely shared with either EUR or EAS at various frequencies (z=0%, z=20%, z=50% and z=80%). Windows that fall within the upper tail of the distribution for each modern-archaic population pair are colored in red (P < 0.001 / number of pairs tested) and those that do not are colored in blue, except for those in the X chromosome, which are in green. Ovals drawn around multiple points contain multiple windows with uniquely shared alleles that are contiguous. For comparison, the number of high frequency uniquely shared sites between Denisova and Tibetans is also shown [33], although Tibetans are not included in the 1000 Genomes data and the region is 32 kb long, so this may be an underestimate.

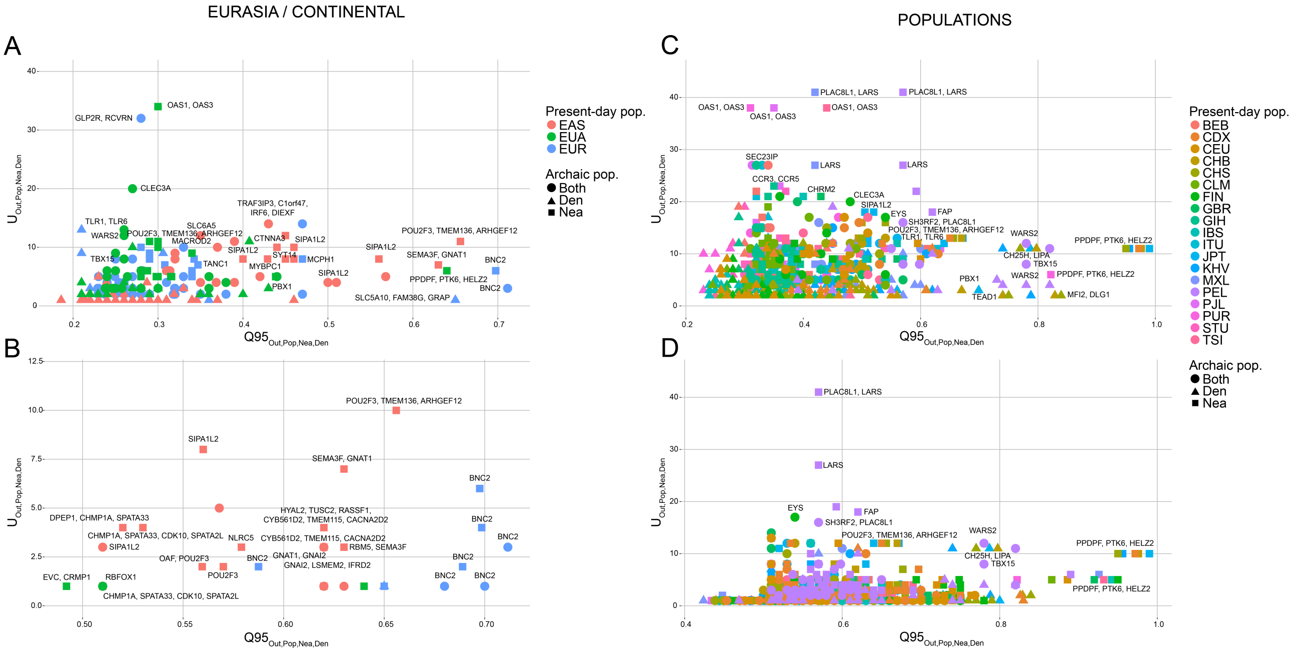


Figure 3.30. We plotted the 40kb regions in the 99.9% highest quantiles of both the $Q95_{Out,Pop,Nea,Den}(1\%, y, z)$ and $U_{Out,Pop,Nea,Den}(1\%, x, y, z)$ statistics for different choices of target introgressed population (Pop) and outgroup non-introgressed population (Out), and different archaic allele frequency cutoffs within the target population (x). A) We plotted the extreme regions for continental populations EUR (Out=EAS+AFR), EAS (Out=EUR+AFR) and Eurasians (EUA, Out=AFR), using a target population archaic allele frequency cutoff x of 20%. B) We plotted the extreme regions from the same statistics as in panel A, but with a more stringent target population archaic allele frequency cutoff x of 50%. C) We plotted the extreme regions for individual non-African populations within the 1000 Genomes data, using all African populations (excluding African-Americans) as the outgroup, and a cutoff x of 20%. D) We plotted the extreme regions from the same statistics as in panel C, but with a more stringent target population archaic allele frequency cutoff x of 50%. Nea-only = $U_{Out,Pop,Nea,Den}(1\%, x, 100\%, 0\%)$ and $Q95_{Out,Pop,Nea,Den}(1\%, 100\%, 0\%)$. Den-only = $U_{Out,Pop,Nea,Den}(1\%, x, 0\%, 100\%)$ and $Q95_{Out,Pop,Nea,Den}(1\%, 0\%, 100\%)$. Both = $U_{Out,Pop,Nea,Den}(1\%, x, 100\%, 100\%)$ and $Q95_{Out,Pop,Nea,Den}(1\%, 100\%, 100\%)$.

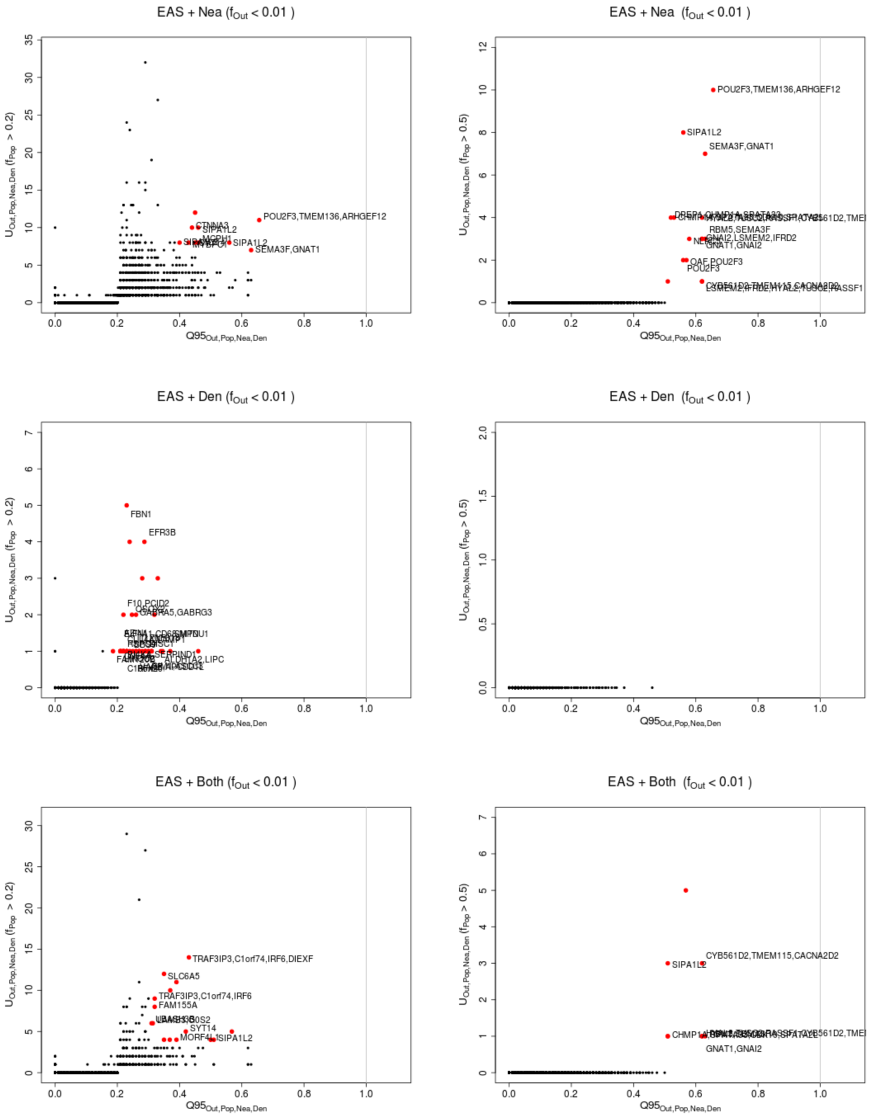


Figure 3.31. Uniquely shared archaic alleles in an East Asian (EAS) panel. Joint distribution of $Q95_{EUR+AFR,EAS,Nea,Den}(1\%, y, z)$ and $U_{EUR+AFR,EAS,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

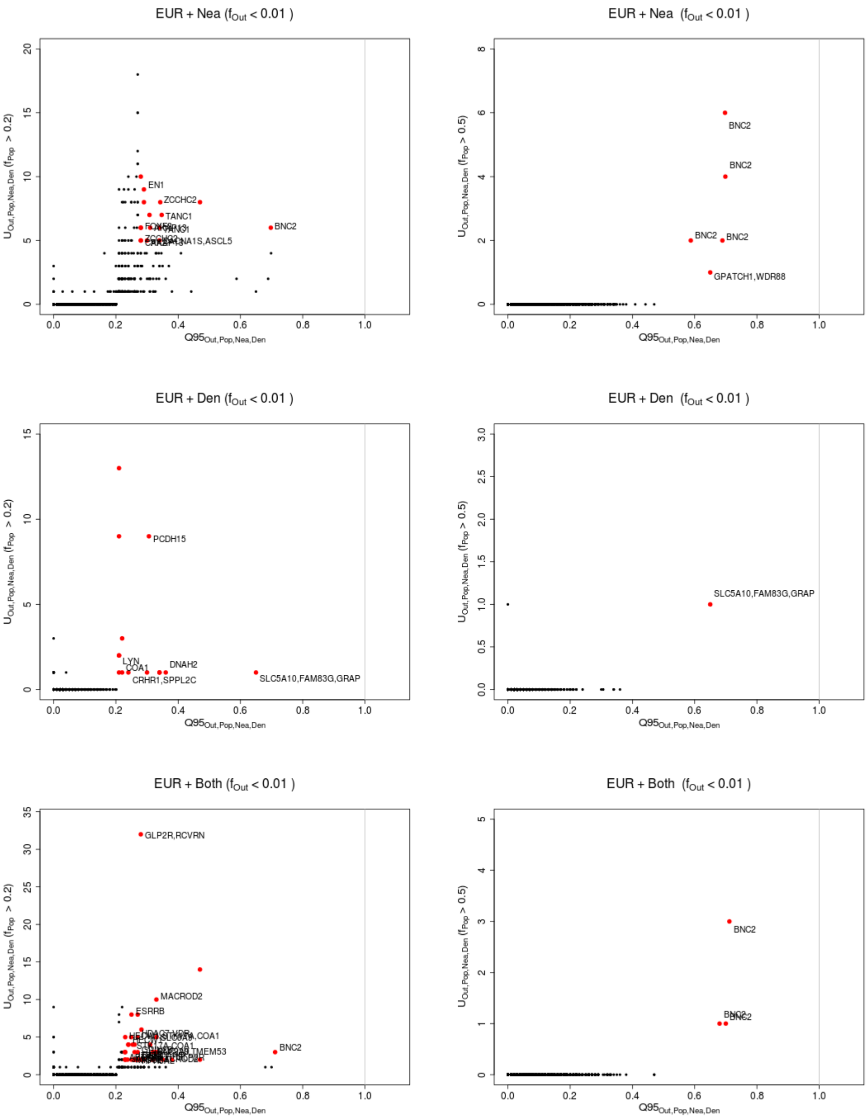


Figure 3.32. Uniquely shared archaic alleles in an European (EUR) panel. Joint distribution of $Q95_{EAS+AFR,EUR,Nea,Den}(1\%, y, z)$ and $U_{EAS+AFR,EUR,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

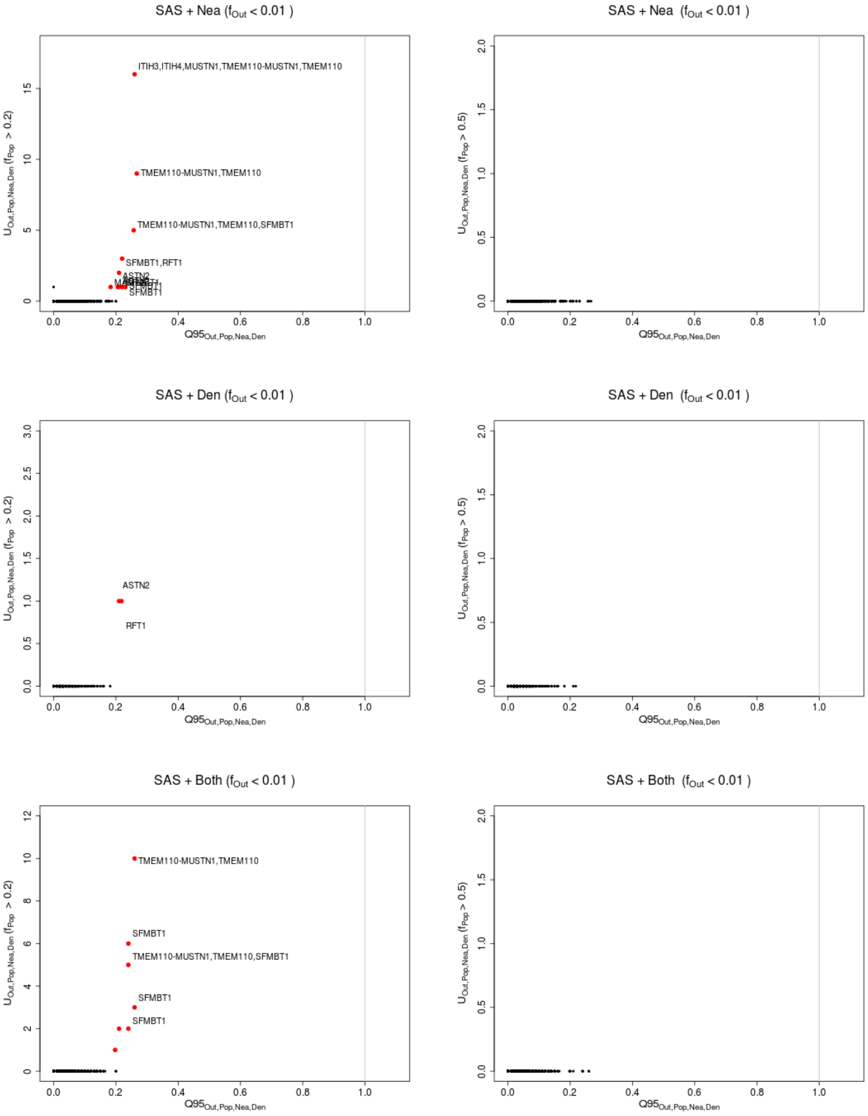


Figure 3.33. Uniquely shared archaic alleles in a South Asian (SAS) panel. Joint distribution of $Q95_{EAS+EUR+AFR,SAS,Nea,Den}(1\%, y, z)$ and $U_{EAS+EUR+AFR,SAS,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

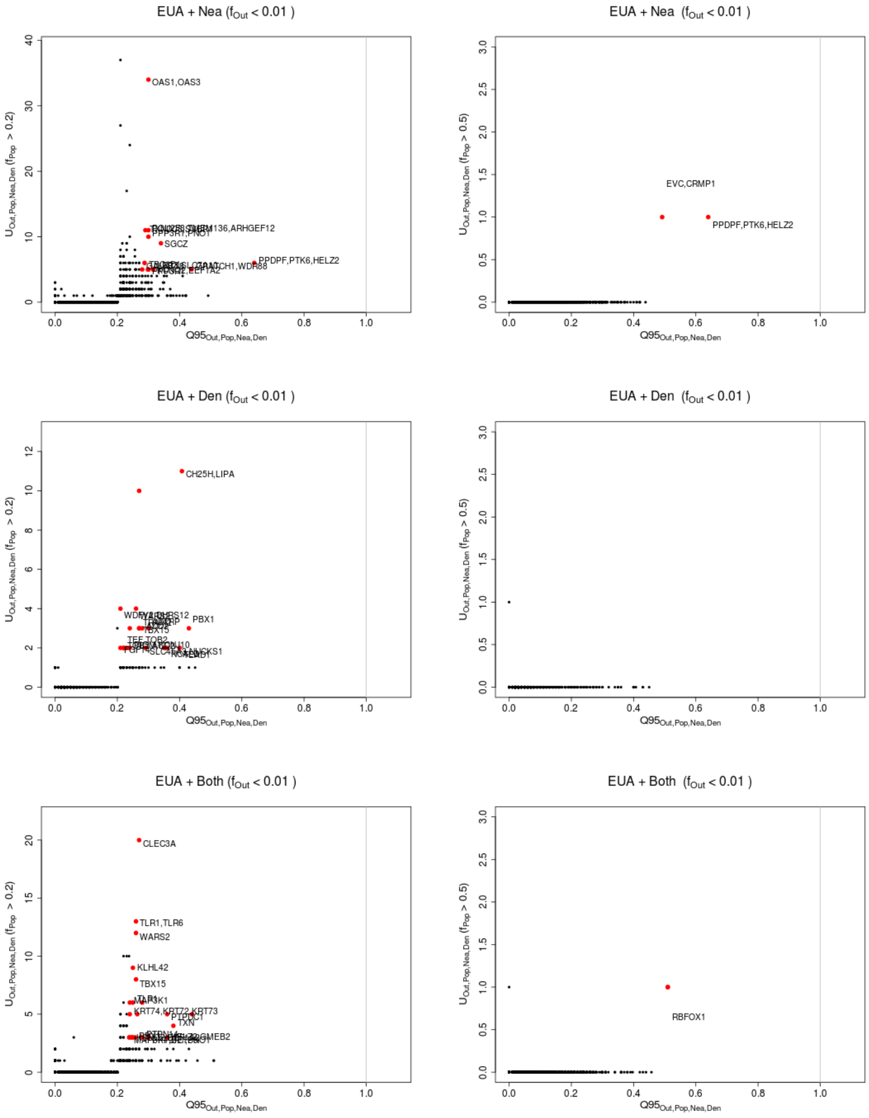


Figure 3.34. Uniquely shared archaic alleles in a Eurasian (EUA=EUR+SAS+EAS) panel. Joint distribution of $Q95_{AFR,EUR+SAS+EAS,Nea,Den}(1\%, y, z)$ and $U_{AFR,EUR+SAS+EAS,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

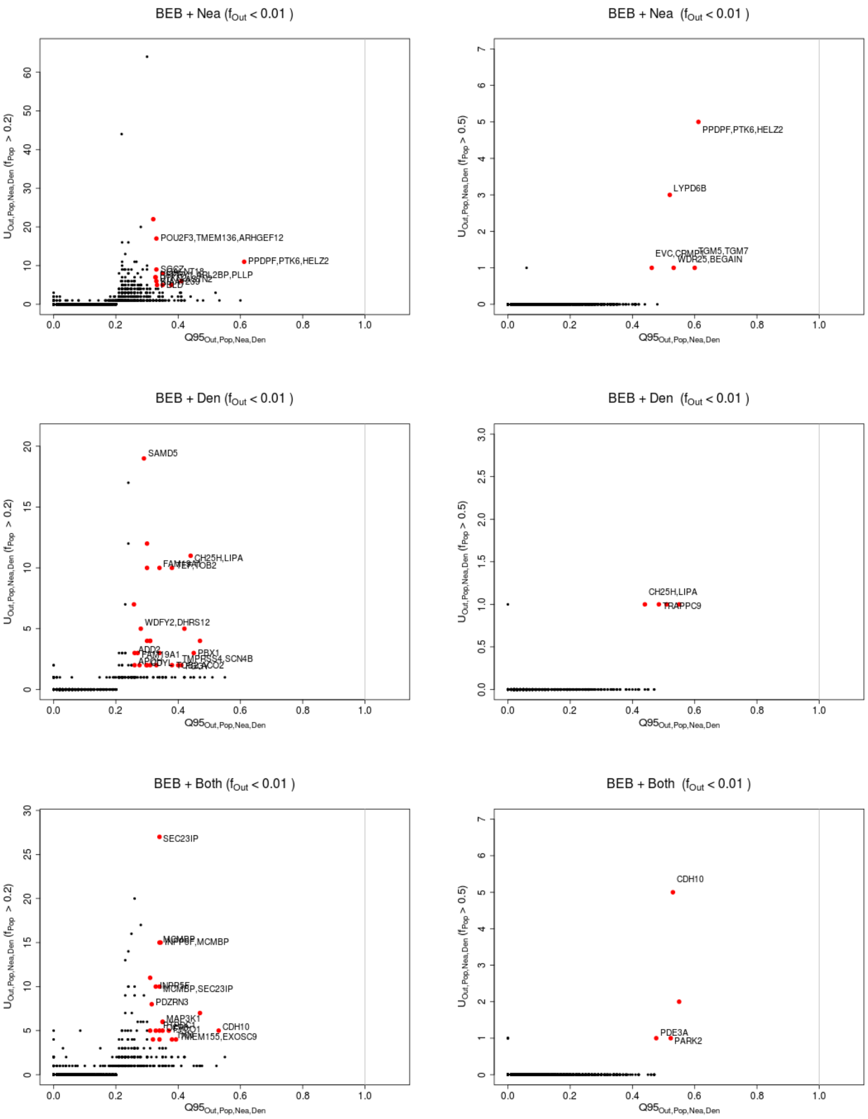


Figure 3.35. Uniquely shared archaic alleles in a Bengali (BEB) panel. Joint distribution of $Q95_{AFR,BEB,Nea,Den}(1\%, y, z)$ and $U_{AFR,BEB,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

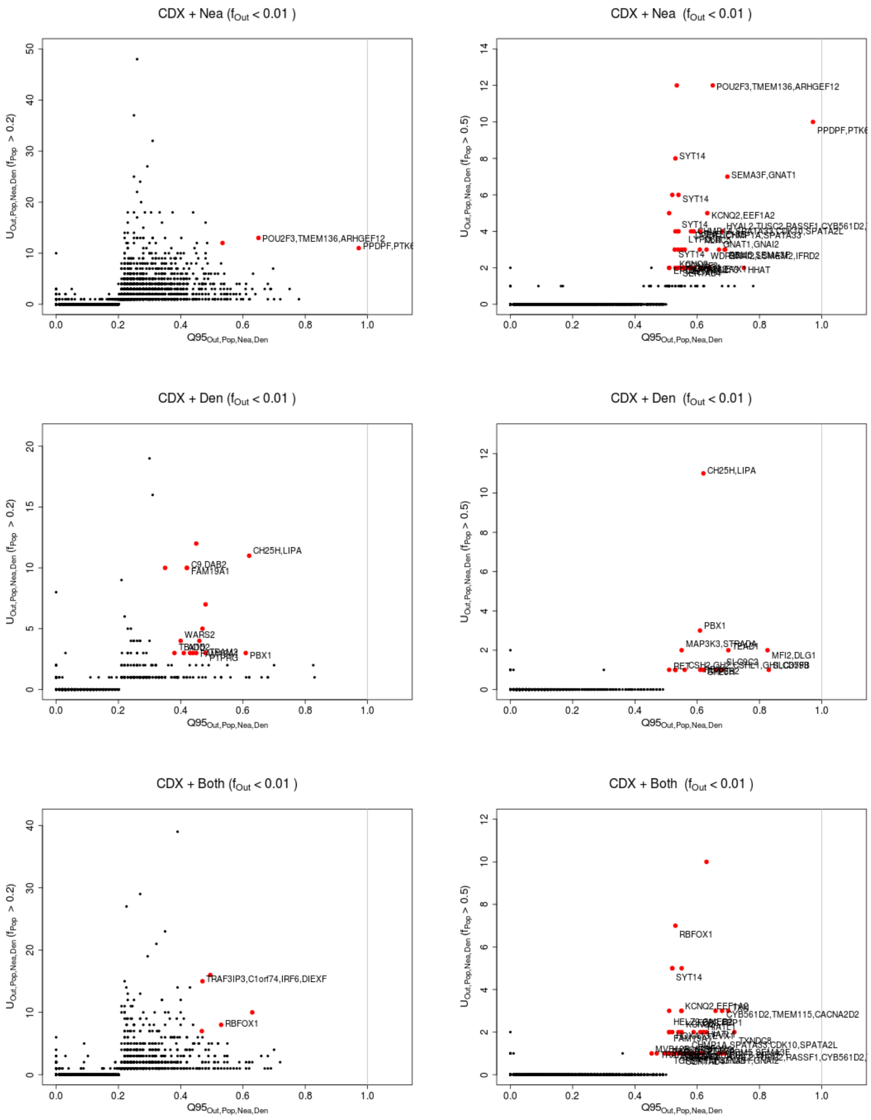


Figure 3.36. Uniquely shared archaic alleles in a Chinese Dai (CDX) panel. Joint distribution of $Q95_{AFR,CDX,Nea,Den}(1\%, y, z)$ and $U_{AFR,CDX,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

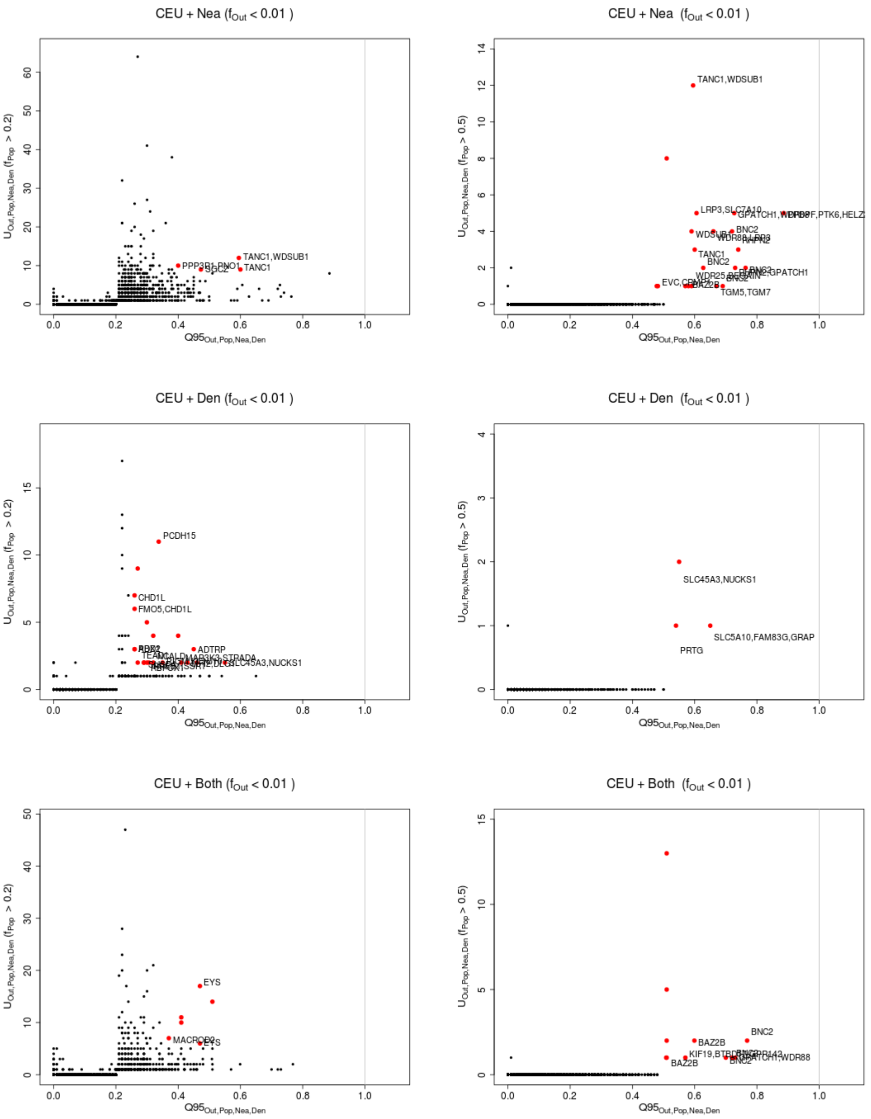


Figure 3.37. Uniquely shared archaic alleles in a Central European (CEU) panel. Joint distribution of $Q95_{AFR,CEU,Nea,Den}(1\%, y, z)$ and $U_{AFR,CEU,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

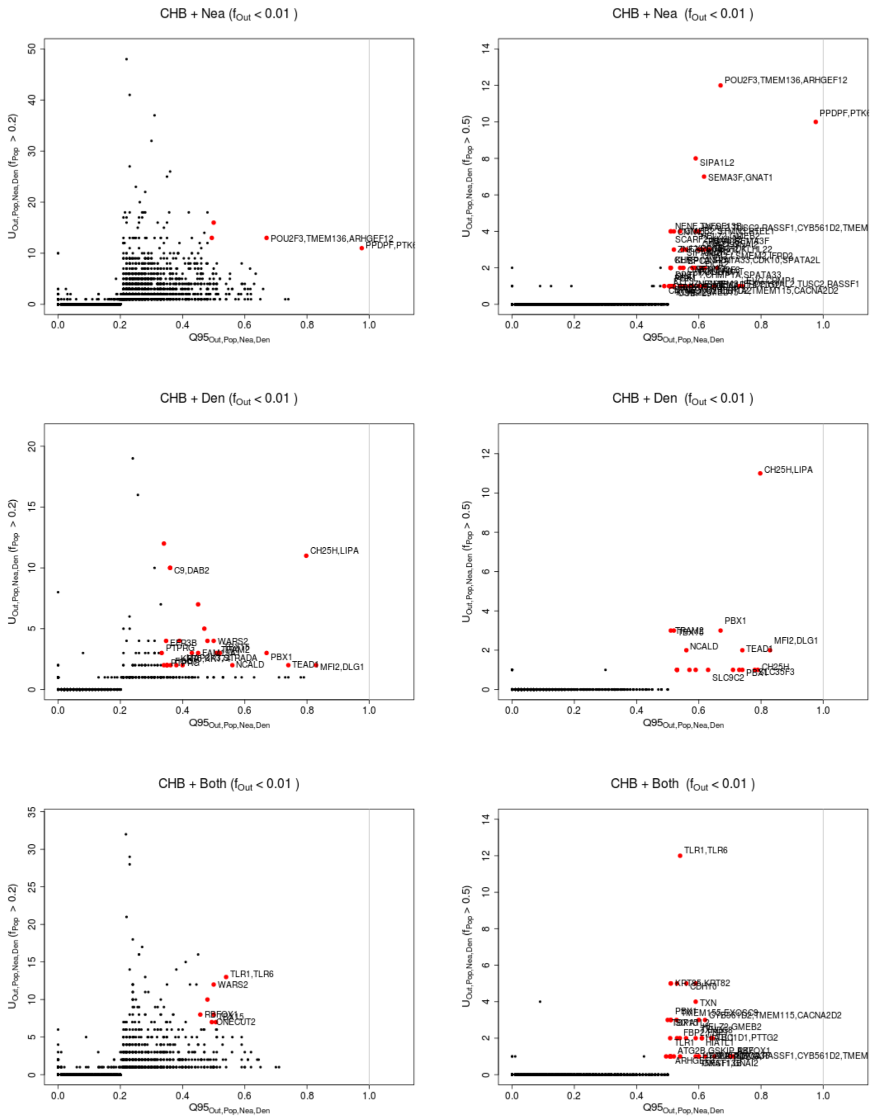


Figure 3.38. Uniquely shared archaic alleles in a Han Chinese (CHB) panel. Joint distribution of $Q95_{AFR,CHB,Nea,Den}(1\%, y, z)$ and $U_{AFR,CHB,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

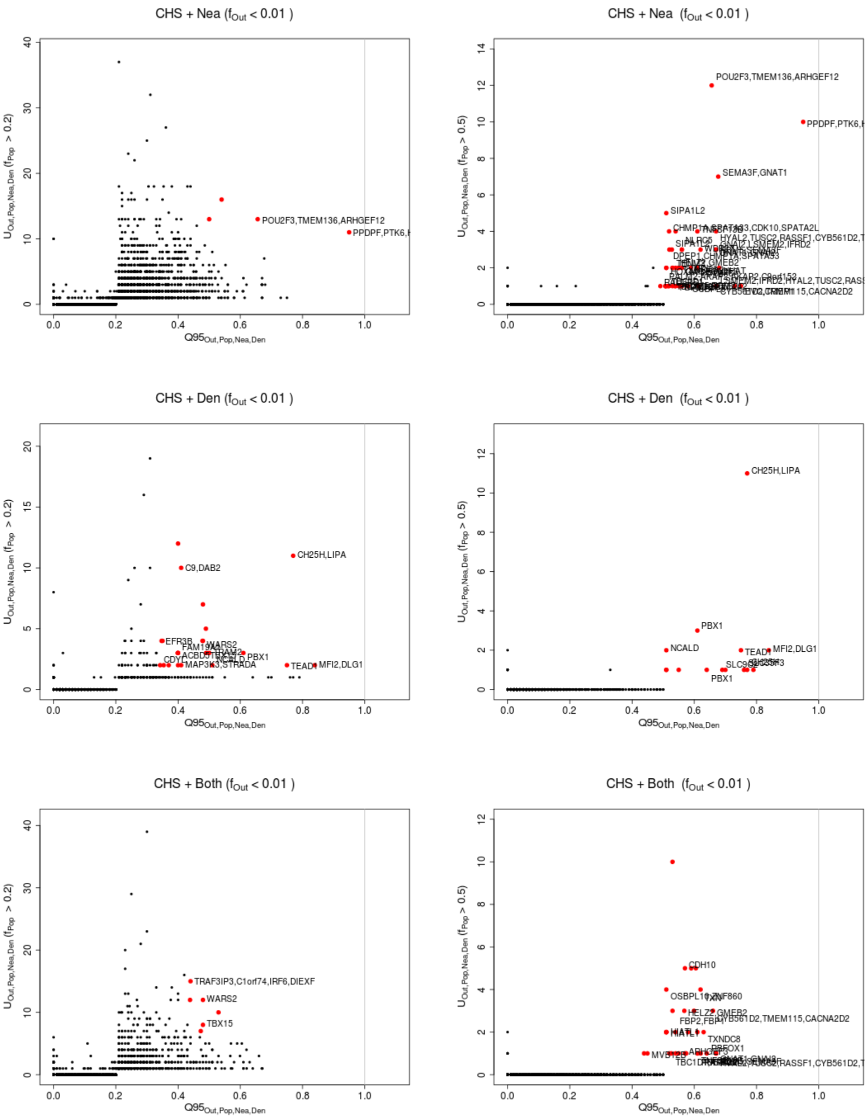


Figure 3.39. Uniquely shared archaic alleles in a Southern Han Chinese (CHS) panel. Joint distribution of $Q95_{AFR,BEB,Nea,Den}(1\%, y, z)$ and $U_{AFR,CHS,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

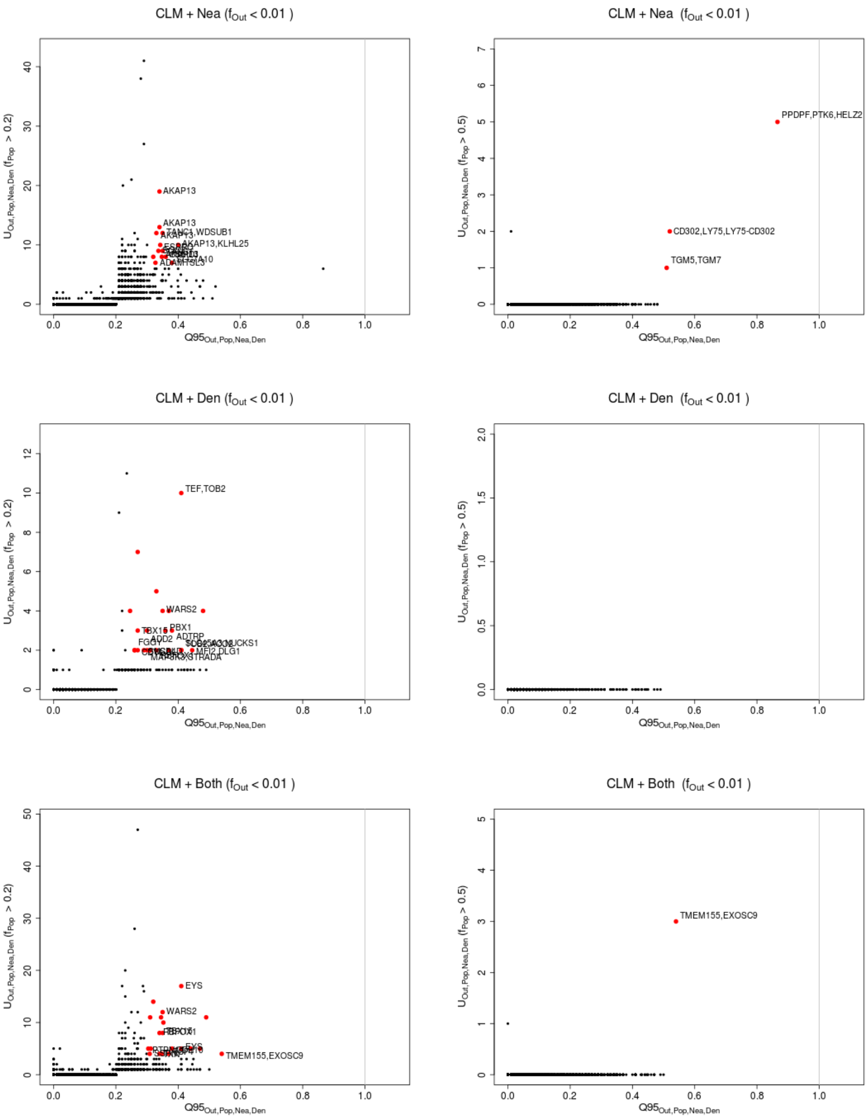


Figure 3.40. Uniquely shared archaic alleles in a Colombian (CLM) panel. Joint distribution of $Q95_{AFR,CLM,Nea,Den}(1\%, y, z)$ and $U_{AFR,CLM,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

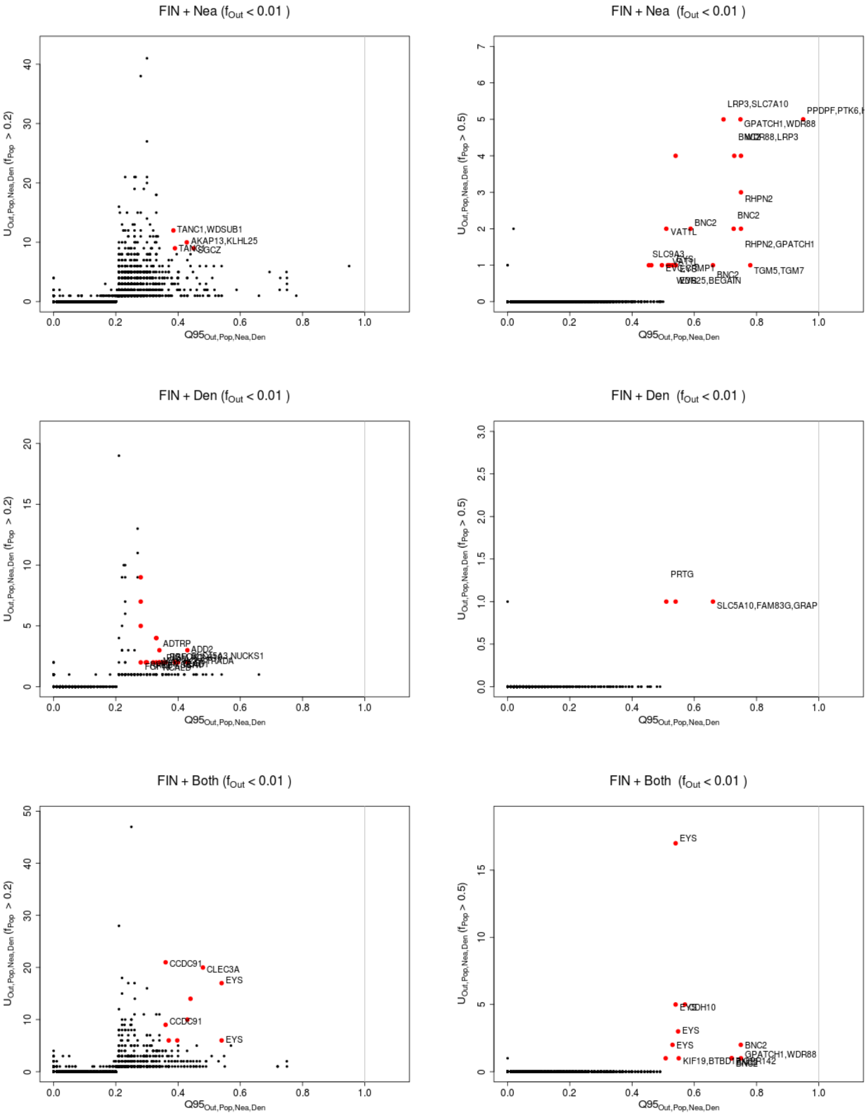


Figure 3.41. Uniquely shared archaic alleles in a Finnish (FIN) panel. Joint distribution of $Q95_{AFR,FIN,Nea,Den}(1\%, y, z)$ and $U_{AFR,FIN,Nea,Den}(1\%$,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

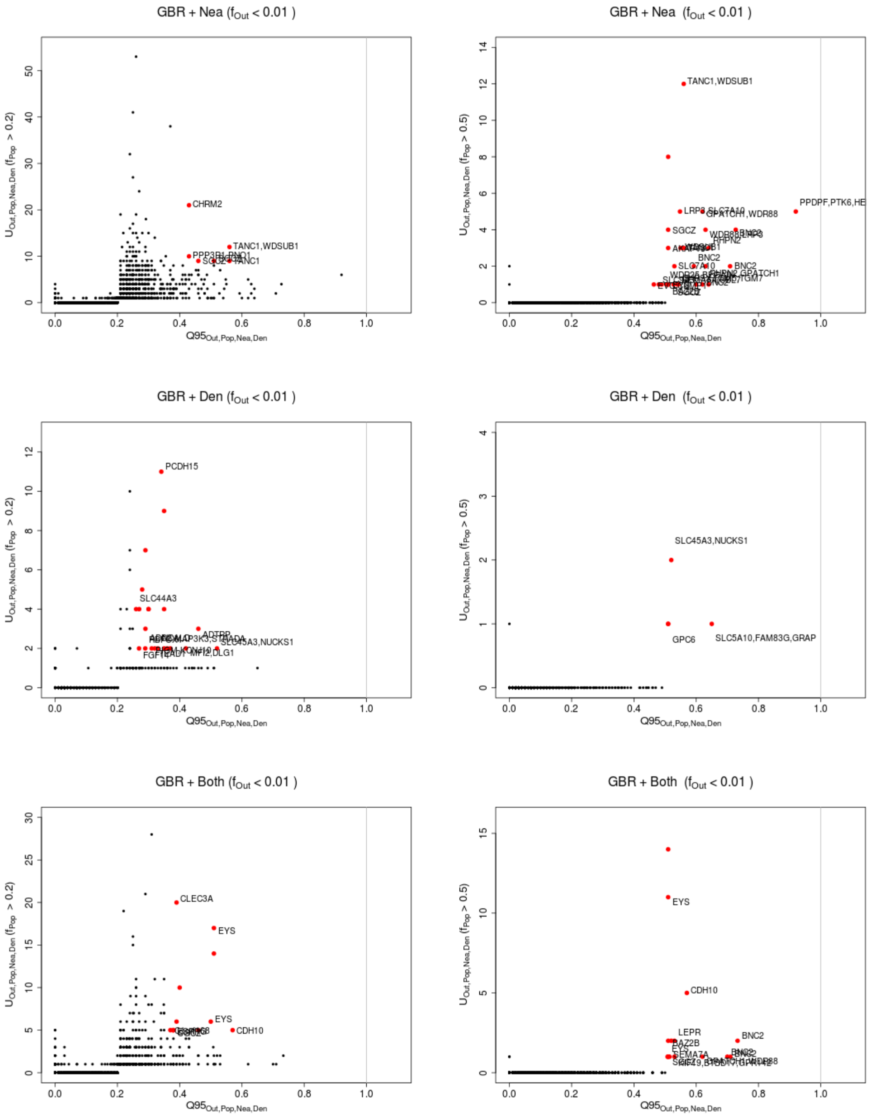


Figure 3.42. Uniquely shared archaic alleles in a British (GBR) panel. Joint distribution of $Q95_{AFR,GBR,Nea,Den}(1\%, y, z)$ and $U_{AFR,GBR,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

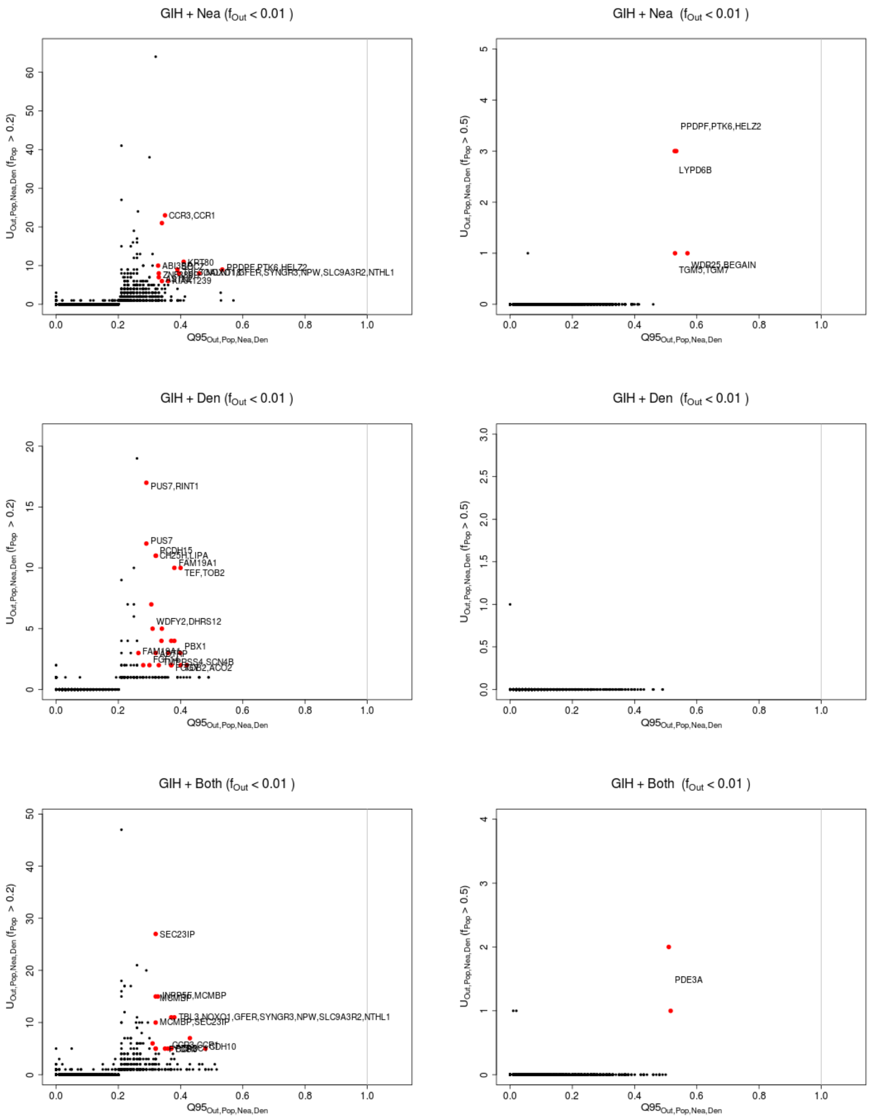


Figure 3.43. Uniquely shared archaic alleles in a Gujarati Indian (GIH) panel. Joint distribution of $Q95_{AFR,GIH,Nea,Den}(1\%, y, z)$ and $U_{AFR,GIH,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

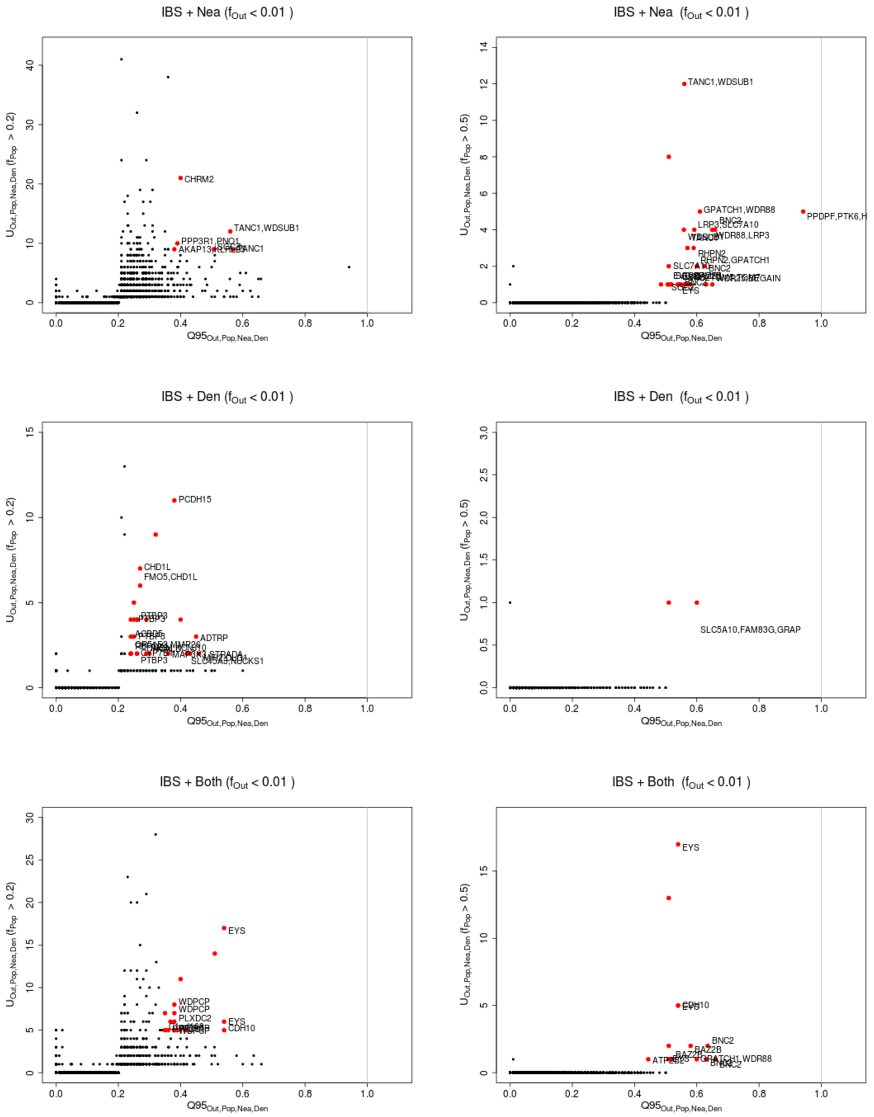


Figure 3.44. Uniquely shared archaic alleles in an Iberian (IBS) panel. Joint distribution of $Q95_{AFR,IBS,Nea,Den}(1\%, y, z)$ and $U_{AFR,IBS,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

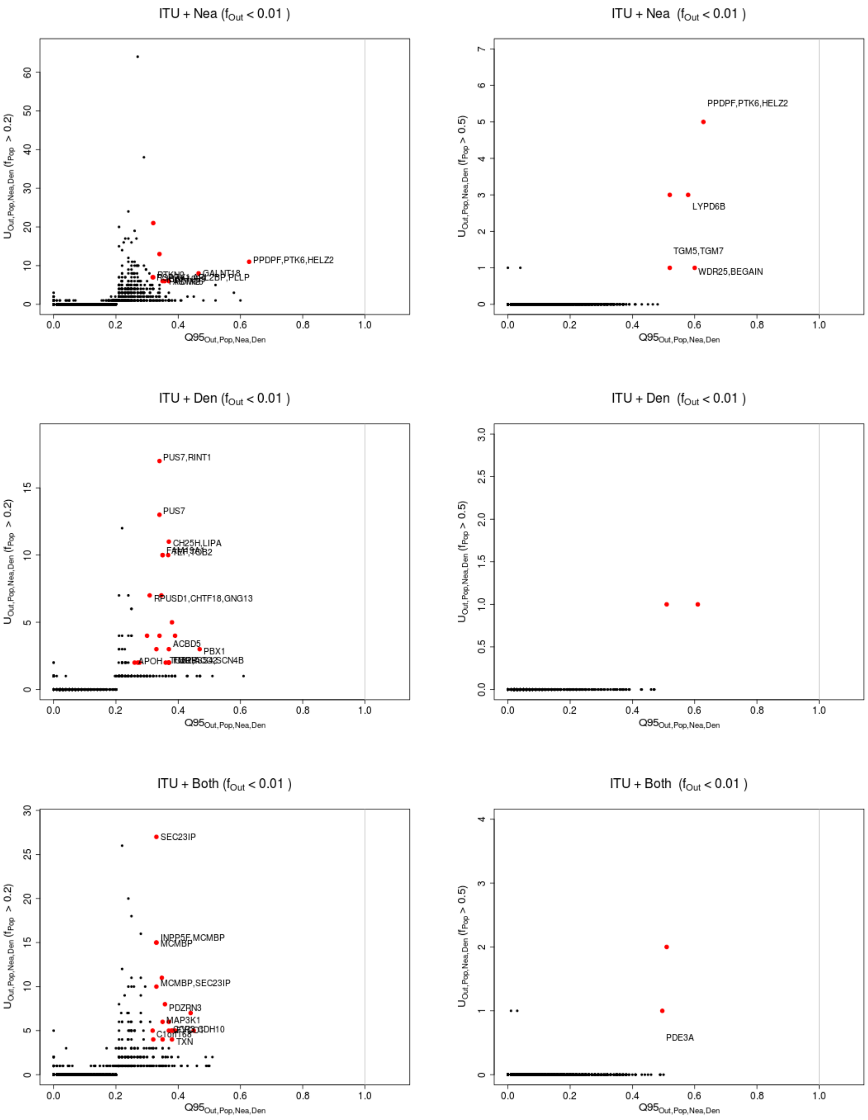


Figure 3.45. Uniquely shared archaic alleles in an Indian Telugu (ITU) panel. Joint distribution of $Q95_{AFR,ITU,Nea,Den}(1\%, y, z)$ and $U_{AFR,ITU,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

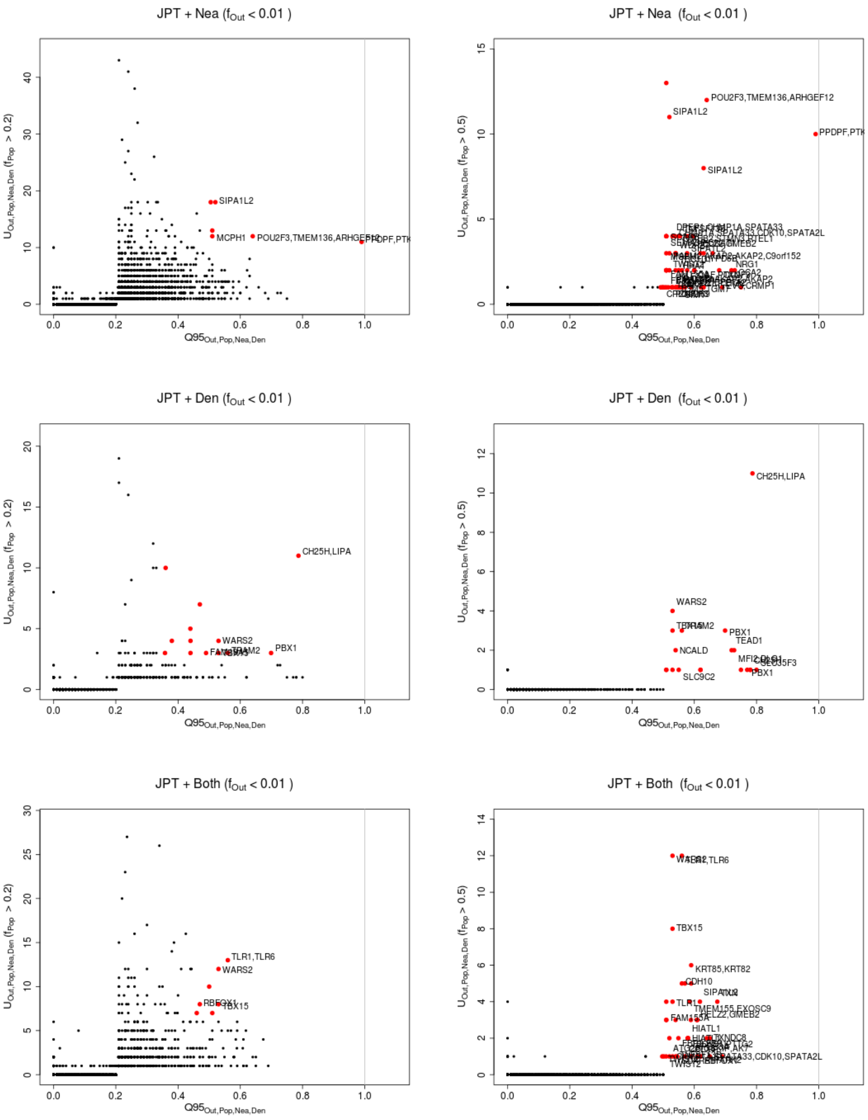


Figure 3.46. Uniquely shared archaic alleles in a Japanese (JPT) panel. Joint distribution of $Q95_{AFR,JPT,Nea,Den}(1\%, y, z)$ and $U_{AFR,JPT,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

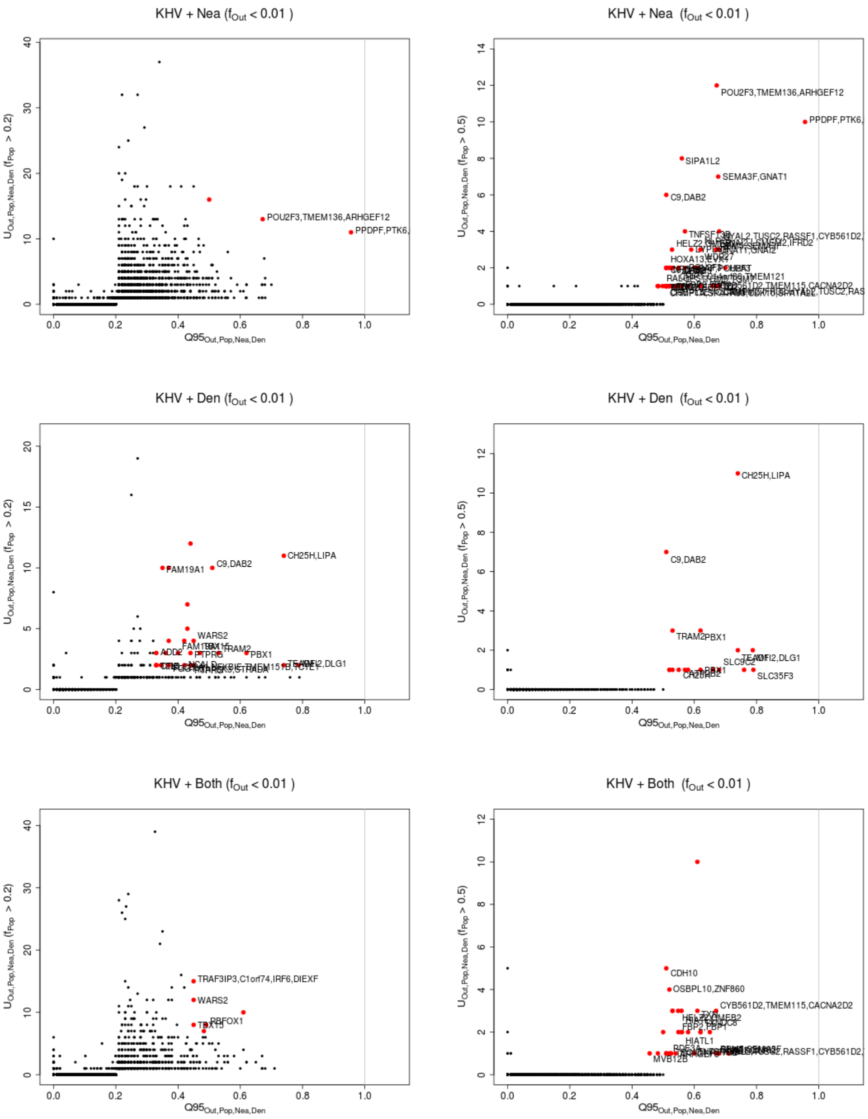


Figure 3.47. Uniquely shared archaic alleles in a Kinh (KHV) panel. Joint distribution of $Q95_{AFR,KHV,Nea,Den}(1\%, y, z)$ and $U_{AFR,KHV,Nea,Den}(1\%,x,y,z)$, for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

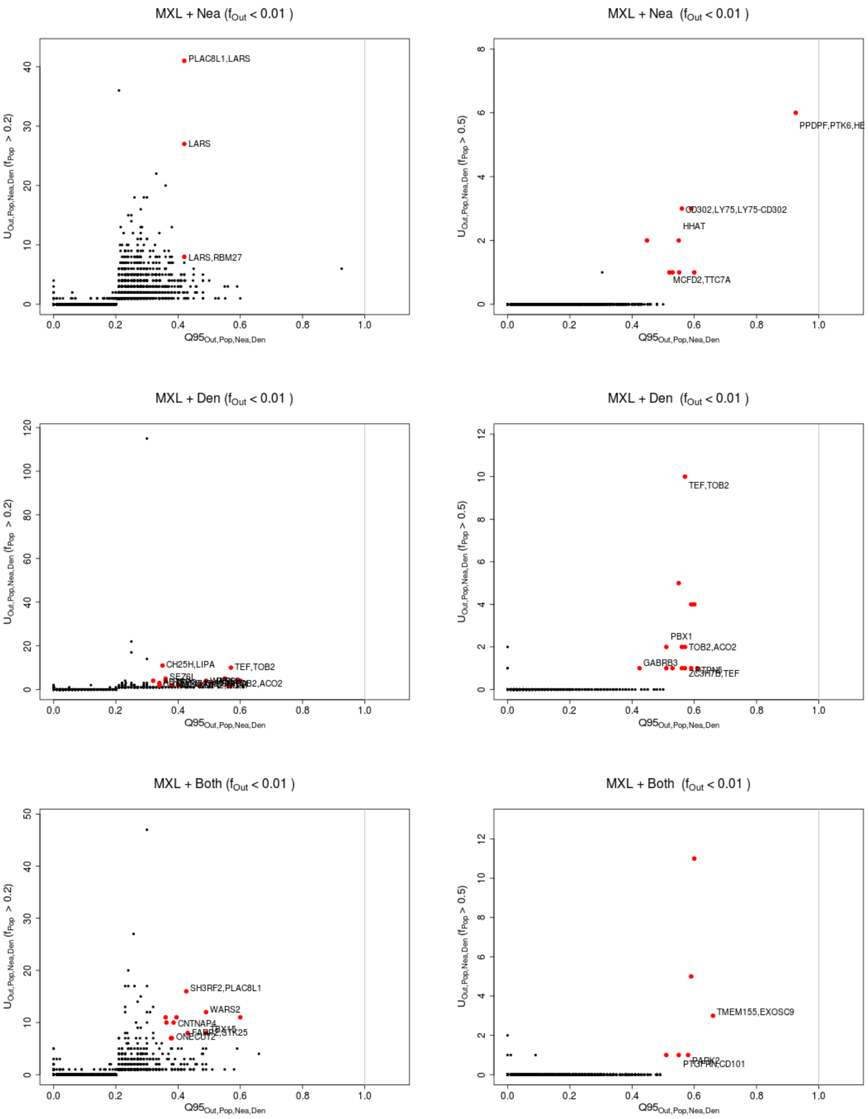


Figure 3.48. Uniquely shared archaic alleles in a Mexican (MXL) panel. Joint distribution of $Q95_{AFR,MXL,Nea,Den}(1\%, y, z)$ and $U_{AFR,MXL,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

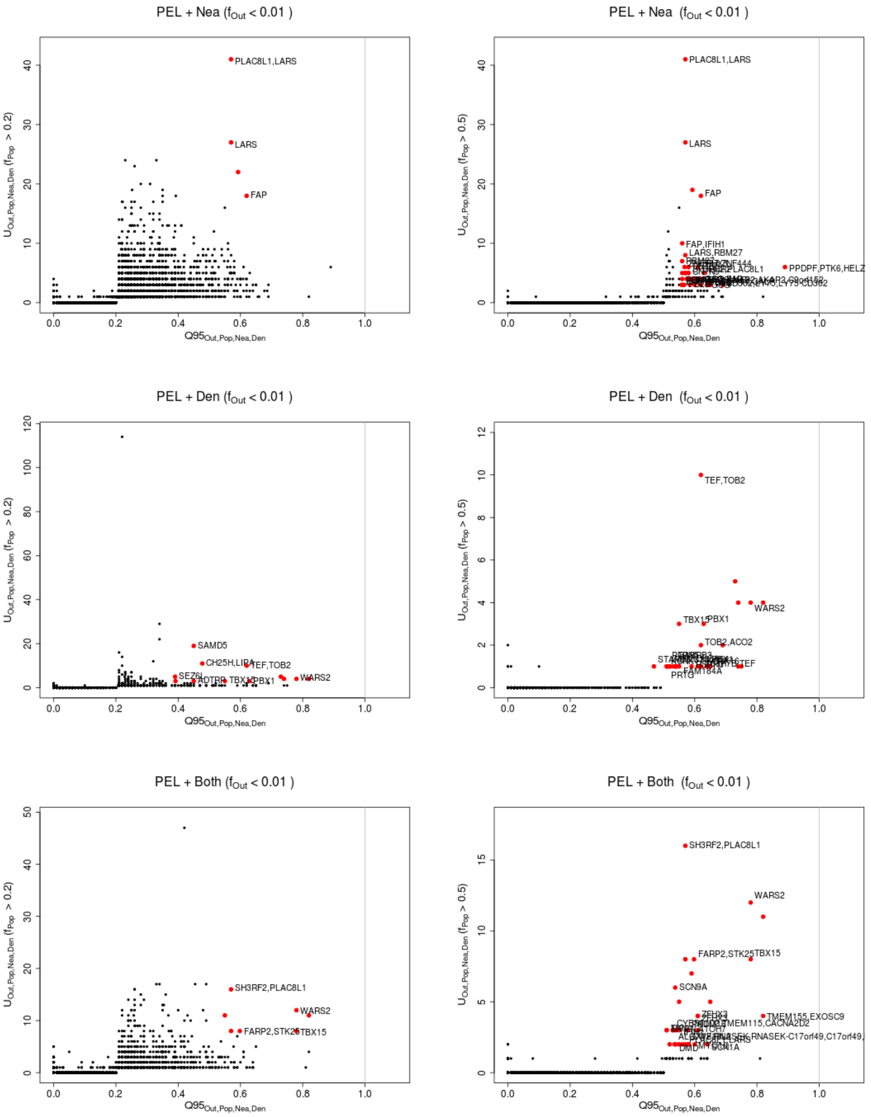


Figure 3.49. Uniquely shared archaic alleles in a Peruvian (PEL) panel. Joint distribution of $Q95_{AFR,PEL,Nea,Den}(1\%, y, z)$ and $U_{AFR,PEL,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

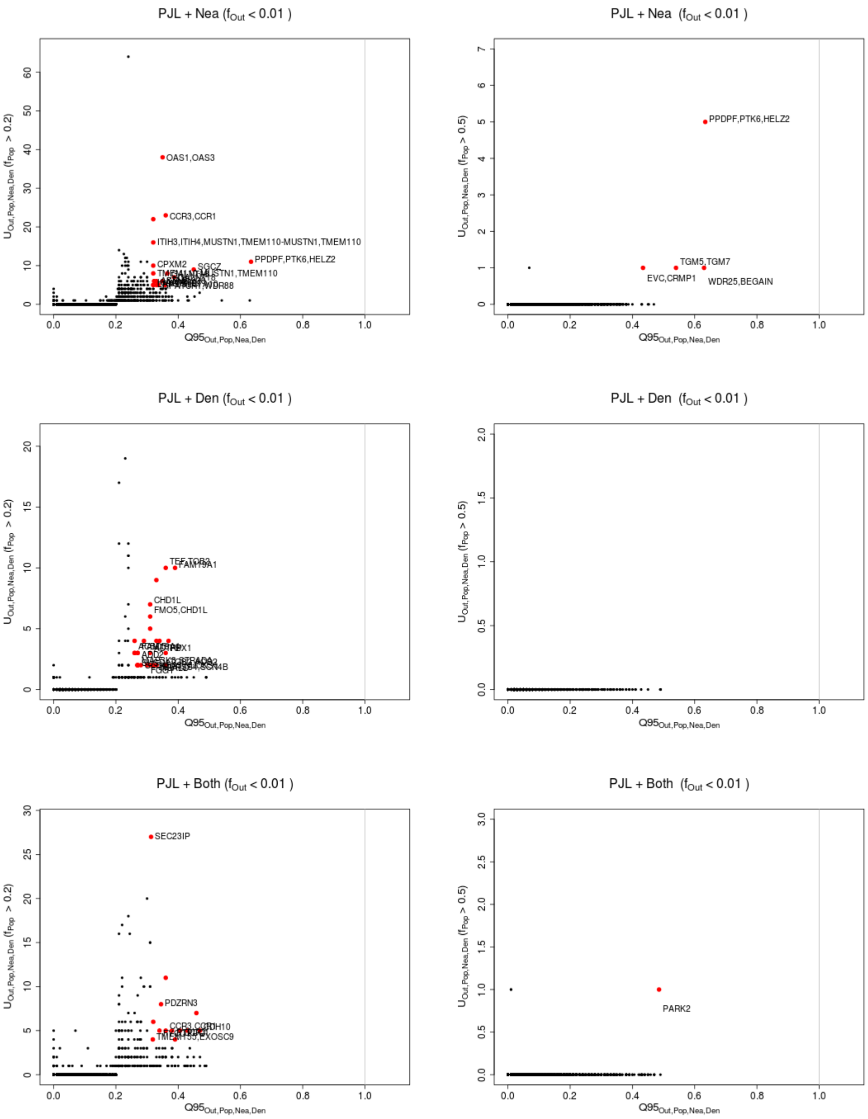


Figure 3.50. Uniquely shared archaic alleles in a Punjabi (PJL) panel. Joint distribution of $Q95_{AFR,PJL,Nea,Den}(1\%, y, z)$ and $U_{AFR,PJL,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

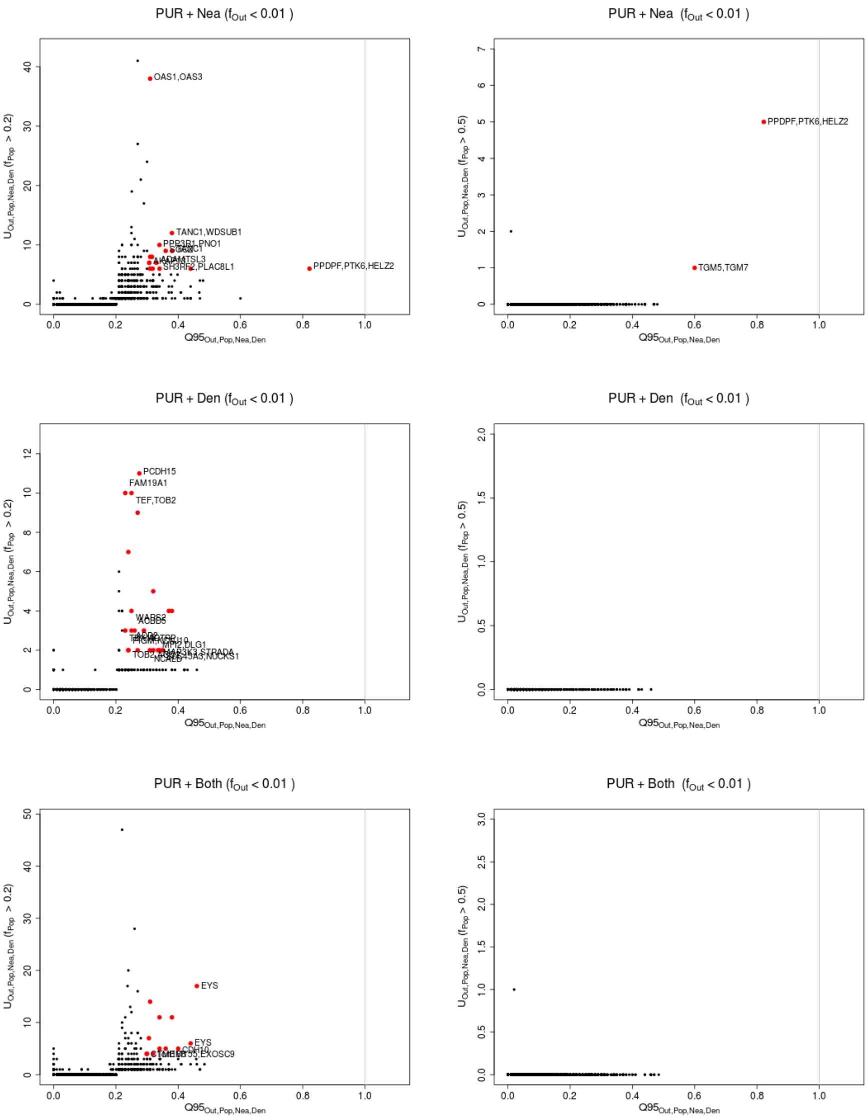


Figure 3.51. Uniquely shared archaic alleles in a Puerto Rican (PUR) panel. Joint distribution of $Q95_{AFR,PUR,Nea,Den}(1\%, y, z)$ and $U_{AFR,PUR,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

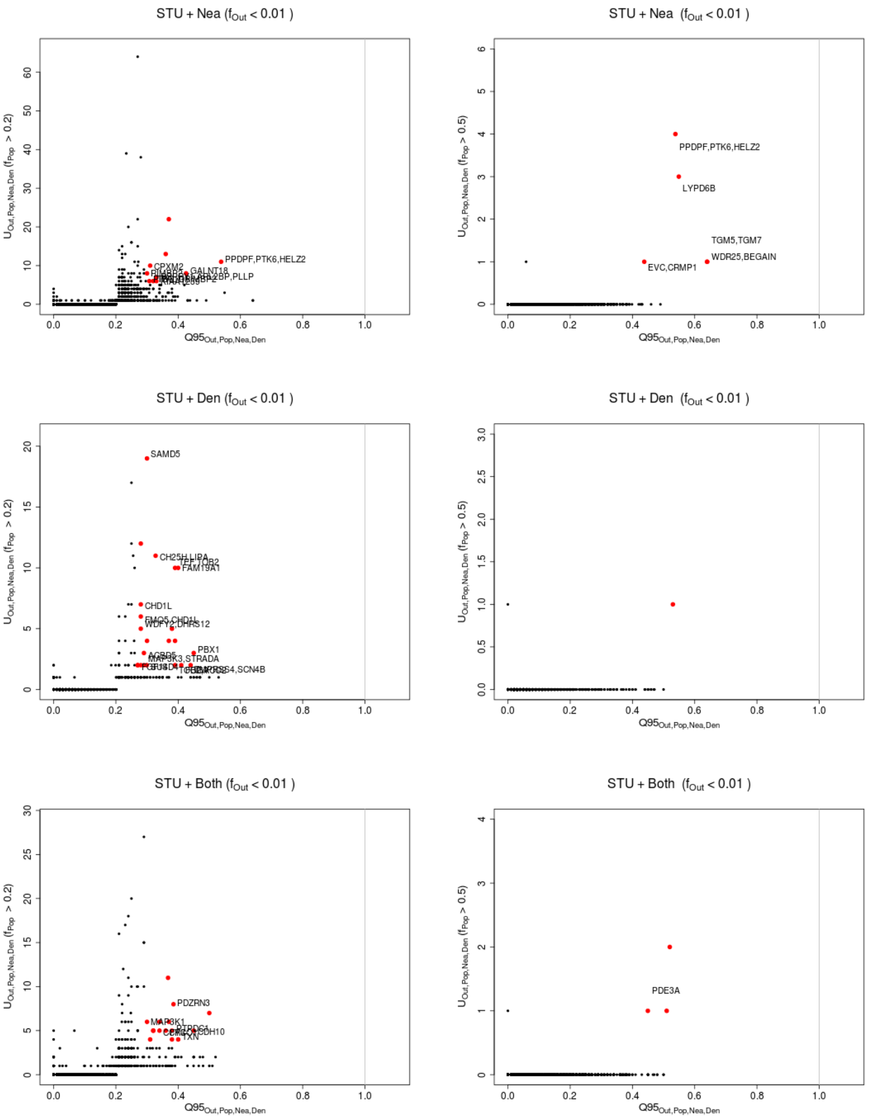


Figure 3.52. Uniquely shared archaic alleles in a Sri Lankan Tamil (STU) panel. Joint distribution of $Q95_{AFR,STU,Nea,Den}(1\%, y, z)$ and $U_{AFR,STU,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

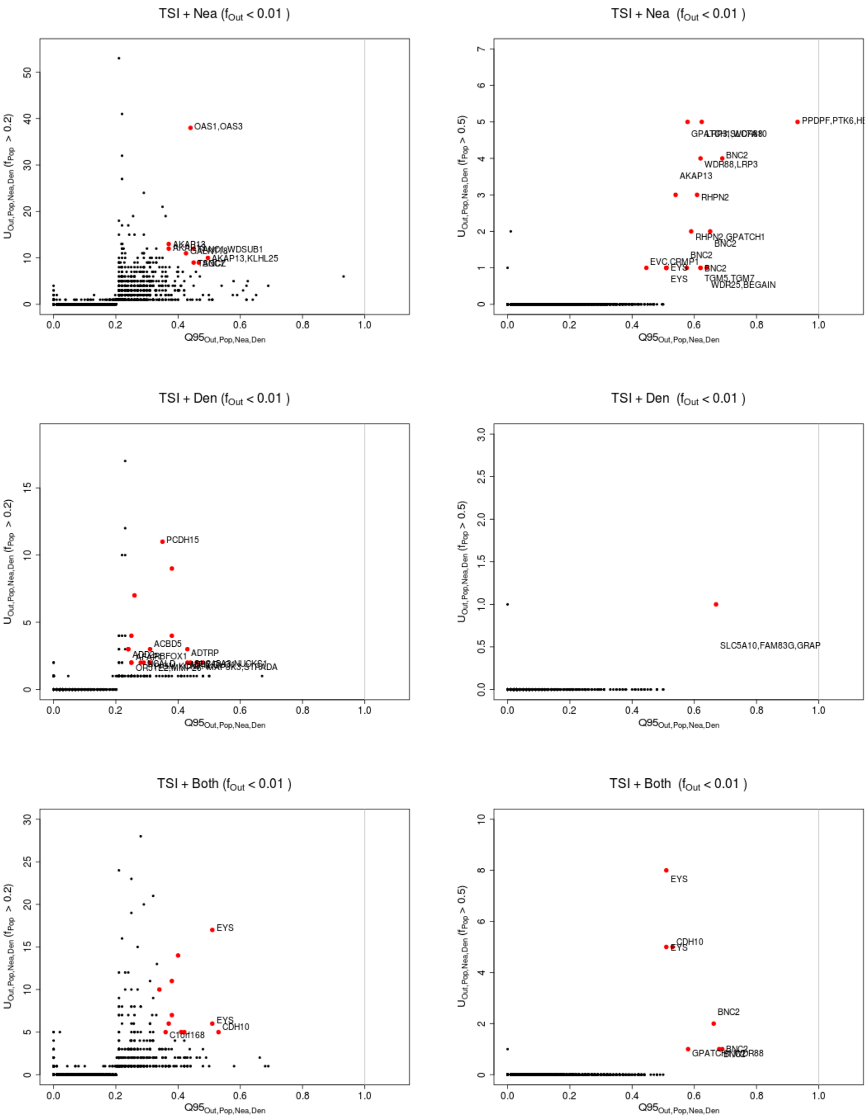


Figure 3.53. Uniquely shared archaic alleles in a Toscani (TSI) panel. Joint distribution of $Q95_{AFR,TSI,Nea,Den}(1\%, y, z)$ and $U_{AFR,TSI,Nea,Den}$(1%,x,y,z), for 40kb non-overlapping regions along the genome, using two choices of x (20% in left column panels,50% in right column panels). Red dots refer to regions that are in the 99.9% quantiles for both statistics. Neanderthal-specific shared alleles are displayed in the top panels, Denisovan-specific shared alleles are displayed in the middle-row panels, and alleles shared with both archaic human genome are displayed in the bottom panels.

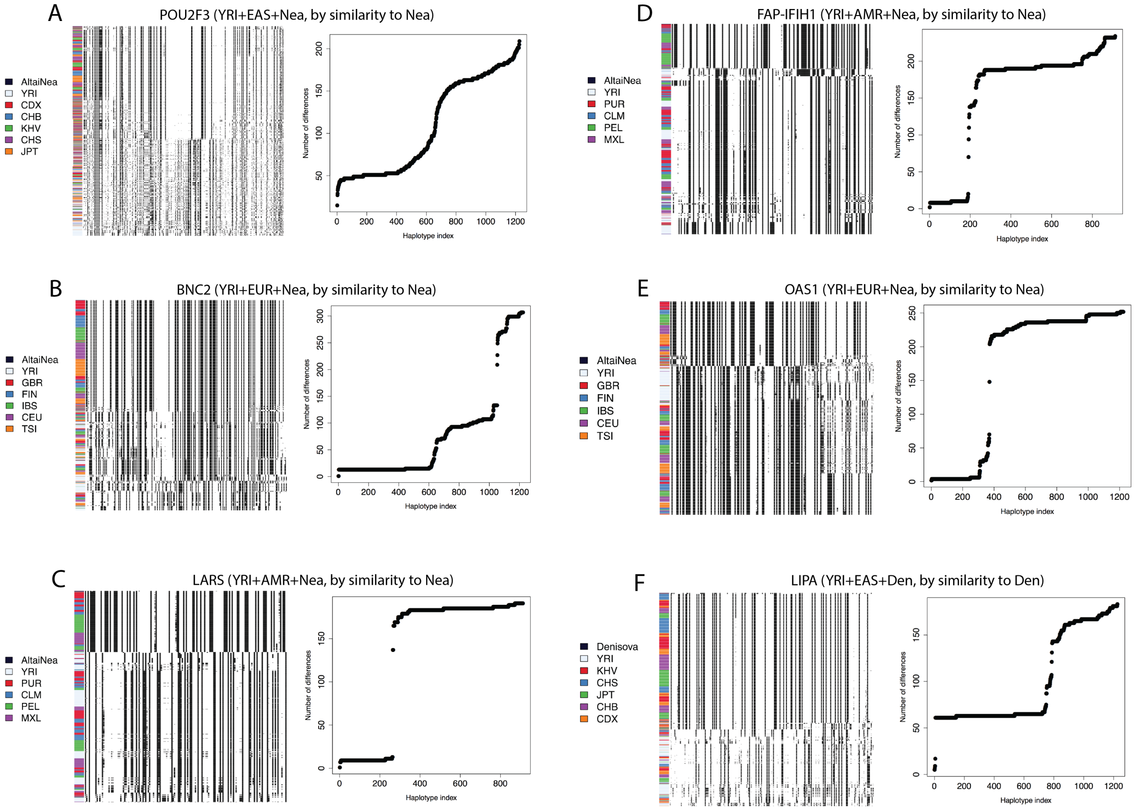


Figure 3.54. We explored the haplotype structure of six candidate regions with strong evidence for AI. For each region, we plotted the haplotypes of particular human populations ordered by similarity to the archaic human genome with the larger number of uniquely shared sites. We also plotted the number of differences to the closest archaic haplotype for each human haplotype and sorted them by decreasing similarity from left to right. *POU2F3*: chr11:120120001-120200000. *BNC2*: chr9:16720001-16760000. *LARS*: chr5:145480001-145520000. *FAP/IFIH1*: chr2:163040001-163120000. *OAS1*: chr12:113360001-113400000. *LIPA*: chr10:90920001-90980000.

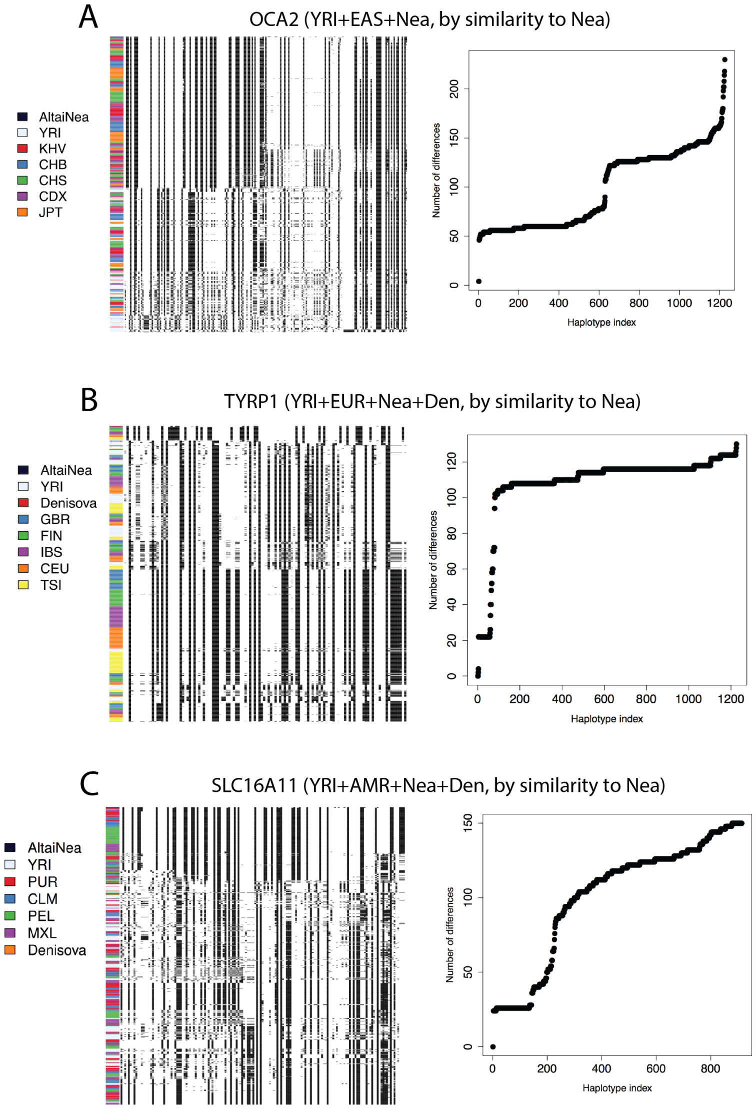


Figure 3.55. We explored the haplotype structure of *OCA2*, *TYRP1* and *SLC16A11*. We plotted the haplotypes of particular human populations ordered by similarity to the archaic human genome with the larger number of uniquely shared sites. We also plotted the number of differences to the closest archaic haplotype for each human haplotype and sorted them by decreasing similarity from left to right. OCA2: chr15:28160001-28200000. TYRP1: chr9:12680001-12720000.SLC16A11: chr17:6880001-6960000.

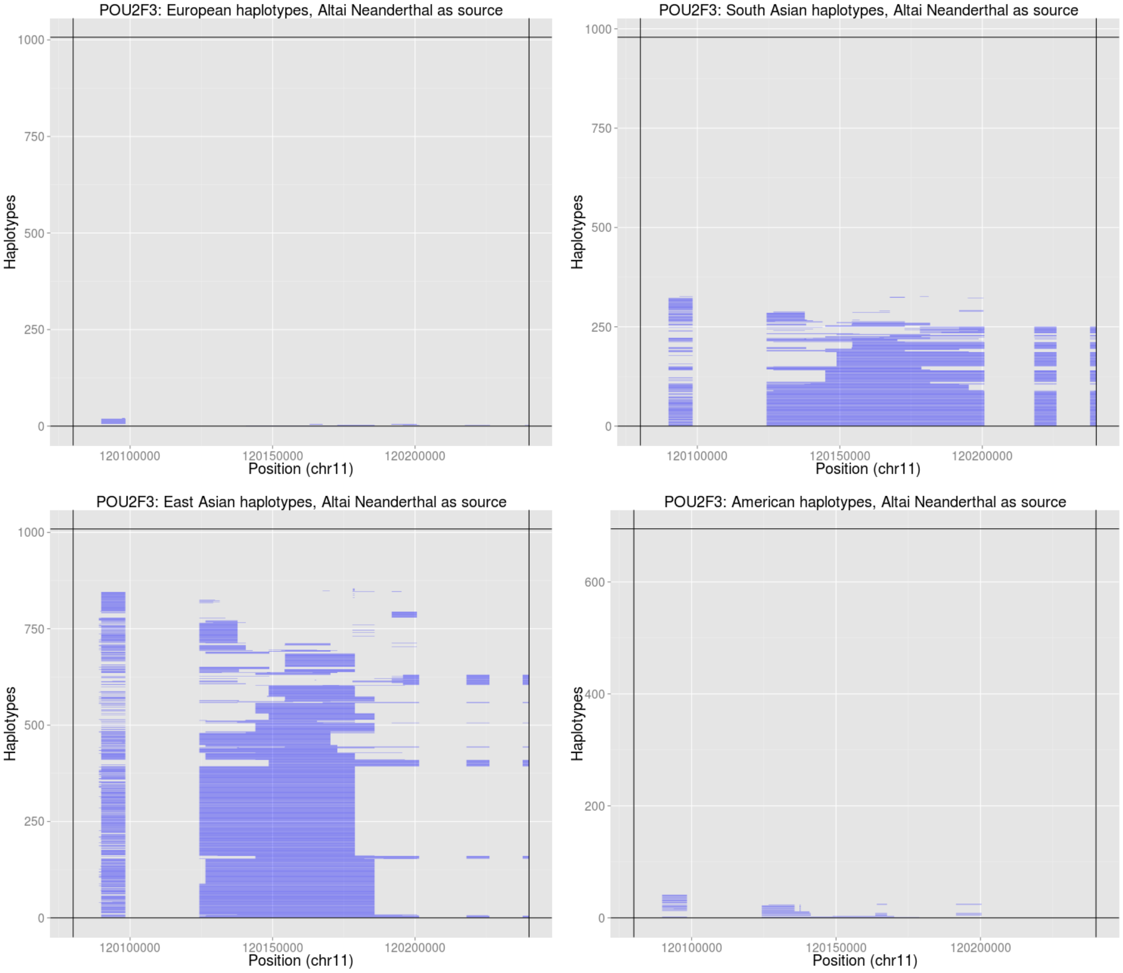


Figure 3.56. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *POU2F3* region, using the Altai Neanderthal genome as the archaic source.

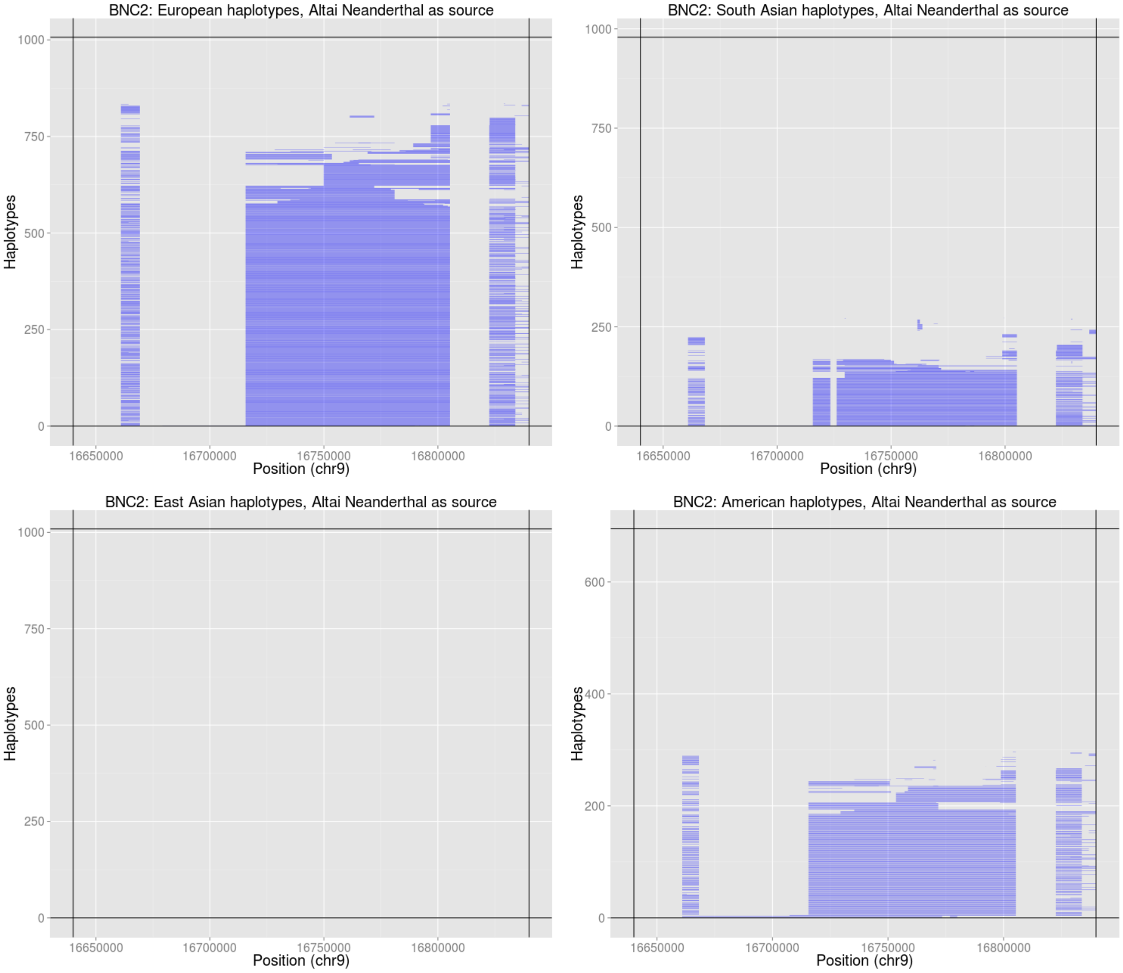


Figure 3.57. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *BNC2* region, using the Altai Neanderthal genome as the archaic source.

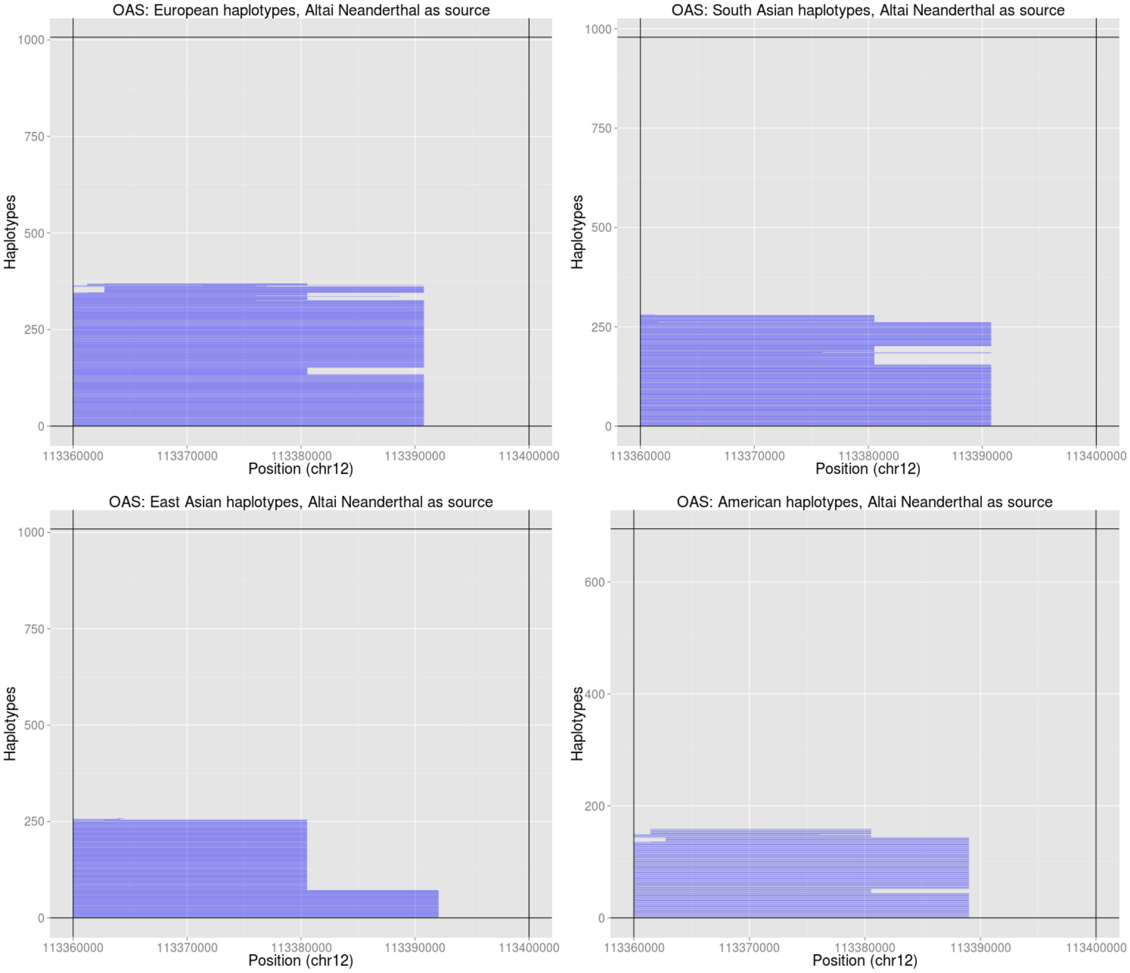


Figure 3.58. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *OAS* region, using the Altai Neanderthal genome as the archaic source.

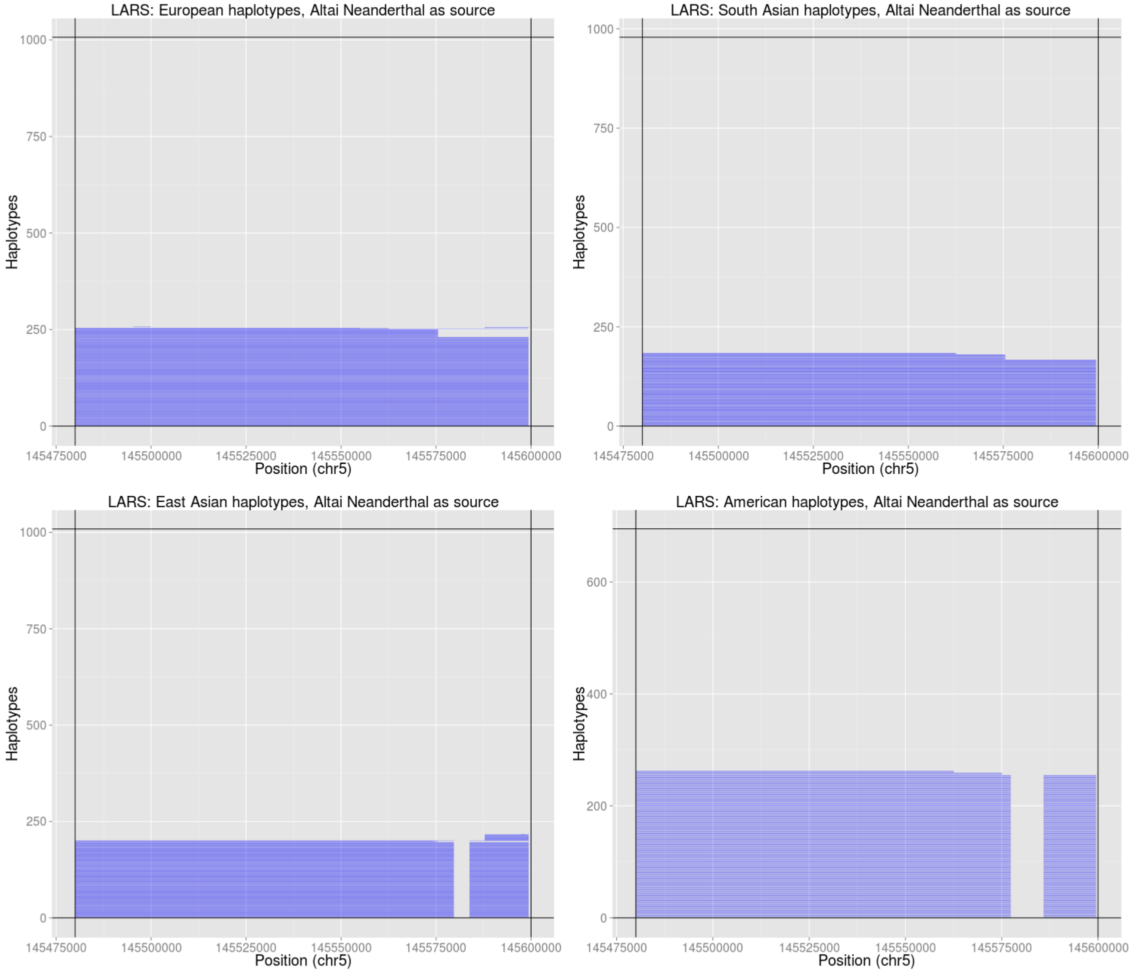


Figure 3.59. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *LARS* region, using the Altai Neanderthal genome as the archaic source.

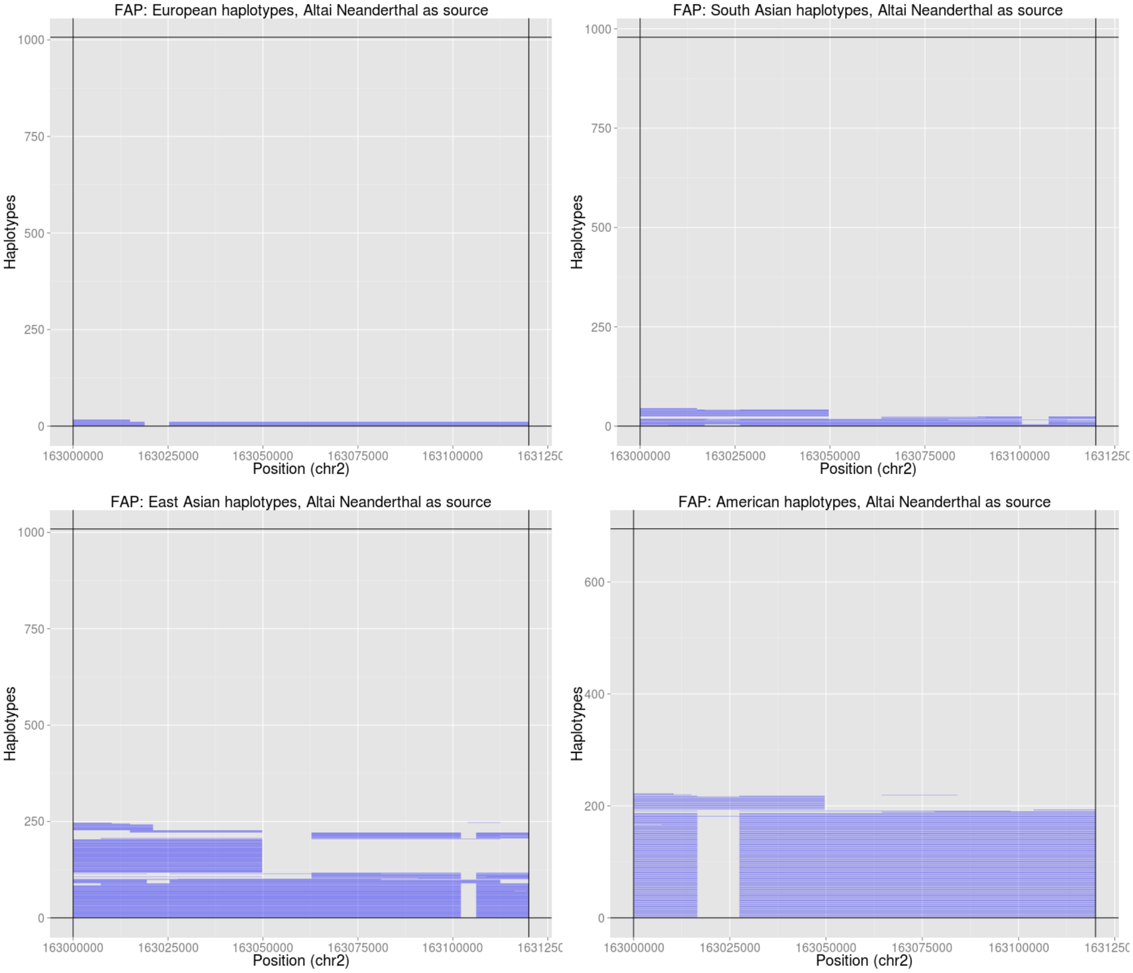


Figure 3.60. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *FAP/IFIH1* region, using the Altai Neanderthal as the archaic source.

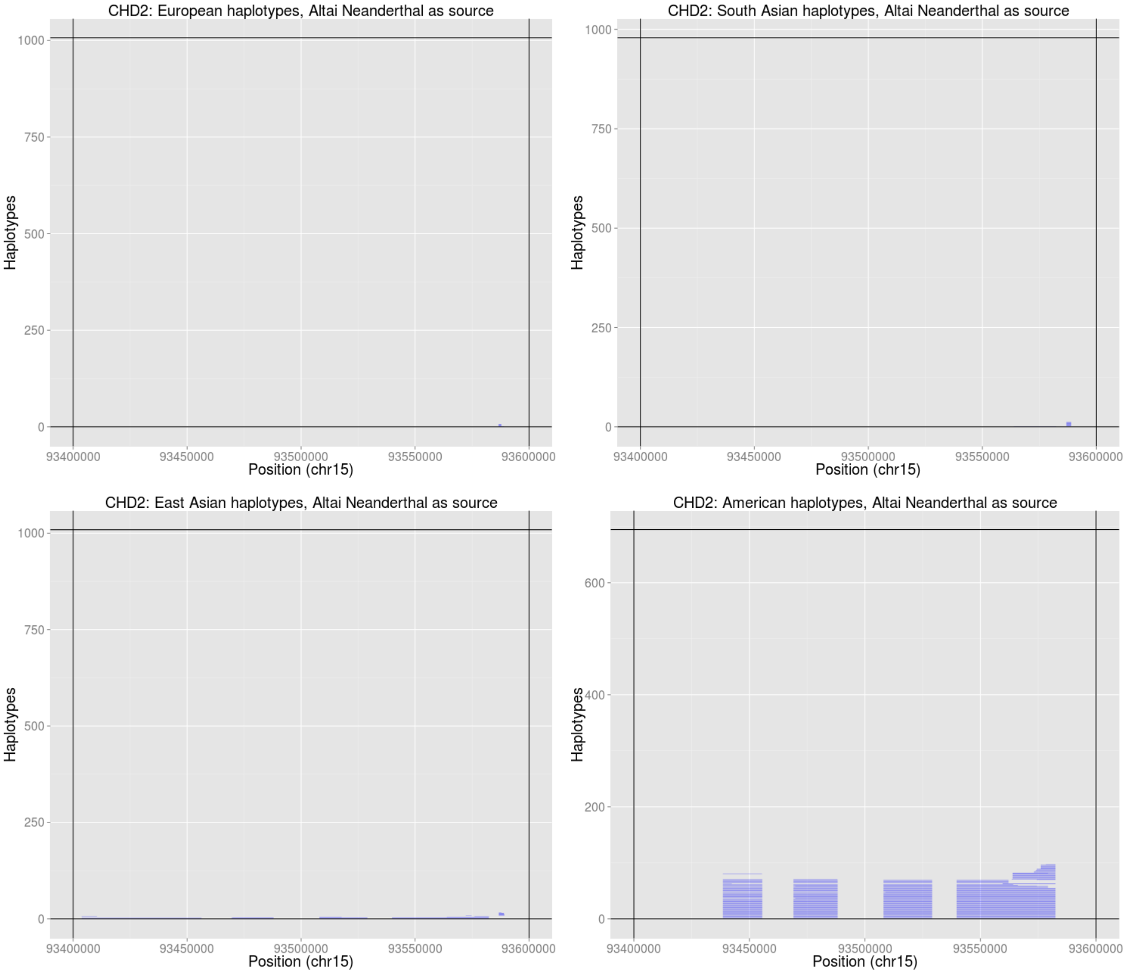


Figure 3.61. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *CHD2* region, using the Altai Neanderthal genome as the archaic source.

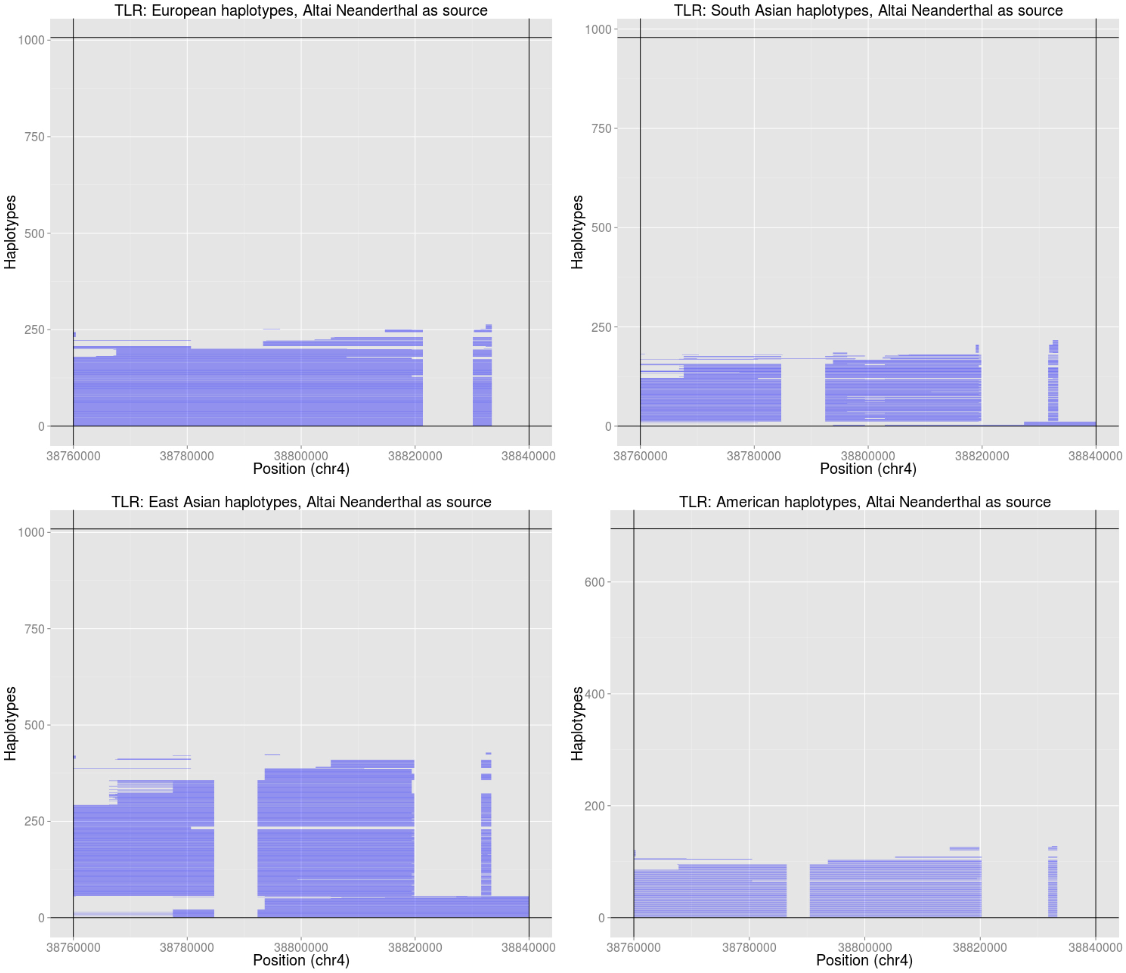


Figure 3.62. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *TLR1-6* region, using the Altai Neanderthal genome as the archaic source.

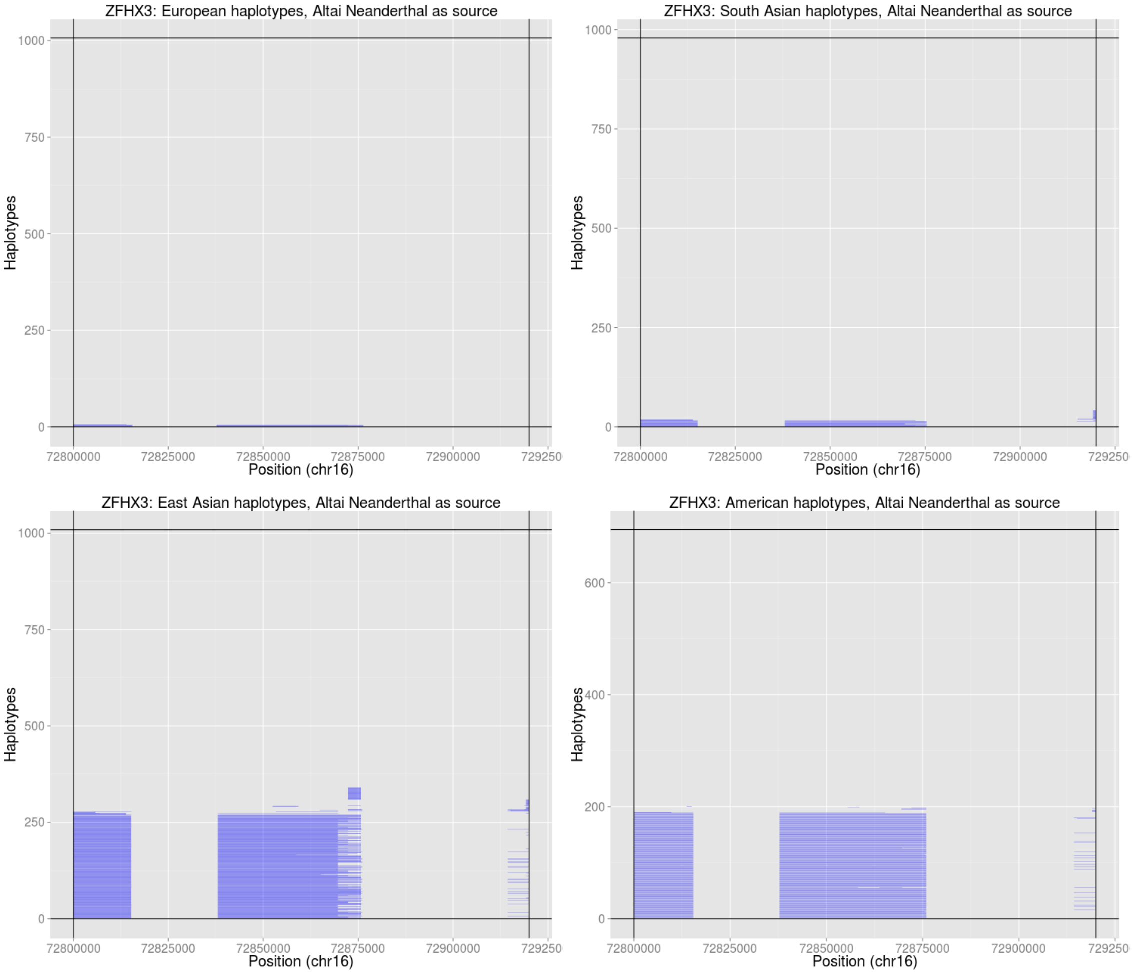


Figure 3.63. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *ZFHX3* region, using the Altai Neanderthal genome as the archaic source.

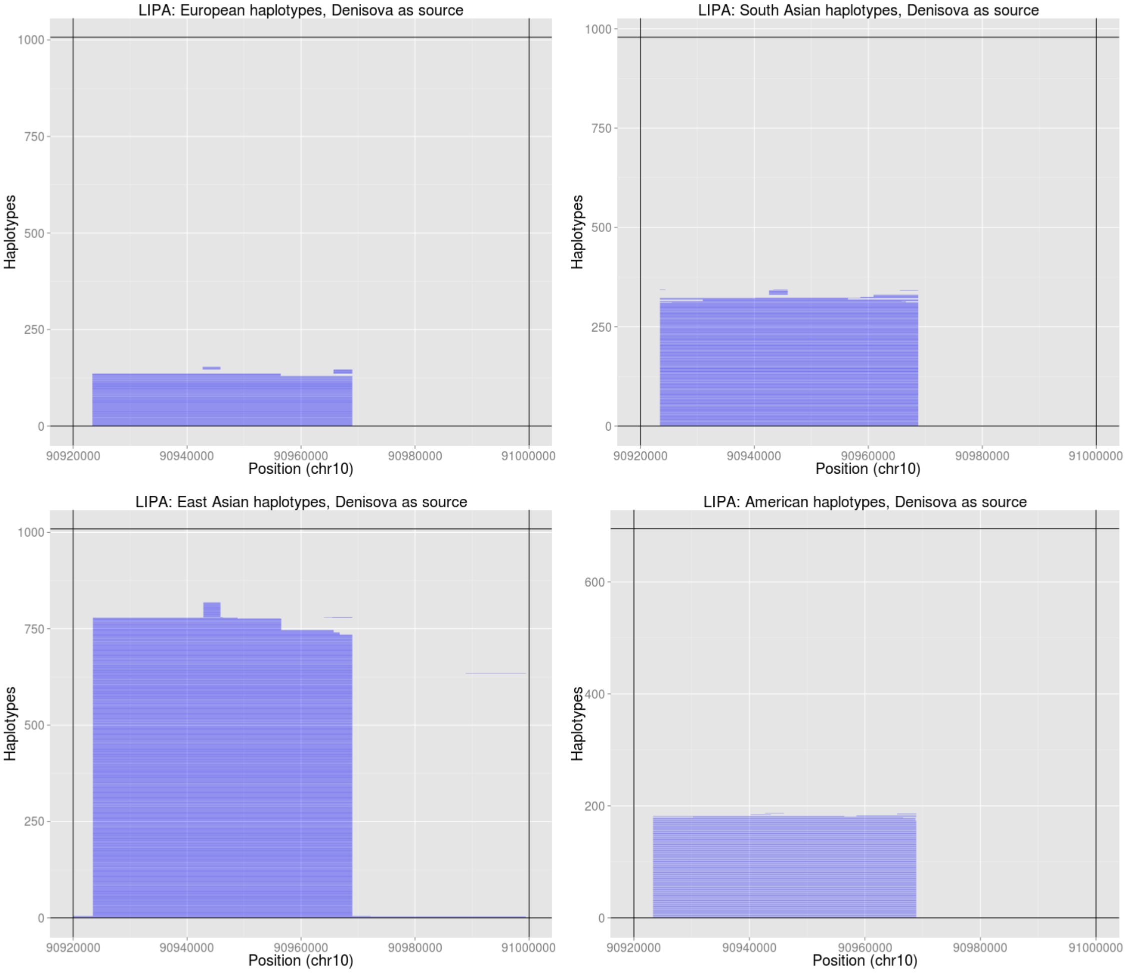


Figure 3.64. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *LIPA* region, using the Denisova genome as the archaic source.

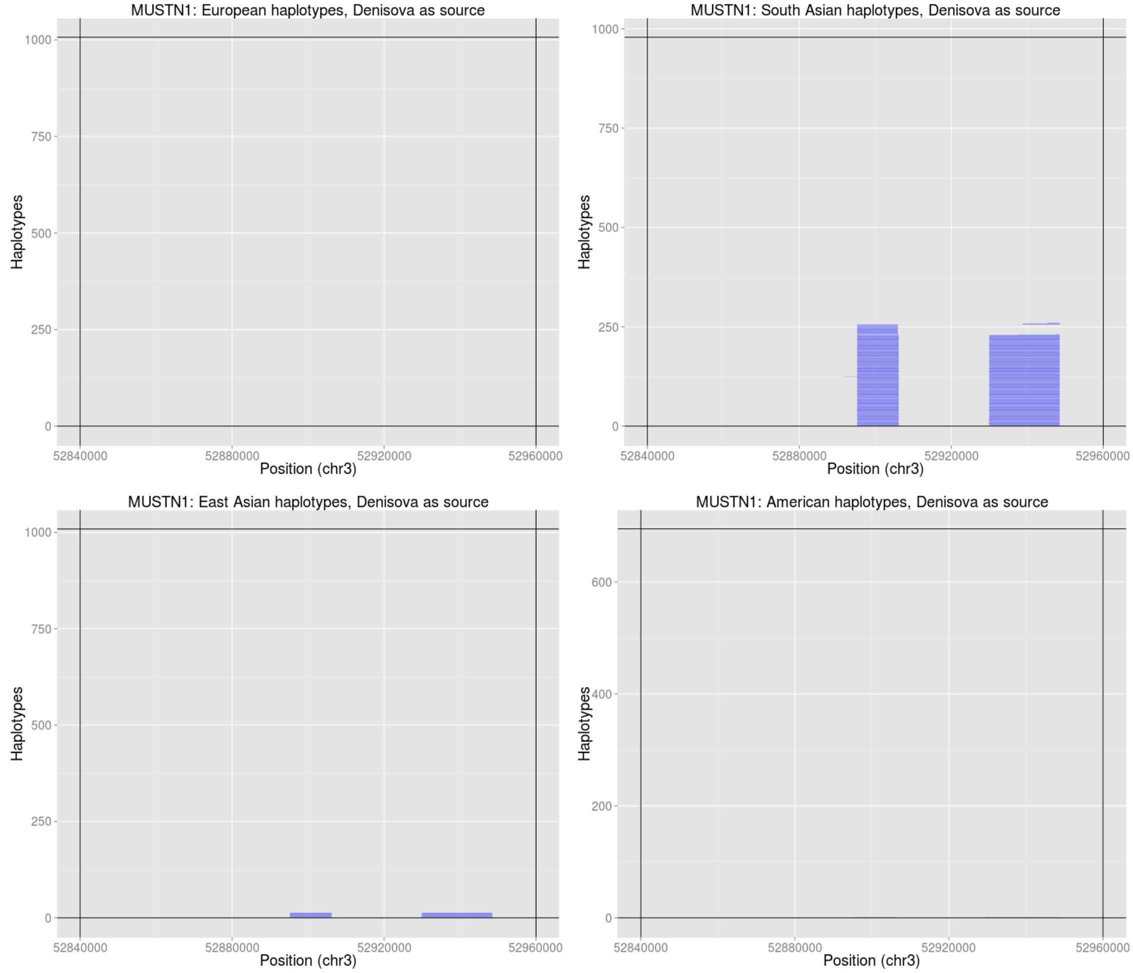


Figure 3.65. Introgressed tracks inferred in the four Non-African 1000 Genomes continental panels by an HMM [23] in the *MUSTN1* region, using the Denisova genome as the archaic source.

4. Testing for ancient selection using cross-population allele frequency differentiation

Fernando Racimo

4.1 Abstract

A powerful way to detect selection in a population is by modeling local allele frequency changes in a particular region of the genome under scenarios of selection and neutrality, and finding which model is most compatible with the data. [147] developed a composite likelihood method called XP-CLR that uses an outgroup population to detect departures from neutrality which could be compatible with hard or soft sweeps, at linked sites near a beneficial allele. However, this method is most sensitive to recent selection and may miss selective events that happened a long time ago. To overcome this, we developed an extension of XP-CLR that jointly models the behavior of a selected allele in a three-population tree. Our method - called 3P-CLR - outperforms XP-CLR when testing for selection that occurred before two populations split from each other, and can distinguish between those events and events that occurred specifically in each of the populations after the split. We applied our new test to population genomic data from the 1000 Genomes Project, to search for selective sweeps that occurred before the split of Yoruba and Eurasians, but after their split from Neanderthals, and that could have led to the spread of modern-human-specific phenotypes. We also searched for sweep events that occurred in East Asians, Europeans and the ancestors of both populations, after their split from Yoruba. In both cases, we are able to confirm a number of regions identified by previous methods, and find several new candidates for selection in recent and ancient times. For some of these, we also find suggestive functional mutations that may have driven the selective events.

4.2 Introduction

Genetic hitchhiking will distort allele frequency patterns at regions of the genome linked to a beneficial allele that is rising in frequency [148]. This is known as a selective sweep. If the sweep is restricted to a particular population and does not affect other closely related populations, one can detect such an event by looking for extreme patterns of localized population differentation, like high values of F_{st} at a specific locus [149]. This and other related statistics have been used to scan the genomes of present-day humans from different populations, so as to detect signals of recent positive selection [150, 151, 152, 32].

Once it became possible to sequence entire genomes of archaic humans (like Neanderthals) [8, 2, 16], researchers also began to search for selective sweeps that occurred in the ancestral population of all present-day humans. For example, [8] searched for genomic regions with a depletion of derived alleles in a low-coverage Neanderthal genome, relative to what would be expected given the derived allele frequency in present-day humans. This is a pattern that would be consistent with a sweep in present-day humans. Later on, [16] developed a hidden Markov model (HMM) that could identify regions where Neanderthals fall outside of all present-day human variation (also called "external regions"), and are therefore likely to have been affected by ancient sweeps in early modern humans. They applied their method to a high-coverage Neanderthal genome. Then, they ranked these regions by their genetic length, to find segments that were extremely long, and therefore highly compatible with a selective sweep. Finally, [37] used summary statistics calculated in the neighborhood of sites that were ancestral in archaic humans but fixed derived in all or almost all present-day humans, to test if any of these sites could be compatible with a selective sweep model. While these methods harnessed different summaries of the patterns of differentiation left by sweeps, they did not attempt to explicitly model the process by which these patterns are generated over time.

[147] developed a method called XP-CLR, which is designed to test for selection in one population after its split from a second, outgroup, population t_{AB} generations ago. It does so by modeling the evolutionary trajectory of an allele under linked selection and under neutrality, and then comparing the likelihood of the data for each of the two models. The method detects local allele frequency differences that are compatible with the linked selection model [148], along windows of the genome.

XP-CLR is a powerful test for detecting selective events restricted to one population. However, it provides little information about when these events happened, as it models all sweeps as if they had immediately occurred in the present generation. Additionally, if one is interested in selective sweeps that took place before two populations a and b split from each other, one would have to run XP-CLR separately on each population, with a third outgroup population c that split from the ancestor of a and b t_{ABC} generations ago (with $t_{ABC} > t_{AB}$). Then, one would need to check that the signal of selection appears in both tests. This may miss important information about correlated allele frequency changes shared by a and b, but not by c, limiting the power to detect ancient events.

To overcome this, we developed an extension of XP-CLR that jointly models the behavior of an allele in all 3 populations, to detect selective events that occurred before or after the closest two populations split from each other. Below we briefly review the modeling framework of XP-CLR and describe our new test, which we call 3P-CLR. In the Results, we show this method outperforms XP-CLR when testing for selection that occurred before the split of two populations, and can distinguish between those events and events that occurred after the split, unlike XP-CLR. We then apply the method to population genomic data from the 1000 Genomes Project [153], to search for selective sweep events that occurred before the split of Yoruba and Eurasians, but after their split from Neanderthals. We also use it to search for selective sweeps that occurred in the Eurasian ancestral population, and to distinguish those from events that occurred specifically in East Asians or specifically in Europeans.

4.3 Materials and Methods

XP-CLR

First, we review the procedure used by XP-CLR to model the evolution of allele frequency changes of two populations a and b that split from each other t_{AB} generations ago (Figure 4.1.A). For neutral SNPs, [147] use an approximation to the Wright-Fisher diffusion dynamics [154]. Namely, the frequency of a SNP in a population a (p_A) in the present is treated as a random variable governed by a normal distribution with mean equal to the frequency in the ancestral population (β) and variance proportional to the drift time ω from the ancestral to the present population:

$$p_A|\beta \sim N(\beta, \omega\beta(1-\beta)) \tag{4.1}$$

where $\omega = t_{AB}/(2N_e)$ and N_e is the effective size of population A.

This is a Brownian motion approximation to the Wright-Fisher model, as the drift increment to variance is constant across generations. If a SNP is segregating in both populations - i.e. has not hit the boundaries of fixation or extinction - this process is time-reversible. Thus, one can model the frequency of the SNP in population a with a normal distribution having mean equal to the frequency in population b and variance proportional to the sum of the drift time (ω) between a and the ancestral population, and the drift time between b and the ancestral population (ψ):

$$p_A|p_B \sim N(p_B, (\omega+\psi)p_B(1-p_B)) \tag{4.2}$$

For SNPs that are linked to a beneficial allele that has produced a sweep in population a only, [147] model the allele as evolving neutrally until the present and then apply a transformation to the normal distribution that depends on the distance to the selected allele r and the strength of selection s [155, 156]. Let $c = 1 - q_0^{r/s}$ where q_0 is the frequency of the beneficial allele in population a before the sweep begins. The frequency of a neutral allele is expected to increase from p to $1 - c + cp$ if the allele is linked to the beneficial allele, and this occurs with probability equal to the frequency of the neutral allele (p) before the sweep begins. Otherwise, the frequency of the neutral allele is expected to decrease from p to cp. This leads to the following transformation of the normal distribution:

$$f(p_A|p_B, r, s, \omega, \psi) = \frac{1}{\sqrt{2\pi}\sigma}\frac{p_A + c - 1}{c^2}e^{-\frac{(p_A+c-1-cp_B)^2}{2c^2\sigma^2}}I_{[1-c,1]}(p_A) + \frac{1}{\sqrt{2\pi}\sigma}\frac{c - p_A}{c^2}e^{-\frac{(p_A-cp_B)^2}{2c^2\sigma^2}}I_{[0,c]}(p_A) \quad (4.3)$$

where $\sigma^2 = (\omega + \psi)p_B(1 - p_B)$ and $I_{[x,y]}(z)$ is 1 on the interval $[x, y]$ and 0 otherwise.

For $s \rightarrow 0$ or $r >> s$, this distribution converges to the neutral case. Let $\mathbf{v}$ be the vector of all drift times that are relevant to the scenario we are studying. In this case, it will be equal to (ω, ψ) but in more complex cases below, it may include additional drift times. Let $\mathbf{r}$ be the vector of recombination fractions between the beneficial alleles and each of the SNPs within a window of arbitrary size. We can then calculate the product of likelihoods over all k SNPs in that window for either the neutral or the linked selection model, after binomial sampling of alleles from the population frequency, and conditioning on the event that the allele is segregating in the population:

$$CL_{XP-CLR}(\mathbf{r}, \mathbf{v}, s) = \prod_{j=1}^{k} \frac{\int_0^1 f(p_A^j|p_B^j, \mathbf{v}, s, r^j)\binom{n}{m_j}(p_A^j)^{m_j}(1 - p_A^j)^{n-m_j}dp_A^j}{\int_0^1 f(p_A^j|p_B^j, \mathbf{v}, s, r^j)dp_A^j} \quad (4.4)$$

This is a composite likelihood [157, 158], because we are ignoring the correlation in frequencies produced by linkage among SNPs that is not strictly due to proximity to the beneficial SNP. We note that the denominator in the above equation is not explicitly stated in [147] for ease of notation, but appears in the published online implementation of the method.

Finally, we obtain a composite likelihood ratio statistic S_{XP-CLR} of the hypothesis of linked selection over the hypothesis of neutrality:

$$S_{XP-CLR} = 2[sup_{\mathbf{r},\mathbf{v},s}log(CL_{XP-CLR}(\mathbf{r}, \mathbf{v}, s)) - sup_{\mathbf{v}}log(CL_{XP-CLR}(\mathbf{r}, \mathbf{v}, s = 0))] \quad (4.5)$$

For ease of computation, [147] assume that $\mathbf{r}$ is given (via a recombination map) instead of maximizing the likelihood with respect to it, and we will do so too. Furthermore, they empirically estimate $\mathbf{v}$ using F_2 statistics [159] calculated over the whole genome, and assume selection is not strong or frequent enough to affect their genome-wide values. Therefore, the likelihoods in the above equation are only maximized with respect to the selection coefficient, using a grid of coefficients on a logarithmic scale.

3P-CLR

We are interested in the case where a selective event occurred more anciently than the split of two populations (a and b) from each other, but more recently than their split from a third population c (Figure 4.1.B). We begin by modeling p_A and p_B as evolving from an unknown common ancestral frequency β:

$$p_A|\beta, \omega \sim N(\beta, \omega\beta(1-\beta)) \tag{4.6}$$

$$p_B|\beta, \psi \sim N(\beta, \psi\beta(1-\beta)) \tag{4.7}$$

Let χ be the drift time separating the most recent common ancestor of a and b from the most recent common ancestor of a, b and c. Additionally, let ν be the drift time separating population c in the present from the most recent common ancestor of a, b and c. Given these parameters, we can treat β as an additional random variable that either evolves neutrally or is linked to a selected allele that swept immediately more anciently than the split of a and b. In both cases, the distribution of β will depend on the frequency of the allele in population c (p_C) in the present. In the neutral case:

$$f_{neut}(\beta|p_C, \nu, \chi) = N(p_C, (\nu+\chi)p_C(1-p_C)) \tag{4.8}$$

In the linked selection case:

$$f_{sel}(\beta|p_C, \nu, \chi, r, s) = \frac{1}{\sqrt{2\pi}\kappa}\frac{\beta+c-1}{c^2}e^{-\frac{(\beta+c-1-cp_C)^2}{2c^2\kappa^2}}I_{[1-c,1]}(\beta) + \frac{1}{\sqrt{2\pi}\kappa}\frac{c-\beta}{c^2}e^{-\frac{(\beta-cp_C)^2}{2c^2\kappa^2}}I_{[0,c]}(\beta) \tag{4.9}$$

where $\kappa^2 = (\nu+\chi)p_C(1-p_C)$

The frequencies in a and b given the frequency in c can be obtained by integrating β out. This leads to a density function that models selection in the ancestral population of a and b.

$$f(p_A, p_B|p_C, \mathbf{v}, r, s) = \int_0^1 f_{neut}(p_A|\beta, \omega)f_{neut}(p_B|\beta, \psi)f_{sel}(\beta|p_C, \nu, \chi, r, s)d\beta \tag{4.10}$$

Additionally, formula 4.10 can be modified to test for selection that occurred specifically in one of the terminal branches that lead to a or b (Figures 4.1.C and 4.1.D), rather than in the ancestral population of a and b. For example, the density of frequencies for a scenario of selection in the branch leading to a can be written as:

$$f(p_A, p_B|p_C, \mathbf{v}, r, s) = \int_0^1 f_{sel}(p_A|\beta, \omega, r, s)f_{neut}(p_B|\beta, \psi)f_{neut}(\beta|p_C, \nu, \chi)d\beta \tag{4.11}$$

We will henceforth refer to the version of 3P-CLR that is tailored to detect selection in the internal branch that is ancestral to a and b as 3P-CLR(Int). In turn, the versions of 3P-CLR that are designed to detect selection in each of the daughter populations a and b will be designated as 3P-CLR(A) and 3P-CLR(B), respectively.

We can now calculate the probability density of specific allele frequencies in populations a and b, given that we observe m_C derived alleles in a sample of size n_C from population c:

$$f(p_A, p_B | m_C, \mathbf{v}, r, s) = \int_0^1 f(p_A, p_B | p_C, \mathbf{v}, r, s) f(p_C | m_C) dp_C \tag{4.12}$$

and

$$f(p_C | m_C) = \frac{1}{B(m_C, n_C - m_C + 1)} p_C^{m_C - 1} (1 - p_C)^{n_C - m_C} \tag{4.13}$$

where B(x,y) is the Beta function. We note that formula 4.13 assumes that the unconditioned density function for the population derived allele frequency $f(p_C)$ comes from the neutral infinite-sites model at equilibrium and is therefore equal to the product of a constant and $1/p_C$ [160].

Conditioning on the event that the site is segregating in the population, we can then calculate the probability of observing m_A and m_B derived alleles in a sample of size n_A from population a and a sample of size n_B from population b, respectively, given that we observe m_C derived alleles in a sample of size n_C from population c, using binomial sampling:

$$P(m_A, m_B | m_C, \mathbf{v}, r, s) = \frac{\int_0^1 \int_0^1 P(m_A | p_A) P(m_B | p_B) f(p_A, p_B | m_C, \mathbf{v}, r, s) dp_A dp_B}{\int_0^1 \int_0^1 f(p_A, p_B | m_C, \mathbf{v}, r, s) dp_A dp_B} \tag{4.14}$$

where

$$P(m_A | p_A) = \binom{n_A}{m_A} p_A^{m_A} (1 - p_A)^{n_A - m_A} \tag{4.15}$$

and

$$P(m_B | p_B) = \binom{n_B}{m_B} p_B^{m_B} (1 - p_B)^{n_B - m_B} \tag{4.16}$$

This allows us to calculate a composite likelihood of the derived allele counts in a and b given the derived allele counts in c:

$$CL_{3P-CLR}(\mathbf{r}, \mathbf{v}, s) = \prod_{j=1}^{k} P(m_A^j, m_B^j | m_C^j, \mathbf{v}, r^j, s) \tag{4.17}$$

As before, we can use this composite likelihood to produce a composite likelihood ratio statistic that can be calculated over regions of the genome to test the hypothesis of linked selection centered on a particular locus against the hypothesis of neutrality. Due to computational costs in numerical integration, we skip the sampling step for population c (formula 4.13) in our implementation of 3P-CLR. In other words, we assume $p_C = m_C/n_C$, but this is also assumed in XP-CLR when computing its corresponding outgroup frequency. To perform the numerical integrations, we used the package Cubature (v.1.0.2). We implemented our method in a freely available C++ program that can be downloaded from here:

`https://github.com/ferracimo`

The program requires the neutral drift parameters α, β and $(\nu+\chi)$ to be specified as input. These can be obtained using F_3 statistics [161, 159], which have previously been implemented in programs like MixMapper [58]. For example, α can be obtained via $F_3(A; B, C)$, while $(\nu+\chi)$ can be obtained via $F_3(C; A, B)$. When computing F_3 statistics, we use only sites where population C is polymorphic, and so we correct for this ascertainment in the calculation. Another way of calculating these drift times is via $\partial a \partial i$ [49]. Focusing on two populations at a time, we can fix one population's size and allow the split time and the other population's size to be estimated by the program, in this case using all polymorphic sites, regardless of which population they are segregating in. We then obtain the two drift times by scaling the inferred split time by the two different population sizes. We provide scripts in our github page for the user to obtain these drift parameters using both of the above ways.

4.4 Results

Simulations

We generated simulations in SLiM [73] to test the performance of XP-CLR and 3P-CLR in a three-population scenario. We first focused on the performance of 3P-CLR(Int) in detecting ancient selective events that occurred in the ancestral branch of two sister populations. We assumed that the population history had been correctly estimated (i.e. the drift parameters and population topology were known). First, we simulated scenarios in which a beneficial mutation arose in the ancestor of populations a and b, before their split from each other but after their split from c (Table 4.1). Although both XP-CLR and 3P-CLR are sensitive to partial or soft sweeps (as they do not rely on extended patterns of haplotype homozygosity [147]), we required the beneficial allele to have fixed before the split (at time t_{ab}) to ensure that the allele had not been lost by then, and also to ensure that the sweep was restricted to the internal branch of the tree. We fixed the effective size of all three populations at $N_e = 10,000$. Each simulation consisted in a 5 cM region and the beneficial mutation occurred in the center of this region. The mutation rate was set at $2.5 * 10^{-8}$ per generation and the recombination rate between adjacent nucleotides was set at 10^{-8} per generation.

To make a fair comparison to 3P-CLR(Int), and given that XP-CLR is a two-population test, we applied XP-CLR in two ways. First, we pretended population b was not sampled, and so the "test" panel consisted of individuals from a only, while the "outgroup" consisted of individuals from c. In the second implementation (which we call "XP-CLR-avg"), we used the same outgroup panel, but pooled the individulas from a and b into a single panel, and this pooled panel was the "test". The window size was set at 0.5 cM and the number of SNPs sampled between each window's central SNP was set at 600 (this number is large because it includes SNPs that are not segregating in the outgroup, which are later discarded). To speed up computation, and because we are largely interested in comparing the relative performance of the three tests under different scenarios, we used only 20 randomly chosen SNPs per window in all tests. We note, however, that the performance of all of these tests can be improved by using more SNPs per window.

Figure 4.2 shows receiver operating characteristic (ROC) curves comparing the sensitivity and specificity of 3P-CLR(Int), 3P-CLR(A), XP-CLR and XP-CLR-avg in the first six demographic scenarios described in Table 4.1. Each ROC curve was made from 100 simulations under selection (with $s = 0.1$ for the central mutation) and 100 simulations under neutrality (with $s = 0$ and no fixation required). In each simulation, 100 haploid individuals (or 50 diploids) were sampled from population a, 100 individuals from population b and 100 individuals from the outgroup population c. For each simulation, we took the maximum value at a region in the neighborhood of the central mutation (+/- 0.5 cM) and used those values to compute ROC curves under the two models.

When the split times are recent or moderately ancient (models A to D), 3P-CLR(Int) outperforms the two versions of XP-CLR. Furthermore, 3P-CLR(A) is the test that is least sensitive to selection in the internal branch as it is only meant to detect selection in the terminal branch leading to population a. When the split times are very ancient (models E and F), none of the tests perform well. The root mean squared error (RMSE) of the genetic distance between the true selected site and the highest scored window is comparable across tests in all six scenarios (Figure 4.3). 3P-CLR(Int) is the best test at finding the true location of the selected site in almost all demographic scenarios. We observe that we lose almost all power if we simulate demographic scenarios where the population size is 10 times smaller ($N_e = 1,000$) (Figure 4.4). Additionally, we observe that the power and specificity of 3P-CLR decrease as the selection coefficient decreases (Figure 4.5).

We also simulated a situation in which only a few individuals are sequenced from the outgroup, while large numbers of sequences are available from the tests. Figures 4.6 and 4.7 show the ROC curves and RMSE plots, respectively, for a scenario in which 100 individuals were sampled from the test populations but only 10 individuals (5 diploids) were sampled from the outgroup. Unsurprisingly, all tests have less power to detect selection when the split times and the selection events are recent to moderately ancient (models A-D). Interestingly though, when the split times and the selective events are very ancient (models E-F), both 3P-CLR and XP-CLR perform better when using a small ougroup panel (Figure 4.6) than when using a large outgroup panel (Figure 4.2). This is due to the Brownian motion approximation that these methods utilize. Under the Wright-Fisher model, the drift increment at generation t is proportional to p(t)*(1-p(t)), where p(t) is the derived allele frequency. The derivative of this function gets smaller the closer p(t) is to 0.5 (and is exactly 0 at that point). Small outgroup panels serve to filter out loci with allele frequencies far from 0.5, and so small changes in allele frequency will not affect the drift increment much, making Brownian motion a good approximation to the Wright-Fisher model. Indeed, when running 3P-ClR(Int) in a demographic scenario with very ancient split times (Model E) and a large outgroup panel (100 sequences) but only restricting to sites that are at intermediate frequencies in the outgroup ($25\% \leq m_C/n_C \leq 75\%$), we find that performance is much improved relative to the case when we use all sites that are segregating in the outgroup (Figure 4.8).

Importantly, the usefulness of 3P-CLR(Int) resides not just in its performance at detecting selective sweeps in the ancestral population, but in its specific sensitivity to that particular type of events. Because the test relies on correlated allele frequency differences in both population a and population b (relative to the outgroup), selective sweeps that are specific to only one of the populations will not lead to high 3P-CLR(Int) scores, but will instead lead to high 3P-CLR(A) scores or 3P-CLR(B) scores, depending on where selection took place. Figure 4.9 shows ROC curves in two scenarios in which a selective sweep occurred only in population a (models I and J in Table 4.1), using 100 sampled individuals from each of the 3 populations. Here, XP-CLR performs well, but is outperformed by 3P-CLR(A). Furthermore, 3P-CLR(Int) shows almost no sensitivity to the recent sweep. For example, in Model I, at a specificity of 90%, 3P-CLR(A) and XP-CLR(A) have 86% and 80% sensitivity, respectively, while at the same specificity, 3P-CLR(Int) only has 18% sensitivity. One can compare this to the same demographic scenario but with selection occurring in the ancestral population of a and b (model C, Figure 4.2), where at 90% specificity, 3P-CLR(A) and XP-CLR(A) have 72% and 84% sensitivity, respectively, while 3P-CLR(Int) has 90% sensitivity. We also observe that 3P-CLR(A) is the best test at finding the true location of the selected site when selection occurs in the terminal branch leading to population a (Figure 4.10).

Finally, we tested the behavior of 3P-CLR under selective scenarios that we did not explicitly model. First, we simulated a selective sweep in the outgroup population. We find that all three types of 3P-CLR statistics (3P-CLR(Int), 3P-CLR(A) and 3P-CLR(B)) are largely insensitive to this type of event, though 3P-CLR(Int) is relatively more sensitive than the other two. Second, we simulated two independent selective sweeps in populations a and b (convergent evolution). This results in elevated 3P-CLR(A) and 3P-CLR(B) statistics, but 3P-CLR(Int) remains largely insensitive (Figure 4.11). We note that 3P-CLR should not be used to detect selective events that occurred before the split of all three populations (i.e. before the split of c and the ancestor of a and b), as it relies on allele frequency differences between the populations.

Selection in Eurasians

We first applied 3P-CLR to modern human data from phase 1 of the 1000 Genomes Project [153]. We used the African-American recombination map [162] to convert physical distances into genetic distances. We focused on Europeans (CEU, FIN, GBR, IBS, TSI) and East Asians (CHB, CHS, JPT) as the two sister populations, using Yoruba (YRI) as the outgroup population (Figure 4.12.A). We randomly sampled 100 individuals from each population and obtained sample derived allele frequencies every 10 SNPs in the genome. We then calculated likelihood ratio statistics by a sliding window approach, where we sampled a "central SNP" once every 10 SNPs. The central SNP in each window was the candidate beneficial SNP for that window. We set the window size to 0.25 cM, and randomly sampled 100 SNPs from each window, centered around the candidate beneficial SNP. In each window, we calculated 3P-CLR to test for selection at three different branches of the population tree: the terminal branch leading to Europeans (3P-CLR Europe), the terminal branch leading to East Asians (3P-CLR East Asia) and the ancestral branch of Europeans and East Asians (3P-CLR Eurasia). Results are shown in Figure 4.13. For each scan, we selected the windows in the top 99.9% quantile of scores and merged them together if their corresponding central SNPs were contiguous, effectively resulting in overlapping windows being merged. Tables 4.2, 4.3 and 4.4 show the top hits for Europeans, East Asians and the ancestral Eurasian branch, respectively

We observe several genes that were identified in previous selection scans. In the East Asian branch, one of the top hits is *EDAR*. Figure 4.14.A shows that this gene appears to be under selection exclusively in this population branch. It codes for a protein involved in hair thickness and incisor tooth morphology [163, 164], and has been repeatedly identified as a candidate for a sweep in East Asians [165, 166].

Furthermore, 3P-CLR allows us to narrow down on the specific time at which selection for previously found candidates occurred in the history of particular populations. For example, [147] performed a scan of the genomes of East Asians using XP-CLR with Yoruba as the outgroup, and identified a number of candidate genes [147]. 3P-CLR confirms recovers several of their loci when looking specifically at the East Asian branch: *OR56A1, OR56B4, OR52B2, SLC30A9, BBX, EPHB1, ACTN1* and *XKR6*. However, when applied to the ancestral Eurasian branch, 3P-CLR finds some genes that were previously found in the XP-CLR analysis of East Asians, but that are not among the top hits in 3P-CLR applied to the East Asian branch: *COMMD3, BMI1, SPAG6, NGLY1, OXSM, CD226, ABCC12, ABCC11, LONP2, SIAH1, PPARA, PKDREJ, GTSE1, TRMU* and *CELSR1*. This suggests selection in these regions occurred earlier, i.e. before the European-East Asian split. Figure 4.14.B shows a comparison between the 3P-CLR scores for the three branches in the region containing genes *BMI1* (a proto-oncogene [167]) and *SPAG6* (involved in sperm motility [168]). Here, the signal of Eurasia-specific selection is evidently stronger than the other two signals. Finally, we also find some candidates from [147] that appear to be under selection in both the ancestral Eurasian branch and the East Asian daughter branch: *SFXN5, EMX1, SPR* and *CYP26B1*. Interestingly, both CYP26B1 and CYP26A1 are very strong candidates for selection in the East Asian branch. These two genes lie in two different chromosomes, so they are not part of a gene cluster, but they both code for proteins that hydrolize retinoic acid, an important signaling molecule [169, 170].

Other selective events that 3P-CLR infers to have occurred in Eurasians include the region containing *HERC2* and *OCA2*, which are major determinants of eye color [171, 172, 173]. There is also evidence that these genes underwent selection more recently in the history of Europeans [174], which could suggest an extended period of selection - perhaps influenced by migrations between Asia and Europe - or repeated selective events at the same locus.

When running 3P-CLR to look for selection specific to Europe, we find that *TYRP1*, which plays a role in human skin pigmentation [175], is among the top hits. This gene has been previously found to be under strong selection in Europe [140], using a statistic called iHS, which measures extended patterns of haplotype homozygosity that are characteristic of selective sweeps. Interestingly, a change in the gene *TYRP1* has also been found to cause a blonde hair phenotype in Melanesians [176]. Another of our top hits is the region containing

SH2B3, which was identified previously as a candidate for selection in Europe based on both iHS and F_{st} [141]. This gene contains a nonsynonymous SNP (rs3184504) segregating in Europeans. One of its alleles (the one in the selected haplotype) has been associated with celiac disease and type 1 diabetes [177, 178] but is also protective against bacterial infection [179].

We used Gowinda (v1.12) [180] to find enriched Gene Ontology (GO) categories among the regions in the 99.5% highest quantile for each branch score, relative to the rest of the genome ($P < 0.05$, $FDR < 0.3$). The significantly enriched categories are listed in Table 4.5. In the East Asian branch, we find categories related to alcohol catabolism, retinol binding, vitamin metabolism and epidermis development, among others. In the European branch, we find cuticle development and hydrogen peroxide metabolic process as enriched categories. We find no enriched categories in the Eurasian branch that pass the above cutoffs.

Selection in ancestral modern humans

We applied 3P-CLR to modern human data combined with recently sequenced archaic human data. We sought to find selective events that occurred in modern humans after their spit from archaic groups. We used the combined Neanderthal and Denisovan high-coverage genomes [2, 16] as the outgroup population, and, for our two test populations, we used Eurasians (CEU, FIN, GBR, IBS, TSI, CHB, CHS, JPT) and Yoruba (YRI), again from phase 1 of the 1000 Genomes Project [153] (Figure 4.12.B). As before, we randomly sampled 100 genomes for each of the two daughter populations at each site, and tested for selective events that occurred more anciently than the split of Yoruba and Eurasians, but more recently than the split from Neanderthals. Figure 4.15 shows an ROC curve for a simulated scenario under these conditions, based on the history of population size changes inferred by PSMC [60, 16], suggesting we should have power to detect strong (s=0.1) selective events in the ancestral branch of present-day humans. We observe that 3P-CLR(Int) has similar power as XP-CLR and XP-CLR-avg at these time-scales, but is less prone to also detect recent (post-split) events, making it more specific to ancestral sweeps.

We ran 3P-CLR using 0.25 cM windows as above (Figure 4.16). As before, we selected the top 99.9% windows and merged them together if their corresponding central SNPs were contiguous (Table 4.6). Figure 4.16 shows that the outliers in the genome-wide distribution of scores are not strong. We wanted to verify that the density of scores was robust to the choice of window size. By using a larger window size (1 cM), we obtained a distribution with slightly more extreme outliers (Figures 4.17 4.16). For that reason, we also show the top hits from this large-window run (Table 4.7), using a smaller density of SNPs (200/1cM rather than 100/0.25cM), due to costs in speed. To find putative candidates for the beneficial variants in each region, we queried the catalogs of modern human-specific high-frequency or fixed derived changes that are ancestral in the Neanderthal and/or the Denisova genomes [16, 61] and overlapped them with our regions.

We found several genes that were identified in previous studies that looked for selection in modern humans after their split from archaic groups [16, 8], including *SIPA1L1, ANAPC10,* *ABCE1, RASA1, CCNH, KCNJ3, HBP1, COG5, CADPS2, FAM172A, POU5F2, FGF7, RABGAP1, SMURF1, GABRA2, ALMS1, PVRL3, EHBP1, VPS54, OTX1, UGP2, GTDC1, ZEB2* and *OIT3*. One of our strongest candidate genes among these is *SIPA1L1* (Figure 4.18.A), which is in the first and the fourth highest-ranking region, when using 1 cM and 0.25 cM windows, respectively. The protein encoded by this gene (E6TP1) is involved in actin cytoskeleton organization and controls neural morphology (UniProt by similarity). Interestingly, it is also a target of degradation of the oncoproteins of high-risk papillomaviruses [181].

Another candidate gene is *ANAPC10* (Figure 4.18.B). This gene codes for a core subunit of the cyclosome, which is involved in progression through the cell cycle [182], and may play a role in oocyte maturation and human T-lymphotropic virus infection (KEGG pathway [183]). *ANAPC10* is noteworthy because it was found to be significantly differentially expressed in humans compared to other great apes and macaques: it is up-regulated in the testes [184]. The gene also contains two intronic changes that are fixed derived in modern humans, ancestral in both Neanderthals and Denisovans and that have evidence for being highly disruptive, based on a composite score that combines conservation and regulatory data (PHRED-scaled C-scores $>$ 11 [185, 16]). The changes, however, appear not to lie in any obvious regulatory region [186, 187].

We also find *ADSL* among the list of candidates. This gene is known to contain a nonsynonymous change that is fixed in all present-day humans but homozygous ancestral in the Neanderthal genome, the Denisova genome and two Neanderthal exomes [61] (Figure 4.19.A). It was previously identified as lying in a region with strong support for positive selection in modern humans, using summary statistics implemented in an ABC method [37]. The gene is interesting because it is one of the members of the Human Phenotype ontology category "aggression / hyperactivity" which is enriched for nonsynonymous changes that occurred in the modern human lineage after the split from archaic humans [188, 61]. *ADSL* codes for adenylosuccinase, an enzyme involved in purine metabolism [189]. A deficiency of adenylosuccinase can lead to apraxia, speech deficits, delays in development and abnormal behavioral features, like hyperactivity and excessive laughter [190]. The nonsynonymous mutation (A429V) is in the C-terminal domain of the protein (Figure 4.19.B) and lies in a highly conserved position (primate PhastCons = 0.953; GERP score = 5.67 [191, 192, 185]). The ancestral amino acid is conserved across the tetrapod phylogeny, and the mutation is only three residues away from the most common causative SNP for severe adenylosuccinase deficiency [193, 194, 195, 196, 197]. The change has the highest probability of being disruptive to protein function, out of all the nonsynonymous modern-human-specific changes that lie in the top-scoring regions (C-score = 17.69). While *ADSL* is an interesting candidate and lies in the center of the inferred selected region (Figure 4.19.A), there are other genes in the region too, including *TNRC6B* and *MKL1*. *TNRC6B* may be involved in miRNA-guided gene silencing [198], while *MKL1* may play a role in smooth muscle differentiation [199], and has been associated with acute megakaryocytic leukemia [200].

RASA1 was also a top hit in a previous scan for selection [8], and was additionally inferred to have evidence in favor of selection in [37]. The gene codes for a protein involved in the control of cellular differentiation [201], and has a modern human-specific fixed nonsynonymous change (G70E). Human diseases associated with *RASA1* include basal cell carcinoma [202] and arteriovenous malformation [203, 204].

The $GABA_A$ gene cluster in chromosome 4p12 is also among the top regions. The gene within the putatively selected region codes for a subunit (*GABRA2*) of the $GABA_A$ receptor, which is a ligand-gated ion channel that plays a key role in synaptic inhibtion in the central nervous system (see review by [205]). *GABRA2* is significantly associated with risk of alcohol dependence in humans [206], perception of pain [207] and asthma [208].

Two other candidate genes that may be involved in brain development are *FOXG1* and *CADPS2*. *FOXG1* was not identified in any of the previous selection scans, and codes for a protein called forkhead box G1, which plays an important role during brain development. Mutations in this gene are associated with a slow-down in brain growth during childhood resulting in microcephaly, which in turn causes various intellectual disabilities [209, 210]. *CADPS2* was identified in [8] as a candidate for selection, and has been associated with autism [211]. The gene has been suggested to be specifically important in the evolution of all modern humans, as it was not found to be selected earlier in great apes or later in particular modern human populations [212].

Finally, we find a signal of selection in a region containing the gene *EHBP1* and *OTX1*. This region was identified in both of the two previous scans for modern human selection [16, 8]. *EHBP1* codes for a protein involved in endocytic trafficking [213] and has been associated with prostate cancer [214]. *OTX1* is a homeobox family gene that may play a role in brain development [215]. Interestingly, *EHBP1* contains a single-nucleotide intronic change (chr2:63206488) that is almost fixed in all present-day humans and homozygous ancestral in Neanderthal and Denisova [16]. This change is also predicted to be highly disruptive (C-score = 13.1) and lies in a position that is extremely conserved across primates (PhastCons = 0.942), mammals (PhastCons = 1) and vertebrates (PhastCons = 1). The change is 18 bp away from the nearest splice site and overlaps a VISTA conserved enhancer region (element 1874) [216], suggesting a putative regulatory role for the change.

We again used Gowinda [180] to find enriched GO categories among the regions with high 3P-CLR scores in the Modern Human branch. The significantly enriched categories ($P < 0.05$, $FDR < 0.3$) are listed in Table 4.5. We find several GO terms related to the regulation of the cell cycle, T cell migration and intracellular transport.

We overlapped the genome-wide association studies (GWAS) database [217, 218] with the list of fixed or high-frequency modern human-specific changes that are ancestral in archaic humans [16] and that are located within our top putatively selected regions in modern humans (Tables 4.8 and 4.9 for the 0.25 cM and 1 cM scans, respectively). None of the resulting SNPs are completely fixed derived, because GWAS can only yield associations from sites that are segregating. We find several SNPs in the *RAB28* gene [186, 187], which are significantly associated with obesity [219]. We also find a SNP with a high C-score (rs10171434) associated with urinary metabolites [220] and suicidal behavior in patients with mood disorders [221]. The SNP is located in an enhancer regulatory freature [186, 187] located between genes *PELI1* and *VPS54*, in the same putatively selected region as genes *EHBP1* and *OTX1* (see above). Finally, there is a highly C-scoring SNP (rs731108) that is associated with renal cell carcinoma [222]. This SNP is also located in an enhancer regulatory feature [186, 187], in an intron of *ZEB2*. In this last case, though, only the Neanderthal genome has the ancestral state, while the Denisova genome carries the modern human variant.

4.5 Discussion

We have developed a new method called 3P-CLR, which allows us to detect positive selection along the genome. The method is based on an earlier test (XP-CLR [147]) that uses linked allele frequency differences between two populations to detect population-specific selection. However, unlike XP-CLR, 3P-CLR can allow us to distinguish between selective events that occurred before and after the split of two populations. Our method has some similiarities to an earlier method developed by [223], which used an F_{st}-like score to detect selection ancestral to two populations. In that case, though, the authors used summary statistics and did not explicitly model the process leading to allele frequency differentiation. It is also similar to a more recent method [224] that models differences in haplotype frequencies between populations, while accounting for population structure.

We used our method to confirm previously found candidate genes in particular human populations, like *EDAR*, *TYRP1* and *CYP26B1*, and find some novel candidates too (Tables 4.2, 4.3, 4.4). Additionally, we can infer that certain genes, which were previously known to have been under selection in East Asians (like *SPAG6*), are more likely to have undergone a sweep in the population ancestral to both Europeans and East Asians than in East Asians only. We find that genes involved in epidermis development and alcohol catabolism are particularly enriched among the East Asian candidate regions, while genes involved in peroxide catabolism and cuticle development are enriched in the European branch. This suggests these biological functions may have been subject to positive selection in recent times.

We also used 3P-CLR to detect selective events that occurred in the ancestors of modern humans, after their split from Neanderthals and Denisovans (Table 4.6). These events could perhaps have led to the spread of phenotypes that set modern humans apart from other hominin groups. We find several intersting candidates, like *SIPA1L1, ADSL, RASA1, OTX1, EHBP1, FOXG1, RAB28* and *ANAPC10*, some of which were previously detected using other types of methods [8, 16, 37]. We also find an enrichment for GO categories related to cell cycle regulation and T cell migration among the candidate regions, suggesting that these biological processes might have been affected by positive selection after the split from archaic humans.

An advantage of differentiation-based tests like XP-CLR and 3P-CLR is that, unlike other patterns detected by tests of neutrality (like extended haplotype homozygostiy, [225]) that are exclusive to hard sweeps, the patterns that both XP-CLR and 3P-CLR are tailored to find are based on regional allele frequency differences between populations. These patterns can also be produced by soft sweeps from standing variation or by partial sweeps [147], and there is some evidence that the latter phenomena may have been more important than classic sweeps during human evolutionary history [226].

Another advantage of both XP-CLR and 3P-CLR is that they do not rely on an arbitrary division of genomic space. Unlike other methods which require the partition of the genome into small windows of fixed size, our composite likelihood ratios can theoretically be computed over windows that are as big as each chromosome, while only switching the central candidate site at each window. This is because the likelihood ratios use the genetic distance to the central SNP as input. SNPs that are very far away from the central SNP will not contribute much to the likelihood function of both the neutral and the selection models, while those that are close to it will. In the interest of speed, we heuristically limit the window size in our implementation, and use less SNPs when calculating likelihoods over larger windows. Nevertheless, these parameters can be arbitrarily adjusted by the user as needed, and if enough computing resources are available. The use of genetic distance in the likelihood function also allows us to take advantage of the spatial distribution of SNPs as an additional source of information, rather than only relying on patterns of population differentiation restricted to tightly linked SNPs.

3P-CLR also has an advantage over HMM-based selection methods, like the one implemented in [16]. The likelihood ratio scores obtained from 3P-CLR can provide an idea of how credible a selection model is for a particular region, relative to the rest of the genome. The HMM-based method previously used to scan for selection in modern humans [16] can only rank putatively selected regions by genetic distance, but cannot output a statistical measure that may indicate how likely each region is to have been under selection in ancient times. In contrast, 3P-CLR provides a composite likelihood ratio score, which allows for a statistically rigorous way to compare the neutral model and a specific selection model (for example, recent or ancient selection).

The outliers from Figure 4.13 have much higher scores (relative to the rest of the genome) than the outliers from Figure 4.16. This may be due to both the difference in time scales in the two sets of tests and to the uncertainty that comes from estimating outgroup allele frequencies using only two archaic genomes. This pattern can also be observed in Figure 4.20, where the densities of the scores looking for patterns of ancient selection (3P-CLR Modern Human and 3P-CLR Eurasia) have much shorter tails than the densities of scores looking for patterns of recent selection (3P-CLR Europe and 3P-CLR East Asia). Simulations show that 3P-CLR(Int) score distributions are naturally shorter than 3P-CLR(A) scores (Figure 4.21), which could explain the short tail of the 3P-CLR Eurasia distribution. Additionally, the even shorter tail in the distribution of 3P-CLR Modern Human scores may be a consequence of the fact that the split times of the demographic history in that case are older than the split times in the Eurasian tree, as simulations show that ancient split times tend to further shorten the tail of the 3P-CLR score distribution (Figure 4.21). We note, though, that using a larger window size produces a larger number of strong outliers (Figure 4.17).

A limitation of composite likelihood ratio tests is that the composite likelihood calculated for each model under comparison is obtained from a product of individual likelihoods at each site, and so it underestimates the correlation that exists between SNPs due to linkage effects

[157, 158, 227, 147]. One way to partially mitigate this problem is by using corrective weights based on linkage disequilibrium (LD) statistics calculated on the outgroup population [147]. Our implementation of 3P-CLR allows the user to incorporate such weights, if appropriate LD statistics are available from the outgroup. However, in cases where these are unreliable, it may not be possible to correct for this (for example, when only a few unphased genomes are available, as in the case of the Neanderthal and Denisova genomes).

While 3P-CLR relies on integrating over the possible allele frequencies in the ancestors of populations a and b (formula 4.10), one could envision using ancient DNA to avoid this step. Thus, if enough genomes could be sampled from that ancestral population that existed in the past, one could use the sample frequency in the ancient set of genomes as a proxy for the ancestral population frequency. This may soon be possible, as several early modern human genomes have already been sequenced in recent years [22, 23, 9].

Though we have focused on a three-population model in this manuscript, it should be straightforward to expand our method to a larger number of populations, albeit with additional costs in terms of speed and memory. 3P-CLR relies on a similar framework to the demographic inference method implemented in TreeMix [228], which can estimate population trees that include migration events, using genome-wide data. With a more complex modeling framework, it may be possible to estimate the time and strength of selective events with better resolution and using more populations, and also to incorporate additional demographic forces, like continuous migration between populations or pulses of admixture.

4.6 Acknowledgments

We thank Montgomery Slatkin, Rasmus Nielsen, Joshua Schraiber, Nicolas Duforet-Frebourg, Emilia Huerta-Sánchez, Hua Chen, Benjamin Peter, Nick Patterson, David Reich, Joachim Hermisson, Graham Coop and members of the Slatkin and Nielsen labs for helpful advice and discussions. We also thank two anonymous reviewers for their helpful comments. This work was supported by NIH grant R01-GM40282 to Montgomery Slatkin.

4.7 Tables

Table 4.1. Description of models tested. All times are in generations. Selection in the "ancestral population" refers to a selective sweep where the beneficial mutation and fixation occurred before the split time of the two most closely related populations. Selection in "daughter population a" refers to a selective sweep that occurred in one of the two most closely related populations (a), after their split from each other. t_{AB}: split time (in generations ago) of populations a and b. t_{ABC}: split time of population c and the ancestral population of a and b. t_M: time at which the selected mutation is introduced. s: selection coefficient. N_e: effective population size.

Model	Population where selection occurred	t_{AB}	t_{ABC}	t_M	s	N_e
A	Ancestral population	500	2,000	1,800	0.1	10,000
B	Ancestral population	1,000	4,000	2,500	0.1	10,000
C	Ancestral population	2,000	4,000	3,500	0.1	10,000
D	Ancestral population	3,000	8,000	5,000	0.1	10,000
E	Ancestral population	2,000	16,000	8,000	0.1	10,000
F	Ancestral population	4,000	16,000	8,000	0.1	10,000
I	Daughter population a	2,000	4,000	1,000	0.1	10,000
J	Daughter population a	3,000	8,000	2,000	0.1	10,000

Table 4.2. Top hits for 3P-CLR run on the European terminal branch, using Yoruba as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if the central SNPs that define them were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
9	125585000	125424000	126089000	362.273	ZBTB26,RABGAP1,GPR21,STRBP,OR1L1,OR1L3,OR1L4, OR1L6,OR5C1,PDCL,OR1K1,RC3H2,ZBTB6
22	35631900	35528100	35754100	309.488	HMGXB4,TOM1
8	52698800	52361800	52932100	289.921	PXDNL,PCMTD1
2	74967500	74450100	74972700	289.019	INO80B,WBP1,MOGS,MRPL53,CCDC142,TTC31,LBX2, PCGF1,TLX2,DQX1,AUP1,HTRA2,LOXL3,DOK1,M1AP, SEMA4F,SLC4A5,DCTN1,WDR54,RTKN
1	35634700	35382000	36592200	263.83	DLGAP3,ZMYM6NB,ZMYM6,ZMYM1,SFPQ,ZMYM4, KIAA0319L,NCDN,TFAP2E,PSMB2,C1orf216,CLSPN,AGO4, AGO1,AGO3,TEKT2,ADPRHL2,COL8A2
15	29279800	29248000	29338300	251.944	APBA2
12	112950000	111747000	113030000	242.067	BRAP,ACAD10,ALDH2,MAPKAPK5,TMEM116,ERP29, NAA25,TRAFD1,RPL6,PTPN11,RPH3A,CUX2,FAM109A, SH2B3,ATXN2
9	90947700	90909300	91210000	219.285	SPIN1,NXNL2
19	33644300	33504200	33705700	213.189	RHPN2,GPATCH1,WDR88,LRP3,SLC7A10
9	30546800	30085400	31031600	207.378	-
4	33865300	33604700	34355600	204.96	-
1	198035000	197943000	198308000	197.96	NEK7
1	204868000	204681000	204873000	194.594	NFASC
10	74613800	73802300	75407100	191.864	SPOCK2,ASCC1,ANAPC16,DDIT4,DNAJB12,MICU1,MCU, OIT3,PLA2G12B,P4HA1,NUDT13,ECD,FAM149B1,DNAJC9,MRPS16, TTC18,ANXA7,MSS51,PPP3CB,USP54,MYOZ1,SYNPO2L
7	138809000	138798000	139136000	180.75	TTC26,UBN2,C7orf55,C7orf55-LUC7L2,LUC7L2
6	95678500	95351800	95831000	180.676	-
2	104752000	104592000	104951000	177.053	-
16	7602450	7528820	7612510	171.615	RBFOX1
10	30568100	30361300	30629500	170.714	KIAA1462,MTPAP
3	137183000	136873000	137250000	166.559	-
1	116731000	116709000	116919000	165.137	ATP1A1
9	135136000	135132000	135298000	165.004	SETX,TTF1,C9orf171
13	89882200	89262100	90103800	158.112	-
2	17094600	16977500	17173100	156.531	-
4	82050400	81981400	82125100	154.54	PRKG2
2	69245100	69147300	69342700	149.948	GKN2,GKN1,ANTXR1
17	46949100	46821000	47137900	147.537	ATP5G1,UBE2Z,SNF8,GIP,IGF2BP1,TTLL6,CALCOCO2
10	83993700	83977100	84328100	147.072	NRG3
14	63893800	63780300	64044700	142.831	PPP2R5E
1	244070000	243645000	244107000	142.335	SDCCAG8,AKT3
14	66636800	66417700	67889500	140.97	GPHN,FAM71D,MPP5,ATP6V1D,EIF2S1,PLEK2
11	38611200	38349600	39004500	138.731	-
3	123368000	123196000	123418000	136.651	PTPLB,MYLK
6	112298000	111392000	112346000	135.167	SLC16A10,KIAA1919,REV3L,TRAF3IP2,FYN
5	109496000	109419000	109608000	132.766	-
5	142160000	142070000	142522000	132.436	FGF1,ARHGAP26
12	39050200	33590600	39618900	130.832	SYT10,ALG10,ALG10B,CPNE8
9	108423000	108410000	108674000	129.893	TAL2,TMEM38B
3	159453000	159263000	159486000	126.462	IQCJ-SCHIP1
2	70182800	70020100	70563900	126.092	FAM136A,ANXA4,GMCL1,SNRNP27,MXD1,ASPRV1,PCBP1,C2orf42,

					TIA1,PCYOX1,SNRPG
3	177605000	177536000	177745000	123.927	-
8	18534300	18515900	18656800	123.593	PSD3
5	123555000	123371000	123603000	122.973	-
17	19287500	18887800	19443300	122.35	SLC5A10,FAM83G,GRAP,GRAPL,EPN2,B9D1,MAPK7, MFAP4,RNF112,SLC47A1
11	42236100	41807600	42311500	122.131	-
13	41623700	41119400	41801600	121.214	FOXO1,MRPS31,SLC25A15,ELF1,WBP4,KBTBD6,KBTBD7, MTRF1
5	10311500	10284000	10481500	118.766	CMBL,MARCH6,ROPN1L
14	65288500	65222500	65472700	118.576	SPTB,CHURC1,FNTB,GPX2,RAB15
1	47651700	47396900	47938300	118.241	CYP4A11,CYP4X1,CYP4Z1,CYP4A22,PDZK1IP1,TAL1,STIL, CMPK1,FOXE3,FOXD2
2	138527000	138428000	138694000	116.881	-
17	42294300	42056700	42351800	115.466	PYY,NAGS,TMEM101,LSM12,G6PC3,HDAC5,C17orf53,ASB16, TMUB2,ATXN7L3,UBTF,SLC4A1
9	12480000	12439900	12776500	115.209	TYRP1,LURAP1L
7	78743000	78688400	78897900	114.946	MAGI2
2	216626000	216556000	216751000	114.901	-
1	65511700	65377500	65611400	114.699	JAK1
5	115391000	115369000	115784000	113.862	ARL14EPL,COMMD10,SEMA6A
15	45402300	45096000	45490700	113.69	C15orf43,SORD,DUOX2,DUOXA2,DUOXA1,DUOX1,SHF
3	25840300	25705200	25934000	113.326	TOP2B,NGLY1,OXSM
2	73086900	72373800	73148200	110.523	CYP26B1,EXOC6B,SPR,EMX1

Table 4.3. Top hits for 3P-CLR run on the East Asian terminal branch, using Yoruba as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if the central SNPs that define them were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
15	64151100	63693900	64188300	266.459	USP3,FBXL22,HERC1
10	94962900	94830500	95093900	241.875	CYP26A1,MYOF
2	73086900	72353500	73170800	218.482	CYP26B1,EXOC6B,SPR,EMX1,SFXN5
10	55988000	55869200	56263600	215.051	PCDH15
1	234359000	234209000	234396000	189.946	SLC35F3
5	117350000	117344000	117714000	189.051	-
17	60964400	60907300	61547900	186.63	TANC2,CYB561
2	44268900	44101400	44315200	185.629	ABCG8,LRPPRC
11	6126830	6028090	6191240	184	OR56A1,OR56B4,OR52B2
2	109318000	108905000	109629000	183.859	LIMS1,RANBP2,CCDC138,EDAR,SULT1C2,SULT1C4,GCC2
4	41882900	41456100	42196500	183.481	LIMCH1,PHOX2B,TMEM33,DCAF4L1,SLC30A9,BEND4
18	5304160	5201440	5314680	183.476	ZBTB14
9	105040000	104779000	105042000	181.781	-
7	105097000	104526000	105128000	181.358	KMT2E,SRPK2,PUS7
3	107609000	107149000	107725000	178.27	BBX
7	101729000	101511000	101942000	169.558	CUX1
6	159274000	159087000	159319000	169.058	SYTL3,EZR,C6orf99
9	90947700	90909300	91202200	163.828	SPIN1,NXNL2
9	92311400	92294400	92495100	162.821	-
15	26885200	26723700	26911100	160.496	GABRB3
5	109197000	108988000	109240000	156.271	MAN2A1
3	12506200	12476600	12819300	151.978	TSEN2,C3orf83,MKRN2,RAF1,TMEM40
2	125998000	125740000	126335000	148.576	-
3	139052000	139033000	139351000	148.572	MRPS22,COPB2,RBP2,RBP1,NMNAT3
3	134739000	134629000	135618000	146.833	EPHB1
2	9766680	9354260	9774110	145.998	ASAP2,ITGB1BP1,CPSF3,IAH1,ADAM17,YWHAQ
3	17873800	17189600	18009400	145.345	TBC1D5
14	69592000	69423900	69791100	144.488	ACTN1,DCAF5,EXD2,GALNT16
22	39747800	39574300	39845300	144.477	PDGFB,RPL3,SYNGR1,TAB1
8	10875300	10731100	11094000	143.754	XKR6
4	99985900	99712200	100322000	143.554	EIF4E,METAP1,ADH5,ADH4,ADH6,ADH1A,ADH1B
4	144235000	143610000	144412000	143.124	INPP4B,USP38,GAB1
2	17596700	16574500	17994400	142.084	FAM49A,RAD51AP2,VSNL1,SMC6,GEN1
2	211707000	211652000	211873000	141.706	-
1	103763000	103353000	103785000	141.473	COL11A1
3	71482600	71372800	71685500	140.75	FOXP1
17	10519000	10280200	10564000	140.243	MYH8,MYH4,MYH1,MYH2,MYH3
4	13283100	13126100	13537100	139.729	RAB28
8	73836900	73815300	73953100	139.423	KCNB2,TERF1
14	50226700	49952500	50426100	139.052	RPS29,LRR1,RPL36AL,MGAT2,DNAAF2,POLE2,KLHDC1, KLHDC2,NEMF,ARF6
2	26167200	25895300	26238100	138.585	KIF3C,DTNB
6	47369600	47312800	47708400	138.112	CD2AP,GPR115
3	102005000	101899000	102361000	137.862	ZPLD1
1	65943500	65891700	66168800	137.68	LEPR,LEPROT
11	25169300	24892400	25274500	137.191	LUZP2
1	28846900	28430000	29177900	136.458	PTAFR,DNAJC8,ATPIF1,SESN2,MED18,PHACTR4,RCC1, TRNAU1AP,TAF12,RAB42,GMEB1,YTHDF2,OPRD1
2	154054000	154009000	154319000	136.247	-

7	108874000	108718000	109226000	135.996	-
1	75471000	75277400	75941000	133.055	LHX8,SLC44A5
1	154824000	154802000	155113000	131.45	KCNN3,PMVK,PBXIP1,PYGO2,SHC1,CKS1B,FLAD1, LENEP,ZBTB7B,DCST2,DCST1,ADAM15,EFNA4, EFNA3,EFNA1,SLC50A1,DPM3
3	58413700	58096400	58550500	130.828	FLNB,DNASE1L3,ABHD6,RPP14,PXK,PDHB,KCTD6, ACOX2,FAM107A
1	36170500	35690600	36592200	130.701	ZMYM4,KIAA0319L,NCDN,TFAP2E,PSMB2,C1orf216, CLSPN,AGO4,AGO1,AGO3,TEKT2,ADPRHL2,COL8A2
17	39768900	39673200	39865400	130.04	KRT15,KRT19,KRT9,KRT14,KRT16,KRT17,JUP,EIF1
15	82080400	81842500	82171400	129.682	-
17	30842700	30613600	30868000	128.36	RHBDL3,C17orf75,ZNF207,PSMD11,CDK5R1,MYO1D
2	107933000	107782000	108041000	128.04	-
3	44917100	44138200	45133100	127.824	TOPAZ1,TCAIM,ZNF445,ZKSCAN7,ZNF660,ZNF197, ZNF35,ZNF502,ZNF501,KIAA1143,KIF15,TMEM42, TGM4,ZDHHC3,EXOSC7,CLEC3B,CDCP1
4	153009000	152902000	153101000	126.503	-
22	43190000	43148300	43455100	126.326	ARFGAP3,PACSIN2,TTLL1
4	168849000	168619000	168995000	126.125	-
5	42286000	41478600	42623200	125.831	PLCXD3,OXCT1,C5orf51,FBXO4,GHR
7	136345000	135788000	136570000	125.551	CHRM2
3	60305100	60226500	60349500	125.16	FHIT
10	59763900	59572200	59825500	124.643	-
3	114438000	114363000	115146000	124.535	ZBTB20
4	160142000	159944000	160359000	123.391	C4orf45,RAPGEF2
2	177717000	177613000	177889000	123.094	-
5	119672000	119639000	119868000	122.93	PRR16
20	43771800	43592200	43969300	122.421	STK4,KCNS1,WFDC5,WFDC12,PI3,SEMG1,SEMG2, SLPI,MATN4,RBPJL,SDC4
1	172928000	172668000	172942000	121.532	-
7	112273000	112126000	112622000	121.336	LSMEM1,TMEM168,C7orf60
1	169523000	169103000	169525000	119.533	NME7,BLZF1,CCDC181,SLC19A2,F5
3	26265100	25931700	26512400	119.052	-

Table 4.4. Top hits for 3P-CLR run on the Eurasian ancestral branch, using Yoruba as the outgroup. We show the windows in the top 99.9% quantile of scores. Windows were merged together if the central SNPs that define them were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
2	72379700	72353500	73170800	617.695	CYP26B1,EXOC6B,SPR,EMX1,SFXN5
20	53879500	53876700	54056200	605.789	-
10	22712400	22309300	22799200	566.463	EBLN1,COMMD3,COMMD3-BMI1,BMI1,SPAG6
3	25856600	25726300	26012000	557.376	NGLY1,OXSM
18	67725100	67523300	67910500	535.743	CD226,RTTN
10	66262400	65794400	66339100	532.732	-
11	39695600	39587400	39934300	518.72	-
7	138927000	138806000	139141000	508.385	TTC26,UBN2,C7orf55,C7orf55-LUC7L2,LUC7L2,KLRG2
9	90934600	90909300	91202200	498.898	SPIN1,NXNL2
4	41554200	41454200	42195300	487.476	LIMCH1,PHOX2B,TMEM33,DCAF4L1,SLC30A9,BEND4
16	61271700	61121600	61458700	485.291	-
17	58509300	58113700	59307700	477.117	HEATR6,CA4,USP32,C17orf64,APPBP2,PPM1D,BCAS3
1	230132000	229910000	230208000	468.258	GALNT2
8	35540400	35533900	35913800	454.601	UNC5D
17	60964400	60907300	61547900	449.203	TANC2,CYB561
16	47972300	33707000	48480500	448.504	SHCBP1,VPS35,ORC6,MYLK3,C16orf87,GPT2,DNAJA2,NETO2,ITFG1,PHKB,ABCC12,ABCC11,LONP2,SIAH1
1	90393900	90329700	90521600	436.002	LRRC8D,ZNF326
8	52698800	52238900	52932100	423.865	PXDNL,PCMTD1
11	106237000	105877000	106256000	419.391	MSANTD4,KBTBD3,AASDHPPT
13	48798100	48722300	49288100	414.218	ITM2B,RB1,LPAR6,RCBTB2,CYSLTR2
3	19240300	19090800	19424900	408.064	KCNH8
2	194986000	194680000	195299000	404.394	-
12	15962600	15690100	16137200	402.558	PTPRO,EPS8,STRAP,DERA
9	125564000	125484000	126074000	400.096	ZBTB26,RABGAP1,GPR21,STRBP,OR1L4,OR1L6,OR5C1,PDCL,OR1K1,RC3H2,ZBTB6
15	28565300	28324600	28611900	398.519	OCA2,HERC2
8	47631700	42502000	49037700	396.687	CHRNB3,CHRNA6,THAP1,RNF170,HOOK3,FNTA,POMK,HGSNAT,SPIDR,CEBPD,MCM4,UBE2V2
1	116994000	116808000	117027000	395.221	ATP1A1
7	99338700	98717600	99376500	393.41	ZSCAN25,CYP3A5,CYP3A7,CYP3A4,SMURF1,KPNA7,ARPC1A,ARPC1B,PDAP1,BUD31,PTCD1,ATP5J2-PTCD1,CPSF4,ATP5J2,ZNF789,ZNF394,ZKSCAN5,FAM200A,ZNF655
7	30343200	30178800	30485700	391.828	MTURN,ZNRF2,NOD1
10	31583000	31430600	31907900	389.863	ZEB1
6	10647900	10583800	10778900	387.883	GCNT2,C6orf52,PAK1IP1,TMEM14C,TMEM14B,SYCP2L,MAK
11	123275000	123156000	123313000	386.485	-
15	64642400	64333700	65204100	385.748	DAPK2,FAM96A,SNX1,SNX22,PPIB,CSNK1G1,KIAA0101,TRIP4,ZNF609,OAZ2,RBPMS2,PIF1,PLEKHO2
2	222560000	222523000	222690000	383.336	-
6	43620800	43398400	43687800	378.463	ABCC10,DLK2,TJAP1,LRRC73,POLR1C,YIPF3,XPO5,POLH,GTPBP2,MAD2L1BP,RSPH9,MRPS18A
14	57643800	57603400	58047900	378.332	EXOC5,AP5M1,NAA30,C14orf105
4	33487100	33294500	34347100	377.815	-
3	188699000	188647000	188856000	373.617	TPRG1
17	46949100	46821000	47137900	371.886	ATP5G1,UBE2Z,SNF8,GIP,IGF2BP1,TTLL6,CALCOCO2
4	172656000	172565000	172739000	369.949	GALNTL6
15	34404500	34212600	34413500	369.949	AVEN,CHRM5,EMC7,PGBD4
1	32888000	32445400	33065900	369.725	KHDRBS1,TMEM39B,KPNA6,TXLNA,CCDC28B,IQCC,DCDC2B,TMEM234,EIF3I,FAM167B,LCK,HDAC1,MARCKSL1,TSSK3,FAM229A,BSDC1,ZBTB8B,ZBTB8A,ZBTB8OS
22	46820900	46593300	46834700	369.511	PPARA,CDPF1,PKDREJ,TTC38,GTSE1,TRMU,CELSR1
10	93143600	93060500	93324900	368.648	HECTD2
6	14845800	14753800	14948200	367.9	-

Table 4.5. Enriched GO categories in the European, East Asian and Modern Human branches. We tested for ontology enrichment among the regions in the 99.5% quantile of the 3P-CLR scores for each population branch (P < 0.05, FDR < 0.3). The Eurasian branch did not have any category that passed these cutoffs.

Population Branch	Raw p-value	FDR	GO category
European	0.00002	0.05977	cuticle development
European	0.00007	0.096085	hydrogen peroxide catabolic process
East Asian	0.00001	0.013385	regulation of cell adhesion mediated by integrin
East Asian	0.00001	0.013385	epidermis development
East Asian	0.00014	0.14102	cell-substrate adhesion
East Asian	0.00023	0.185135	nucleosomal DNA binding
East Asian	0.0003	0.185135	nuclear chromosome
East Asian	0.00033	0.185135	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
East Asian	0.00048	0.2023525	negative regulation of vitamin metabolic process
East Asian	0.00048	0.2023525	substrate adhesion-dependent cell spreading
East Asian	0.00058	0.219074444	regulation of ERK1 and ERK2 cascade
East Asian	0.00077	0.258110909	retinol binding
East Asian	0.00084	0.258110909	primary alcohol catabolic process
East Asian	0.00112	0.296474	D1 dopamine receptor binding
East Asian	0.00125	0.296474	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription
East Asian	0.00127	0.296474	positive regulation of protein kinase B signaling
East Asian	0.0013	0.296474	gap junction assembly
Modern Human	0.00002	0.031153333	nuclear division
Modern Human	0.00003	0.031153333	organelle fission
Modern Human	0.00003	0.031153333	mitosis
Modern Human	0.00006	0.0490675	intra-Golgi vesicle-mediated transport
Modern Human	0.00012	0.069241429	regulation of cell cycle
Modern Human	0.00014	0.069241429	retinoic acid-responsive element binding
Modern Human	0.00015	0.069241429	cell cycle process
Modern Human	0.00029	0.12784125	T cell migration
Modern Human	0.00041	0.162306667	chromosomal part
Modern Human	0.00055	0.198124	'de novo' IMP biosynthetic process
Modern Human	0.00072	0.237017273	intracellular organelle
Modern Human	0.00081	0.24451	SNAP receptor activity
Modern Human	0.00113	0.294514286	ATP-dependent protein binding
Modern Human	0.00114	0.294514286	RNA biosynthetic process

Table 4.6. Top hits for 3P-CLR run on the ancestral branch to Eurasians and Yoruba, using archaic humans as the outgroup and 0.25 cM windows. We show the windows in the top 99.9% quantile of scores. Windows were merged together if the central SNPs that define them were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
2	95724900	95561200	96793700	859.783	ZNF514,ZNF2,PROM2,KCNIP3,FAHD2A,TRIM43,GPAT2,ADRA2B, ASTL,MAL,MRPS5
5	87054300	86463700	87101400	852.543	RASA1,CCNH
17	61538200	60910700	61557700	849.335	TANC2,CYB561,ACE
14	72207400	71649200	72283600	849.304	SIPA1L1
18	19089800	15012100	19548600	846.182	ROCK1,GREB1L,ESCO1,SNRPD1,ABHD3,MIB1
3	110675000	110513000	110932000	841.499	PVRL3
2	37990900	37917900	38024200	841.339	CDC42EP3
3	36938000	36836900	37517500	839.211	TRANK1,EPM2AIP1,MLH1,LRRFIP2,GOLGA4,C3orf35,ITGA9
7	107246000	106642000	107310000	838.948	PRKAR2B,HBP1,COG5,GPR22,DUS4L,BCAP29,SLC26A4
12	96986900	96823000	97411500	835	NEDD1
2	201056000	200639000	201340000	832.4	C2orf69,TYW5,C2orf47,SPATS2L
1	66851800	66772600	66952600	832.221	PDE4B
10	37795700	37165100	38978800	831.353	ANKRD30A,MTRNR2L7,ZNF248,ZNF25,ZNF33A,ZNF37A
2	156129000	155639000	156767000	827.839	KCNJ3
17	56516700	56379200	57404800	826.026	BZRAP1,SUPT4H1,RNF43,HSF5,MTMR4,SEPT4,C17orf47, TEX14,RAD51C,PPM1E, TRIM37,SKA2,PRR11,SMG8,GDPD1
5	18755900	18493900	18793500	825.858	-
2	61190300	61050900	61891900	824.962	REL,PUS10,PEX13,KIAA1841,AHSA2,USP34,XPO1
22	40392200	40360300	41213400	824.52	GRAP2,FAM83F,TNRC6B,ADSL,SGSM3,MKL1,MCHR1,SLC25A17
2	99013400	98996400	99383400	821.891	CNGA3,INPP4A,COA5,UNC50,MGAT4A
4	13294400	13137000	13533100	820.222	RAB28
18	32975600	32604100	33002800	819.128	MAPRE2,ZNF397,ZSCAN30,ZNF24,ZNF396
21	35204700	34737300	35222100	818.754	IFNGR2,TMEM50B,DNAJC28,GART,SON,DONSON,CRYZL1,ITSN1
12	73048100	72740100	73160400	816.903	TRHDE
1	213511000	213150000	213563000	814.632	VASH2,ANGEL2,RPS6KC1
1	27500300	26913700	27703900	814.332	ARID1A,PIGV,ZDHHC18,SFN,GPN2,GPATCH3,NUDC,NR0B2,C1orf172, TRNP1,FAM46B,SLC9A1,WDTC1,TMEM222,SYTL1,MAP3K6,FCN3
8	79219300	78698200	79558000	813.796	PKIA
12	116455000	116380000	116760000	809.406	MED13L
11	72857900	72416300	72912800	809.274	ARAP1,STARD10,ATG16L2,FCHSD2
4	22941400	22827300	23208900	808.696	-
12	79783400	79748800	80435300	804.117	SYT1,PAWR,PPP1R12A
13	35534800	35429700	36097500	801.815	NBEA,MAB21L1
4	146141000	145514000	146214000	799.686	HHIP,ANAPC10,ABCE1,OTUD4
16	61429300	61124400	61458700	798.318	-
4	46530000	46360000	46881700	797.876	GABRA2,COX7B2
2	133038000	132930000	133117000	796.277	-
17	28980100	28549700	29407200	796.136	SLC6A4,BLMH,TMIGD1,CPD,GOSR1,TBC1D29,CRLF3,ATAD5, TEFM,ADAP2,RNF135
5	127332000	127156000	127607000	789.339	SLC12A2,FBN2
5	27208300	27072700	27352900	788.924	CDH9
7	122294000	121973000	122559000	787.777	CADPS2,RNF133,RNF148
10	38218900	37175000	43224100	786.651	ANKRD30A,MTRNR2L7,ZNF248,ZNF25,ZNF33A,ZNF37A,ZNF33B
7	23100200	22888500	23114300	785.919	FAM126A
1	228050000	227587000	228112000	785.53	SNAP47,JMJD4,PRSS38,WNT9A
4	74891400	74846600	75086500	781.895	PF4,PPBP,CXCL5,CXCL3,CXCL2,MTHFD2L
22	34588400	34516300	34811800	781.522	-
2	63899700	62767900	64395700	778.951	EHBP1,OTX1,WDPCP,MDH1,UGP2,VPS54,PELI1

6	136666000	136527000	136967000	778.233	MTFR2,BCLAF1,MAP7,MAP3K5
16	75738400	75522400	75968000	778.171	CHST6,CHST5,TMEM231,GABARAPL2,ADAT1,KARS,TERF2IP
14	63446800	63288600	63597500	776.567	KCNH5
6	117528000	117080000	117579000	775.402	FAM162B,GPRC6A,RFX6
11	30206400	29986200	30443900	775.051	KCNA4,FSHB,ARL14EP,MPPED2
12	67533400	67436200	67639400	772.731	-
20	35460500	35049400	35710900	772.319	DLGAP4,MYL9,TGIF2,TGIF2-C20orf24,C20orf24,SLA2,NDRG3, DSN1,SOGA1,TLDC2,SAMHD1,RBL1
13	80131900	79801800	80268900	771.976	RBM26,NDFIP2
11	121408000	121310000	121493000	771.669	SORL1
4	105305000	104931000	105454000	770.437	CXXC4
5	93218900	92677500	93647600	769.192	NR2F1,FAM172A,POU5F2,KIAA0825
15	49975000	49247500	50040200	768.997	SECISBP2L,COPS2,GALK2,FAM227B,FGF7,DTWD1,SHC4
1	243669000	243505000	244087000	767.303	SDCCAG8,AKT3
21	36822500	36691000	36883300	762.715	RUNX1
1	154133000	153745000	154280000	762.43	INTS3,SLC27A3,GATAD2B,DENND4B,CRTC2,SLC39A1, CREB3L4,JTB,RAB13,RPS27,NUP210L,TPM3,C1orf189, C1orf43,UBAP2L,HAX1
7	144655000	144465000	144700000	762.429	TPK1
12	69177500	68890300	69290800	762.399	RAP1B,NUP107,SLC35E3,MDM2,CPM
2	145116000	144689000	145219000	757.235	GTDC1,ZEB2
1	176195000	175890000	176437000	755.81	RFWD2,PAPPA2
7	152155000	151699000	152199000	754.754	GALNTL5,GALNT11,KMT2C
7	116575000	116324000	116788000	754.606	MET,CAPZA2,ST7
14	29571400	29264600	29691100	754.435	-
1	226323000	226140000	226575000	754.04	SDE2,H3F3A,ACBD3,MIXL1,LIN9,PARP1
7	73051800	72317200	73134700	752.285	POM121,TRIM74,NSUN5,TRIM50,FKBP6,FZD9,BAZ1B,BCL7B, TBL2,MLXIPL,VPS37D,DNAJC30,WBSCR22,STX1A
5	89578700	89408400	89654700	751.498	-
8	22999100	22926500	23113900	749.992	TNFRSF10B,TNFRSF10C,TNFRSF10D,TNFRSF10A,CHMP7
15	75883900	75462000	76038100	749.953	C15orf39,GOLGA6C,GOLGA6D,COMMD4,NEIL1,MAN2C1, SIN3A,PTPN9,SNUPN,IMP3,SNX33,CSPG4,ODF3L1
7	98978400	98719400	99376100	749.35	ZSCAN25,CYP3A5,CYP3A7,CYP3A4,SMURF1,KPNA7,ARPC1A, ARPC1B,PDAP1,BUD31,PTCD1,ATP5J2-PTCD1,CPSF4,ATP5J2, ZNF789,ZNF394,ZKSCAN5,FAM200A,ZNF655
1	96340100	96155200	96608300	748.253	-
2	73508400	73482800	74054300	745.963	FBXO41,EGR4,ALMS1,NAT8,TPRKB,DUSP11,C2orf78
1	150868000	150224000	151137000	745.222	CA14,APH1A,C1orf54,C1orf51,MRPS21,PRPF3,RPRD2,TARS2, ECM1,ADAMTSL4,MCL1,ENSA,GOLPH3L,HORMAD1,CTSS, CTSK,ARNT,SETDB1,CERS2,ANXA9,FAM63A,PRUNE,BNIPL, C1orf56,CDC42SE1,MLLT11,GABPB2,SEMA6C,TNFAIP8L2, SCNM1,LYSMD1
3	99877600	99374500	100207000	744.933	COL8A1,CMSS1,FILIP1L,TBC1D23,NIT2,TOMM70A,LNP1
12	56244900	56086600	56360700	743.698	PMEL,CDK2,ITGA7,BLOC1S1,RDH5,CD63,GDF11,SARNP, ORMDL2,DNAJC14,MMP19,WIBG,DGKA
3	44843200	44139200	45128900	743.157	TOPAZ1,TCAIM,ZNF445,ZKSCAN7,ZNF660,ZNF197,ZNF35, ZNF502,ZNF501,KIAA1143,KIF15,TMEM42,TGM4,ZDHHC3, EXOSC7,CLEC3B,CDCP1
12	102922000	102388000	102964000	741.338	DRAM1,CCDC53,NUP37,PARPBP,PMCH,IGF1
1	21114300	21012100	21636800	740.553	KIF17,SH2D5,HP1BP3,EIF4G3,ECE1
11	108770000	108492000	108830000	740.463	DDX10
3	51678700	50188500	51919700	740.272	SEMA3F,GNAT1,GNAI2,LSMEM2,IFRD2,HYAL3,NAT6,HYAL1, HYAL2,TUSC2,RASSF1,ZMYND10,NPRL2,CYB561D2,TMEM115, CACNA2D2,C3orf18,HEMK1,CISH,MAPKAPK3,DOCK3,MANF, RBM15B,RAD54L2,TEX264,GRM2,IQCF6,IQCF3,IQCF2,IQCF5
11	64581900	64293300	64589300	738.648	RASGRP2,PYGM,SF1,MAP4K2,MEN1,SLC22A11,SLC22A12,NRXN2
9	126023000	125542000	126076000	738.221	ZBTB26,RABGAP1,GPR21,STRBP,OR5C1,PDCL,OR1K1,RC3H2,ZBTB6

Table 4.7. Top hits for 3P-CLR run on the ancestral branch to Eurasians and Yoruba, using archaic humans as the outgroup and 1 cM windows. We show the windows in the top 99.9% quantile of scores. Windows were merged together if the central SNPs that define them were contiguous. Win max = Location of window with maximum score. Win start = left-most end of left-most window for each region. Win end = right-most end of right-most window for each region. All positions were rounded to the nearest 100 bp. Score max = maximum score within region.

chr	Win max	Win start	Win end	Score max	Genes within region
14	71698500	71349200	72490300	1210.24	PCNX,SIPA1L1,RGS6
4	145534000	145023000	146522000	1157.25	GYPB,GYPA,HHIP,ANAPC10,ABCE1,OTUD4,SMAD1
2	156103000	155391000	156992000	1100.35	KCNJ3
5	93425300	92415600	94128600	1065.66	NR2F1,FAM172A,POU5F2,KIAA0825,ANKRD32,MCTP1
7	106717000	106401000	107461000	1049.82	PIK3CG,PRKAR2B,HBP1,COG5,GPR22,DUS4L,BCAP29,SLC26A4,CBLL1,SLC26A3
7	151831000	151651000	152286000	1028.93	GALNTL5,GALNT11,KMT2C
2	145008000	144393000	145305000	1027.28	ARHGAP15,GTDC1,ZEB2
19	16578500	16387600	16994000	991.083	KLF2,EPS15L1,CALR3,C19orf44,CHERP,SLC35E1,MED26,SMIM7,TMEM38A,NWD1,SIN3B
2	37996300	37730400	38054600	989.901	CDC42EP3
2	63467700	62639800	64698300	989.891	TMEM17,EHBP1,OTX1,WDPCP,MDH1,UGP2,VPS54,PELI1,LGALSL
10	38074100	36651400	44014800	988.663	ANKRD30A,MTRNR2L7,ZNF248,ZNF25,ZNF33A,ZNF37A,ZNF33B,BMS1,RET,CSGALNACT2,RASGEF1A,FXYD4,HNRNPF
1	27203100	26703800	27886000	988.598	LIN28A,DHDDS,HMGN2,RPS6KA1,ARID1A,PIGV,ZDHHC18,SFN,GPN2,GPATCH3,NUDC,NR0B2,C1orf172,TRNP1,FAM46B,SLC9A1,WDTC1,TMEM222,SYTL1,MAP3K6,FCN3,CD164L2,GPR3,WASF2,AHDC1
12	102906000	102308000	103125000	966.591	DRAM1,CCDC53,NUP37,PARPBP,PMCH,IGF1
2	133034000	132628000	133270000	941.856	GPR39
15	43507200	42284300	45101400	938.129	PLA2G4E,PLA2G4D,PLA2G4F,VPS39,TMEM87A,GANC,CAPN3,ZNF106,SNAP23,LRRC57,HAUS2,STARD9,CDAN1,TTBK2,UBR1,EPB42,TMEM62,CCNDBP1,TGM5,TGM7,LCMT2,ADAL,ZSCAN29,TUBGCP4,TP53BP1,MAP1A,PPIP5K1,CKMT1B,STRC,CATSPER2,CKMT1A,PDIA3,ELL3,SERF2,SERINC4,HYPK,MFAP1,WDR76,FRMD5,CASC4,CTDSPL2,EIF3J,SPG11,PATL2,B2M,TRIM69
2	73848400	73178500	74194400	934.997	SFXN5,RAB11FIP5,NOTO,SMYD5,PRADC1,CCT7,FBXO41,EGR4,ALMS1,NAT8,TPRKB,DUSP11,C2orf78,STAMBP,ACTG2,DGUOK
5	54861800	54193000	55422100	927.745	ESM1,GZMK,GZMA,CDC20B,GPX8,MCIDAS,CCNO,DHX29,SKIV2L2,PPAP2A,SLC38A9,DDX4,IL31RA,IL6ST,ANKRD55
3	52356200	50184000	53602300	925.895	SEMA3F,GNAT1,GNAI2,LSMEM2,IFRD2,HYAL3,NAT6,HYAL1,HYAL2,TUSC2,RASSF1,ZMYND10,NPRL2,CYB561D2,TMEM115,CACNA2D2,C3orf18,HEMK1,CISH,MAPKAPK3,DOCK3,MANF,RBM15B,RAD54L2,TEX264,GRM2,IQCF6,IQCF3,IQCF2,IQCF5,IQCF1,RRP9,PARP3,GPR62,PCBP4,ABHD14B,ABHD14A,ACY1,RPL29,DUSP7,POC1A,ALAS1,TLR9,TWF2,PPM1M,WDR82,GLYCTK,DNAH1,BAP1,PHF7,SEMA3G,TNNC1,NISCH,STAB1,NT5DC2,SMIM4,PBRM1,GNL3,GLT8D1,SPCS1,NEK4,ITIH1,ITIH3,ITIH4,MUSTN1,TMEM110-MUSTN1,TMEM110,SFMBT1,RFT1,PRKCD,TKT,CACNA1D
13	96364900	96038900	97500100	923.257	CLDN10,DZIP1,DNAJC3,UGGT2,HS6ST3
18	19248800	14517500	19962400	920.641	POTEC,ANKRD30B,ROCK1,GREB1L,ESCO1,SNRPD1,ABHD3,MIB1,GATA6
7	116587000	116214000	117339000	918.567	MET,CAPZA2,ST7,WNT2,ASZ1,CFTR

14	29544300	29031800	29913200	918.292	FOXG1
7	94710700	93964000	95170200	910.235	COL1A2,CASD1,SGCE,PEG10,PPP1R9A,PON1,PON3,PON2, ASB4
12	79783400	79231600	80435300	906.28	SYT1,PAWR,PPP1R12A
19	19290700	18936200	19885600	905.94	UPF1,CERS1,GDF1,COPE,DDX49,HOMER3,SUGP2,ARMC6, SLC25A42,TMEM161A,MEF2BNB-MEF2B,MEF2B,MEF2BNB, RFXANK,NR2C2AP,NCAN,HAPLN4,TM6SF2,SUGP1,MAU2, GATAD2A,TSSK6,NDUFA13,YJEFN3,CILP2,PBX4,LPAR2, GMIP,ATP13A1,ZNF101,ZNF14
11	72551000	72182800	72952400	902.837	PDE2A,ARAP1,STARD10,ATG16L2,FCHSD2,P2RY2
14	31685700	31255700	32384600	895.417	COCH,STRN3,AP4S1,HECTD1,DTD2,NUBPL

Table 4.8. Overlap between GWAS catalog and catalog of modern human-specific high-frequency changes in the top modern human selected regions (0.25 cM scan). Chr = chromosome. Pos = position (hg19). ID = SNP rs ID. Hum = Present-day human major allele. Anc = Human-Chimpanzee ancestor allele. Arch = Archaic human allele states (Altai Neanderthal, Denisova) where H=human-like allele and A=ancestral allele. Freq = present-day human derived frequency. Cons = consequence. C = C-score. PubMed = PubMed article ID for GWAS study.

Chr	Pos	Hum	Anc	Arch	Freq	Gene	Cons	C	GWAS trait
1	27138393	C	T	A/A,A/A	0.95	Metazoa SRP	upstream	4.193	HDL cholesterol
1	27138393	C	T	A/A,A/A	0.95	Metazoa SRP	upstream	4.193	LDL cholesterol
1	27138393	C	T	A/A,A/A	0.95	Metazoa SRP	upstream	4.193	Triglycerides
1	151009719	A	G	A/A,A/A	0.92	BNIPL	intron	7.111	DNA methylation, in blood cell lines
1	244044810	A	C	A/A,A/A	0.94	NA	intergenic	2.376	Response to taxane treatment (placlitaxel)
2	64279606	C	T	A/A,A/A	0.92	NA	intergenic	8.324	Suicide attempts in bipolar disorder
2	64279606	C	T	A/A,A/A	0.92	NA	intergenic	8.324	Urinary metabolites
2	144783214	T	C	A/A,A/A	0.93	GTDC1	intron	4.096	Body mass index
2	144783214	T	C	A/A,A/A	0.93	GTDC1	intron	4.096	Body mass index
2	145213638	G	C	A/A,H/H	0.92	ZEB2	intron,nc	12.16	Renal cell carcinoma
2	156506516	C	T	A/A,A/A	0.92	NA	intergenic	2.077	Alcohol consumption
3	51142359	T	C	A/A,A/A	0.91	DOCK3	intron	2.344	Multiple complex diseases
3	51824167	G	C	A/A,A/A	0.94	NA	intergenic	2.285	Response to taxane treatment (placlitaxel)
4	13325741	G	C	A/A,A/A	0.91	NA	intergenic	0.56	Obesity (extreme)
4	13328373	T	C	A/A,A/A	0.92	NA	intergenic	3.609	Obesity (extreme)
4	13330095	C	T	A/A,A/A	0.92	NA	intergenic	0.303	Multiple complex diseases
4	13330095	C	T	A/A,A/A	0.92	NA	intergenic	0.303	Obesity (extreme)
4	13333413	A	G	A/A,A/A	0.92	HSP90AB2P	upstream	4.041	Obesity (extreme)
4	13338465	C	T	A/A,A/A	0.92	HSP90AB2P	intron,nc	10.31	Obesity (extreme)
4	13340249	T	C	A/A,A/A	0.92	HSP90AB2P	non coding exon,nc	0.873	Obesity (extreme)
4	13346602	C	T	A/A,A/A	0.92	NA	intergenic	0.22	Obesity (extreme)
4	13350973	T	C	A/A,A/A	0.92	NA	regulatory	3.346	Obesity (extreme)
4	13356393	G	A	A/A,A/A	0.94	NA	intergenic	1.347	Obesity (extreme)
4	13357274	A	G	A/A,A/A	0.94	NA	intergenic	20	Obesity (extreme)
4	13360622	T	A	A/A,A/A	0.93	RAB28	downstream	4.509	Obesity (extreme)
4	13363958	A	G	A/A,A/A	0.97	RAB28	intron	1.536	Obesity (extreme)
4	13363974	C	T	A/A,A/A	0.97	RAB28	intron	0.363	Obesity (extreme)
4	13366481	C	T	A/A,A/A	0.93	RAB28	intron	3.083	Obesity (extreme)
4	13370308	T	C	A/A,A/A	0.93	RAB28	intron	14.23	Obesity (extreme)
4	13373583	C	T	A/A,A/A	0.97	RAB28	intron	0.402	Obesity (extreme)
4	13374462	G	A	A/A,A/A	0.93	RAB28	intron	0.826	Obesity (extreme)
4	13393897	A	T	A/A,A/A	0.96	RAB28	intron	2.579	Obesity (extreme)
4	13403855	G	A	A/A,A/A	0.94	RAB28	intron	0.842	Multiple complex diseases
4	13403855	G	A	A/A,A/A	0.94	RAB28	intron	0.842	Obesity (extreme)
4	13403998	G	A	A/A,A/A	0.93	RAB28	intron	1.179	Obesity (extreme)
4	13404130	G	T	A/A,A/A	0.94	RAB28	intron	0.385	Obesity (extreme)
4	13404717	A	C	A/A,A/A	0.93	RAB28	intron	1.116	Obesity (extreme)

4	13440031	C	G	A/A,A/A	0.93	RAB28	intron	0.138	Obesity (extreme)
4	13440271	C	T	A/A,A/A	0.94	RAB28	intron	0.54	Obesity (extreme)
4	13449532	A	C	A/A,A/A	0.94	RAB28	intron	0.905	Obesity (extreme)
4	13452022	C	A	A/A,A/A	0.91	RAB28	intron	3.789	Obesity (extreme)
4	13463991	T	C	A/A,A/A	0.93	RAB28	intron	3.377	Obesity (extreme)
4	13465710	T	A	A/A,A/A	0.93	RAB28	intron	1.709	Obesity (extreme)
4	23095293	C	T	A/A,A/A	0.96	NA	intergenic	0.797	Multiple complex diseases
5	89540468	C	T	A/A,A/A	0.97	RP11-61G23.1	intron,nc	3.627	Multiple complex diseases
6	136947540	A	G	A/A,A/A	0.93	MAP3K5	intron	0.586	Blood pressure, CVD RF and other traits
7	72746648	C	T	A/A,A/A	0.97	TRIM50	upstream	1.88	Immune reponse to smallpox
7	106720932	G	A	A/A,A/A	0.93	NA	regulatory	3.447	Multiple complex diseases
7	116668662	C	T	A/A,A/A	0.93	ST7-OT4	intron,nc	8.279	Response to gemcitabine or arabinosylcytosin
7	116668662	C	T	A/A,A/A	0.93	ST7-OT4	intron,nc	8.279	Response to gemcitabine or arabinosylcytosin
10	37579117	A	C	A/A,A/A	0.94	ATP8A2P1	intron,nc	2.346	Multiple complex diseases
10	37579117	A	C	A/A,A/A	0.94	ATP8A2P1	intron,nc	2.346	Multiple complex diseases
12	56308562	G	T	A/A,A/A	0.96	NA	regulatory	1.192	Multiple complex diseases
12	72889122	A	T	A/A,A/A	0.93	TRHDE	intron	4.133	Multiple complex diseases
13	35811439	C	T	A/A,A/A	0.93	NBEA	intron	3.514	Body mass index
16	61340362	G	C	A/A,A/A	0.93	NA	intergenic	4.37	Multiple complex diseases
22	34557399	T	G	A/A,A/A	0.93	LL22NC03-86D4.1	intron,nc	1.126	HIV-1 viral setpoint

Table 4.9. Overlap between GWAS catalog and catalog of modern human-specific high-frequency changes in the top modern human selected regions (1 cM scan). Chr = chromosome. Pos = position (hg19). ID = SNP rs ID. Hum = Present-day human major allele. Anc = Human-Chimpanzee ancestor allele. Arch = Archaic human allele states (Altai Neanderthal, Denisova) where H=human-like allele and A=ancestral allele. Freq = present-day human derived frequency. Cons = consequence. C = C-score. PubMed = PubMed article ID for GWAS study.

Chr	Pos	Hum	Anc	Arch	Freq	Gene	Cons	C	GWAS trait
1	27138393	C	T	A/A,A/A	0.95	Metazoa SRP	upstream	4.193	HDL cholesterol
1	27138393	C	T	A/A,A/A	0.95	Metazoa SRP	upstream	4.193	LDL cholesterol
1	27138393	C	T	A/A,A/A	0.95	Metazoa SRP	upstream	4.193	Triglycerides
2	64279606	C	T	A/A,A/A	0.92	NA	intergenic	8.324	Suicide attempts in bipolar disorder
2	64279606	C	T	A/A,A/A	0.92	NA	intergenic	8.324	Urinary metabolites
2	144783214	T	C	A/A,A/A	0.93	GTDC1	intron	4.096	Body mass index
2	144783214	T	C	A/A,A/A	0.93	GTDC1	intron	4.096	Body mass index
2	145213638	G	C	A/A,H/H	0.92	ZEB2	intron,nc	12.16	Renal cell carcinoma
2	156506516	C	T	A/A,A/A	0.92	NA	intergenic	2.077	Alcohol consumption
3	51142359	T	C	A/A,A/A	0.91	DOCK3	intron	2.344	Multiple complex diseases
3	51824167	G	C	A/A,A/A	0.94	NA	intergenic	2.285	Response to taxane treatment (placlitaxel)
3	52506426	T	C	A/A,A/A	0.96	NA	regulatory	0.316	Waist-hip ratio
5	54558972	G	A	A/A,A/A	0.92	DHX29	intron	5.673	Alcohol and nicotine co-dependence
7	106720932	G	A	A/A,A/A	0.93	NA	regulatory	3.447	Multiple complex diseases
7	116668662	C	T	A/A,A/A	0.93	ST7-OT4	intron,nc	8.279	Response to gemcitabine or arabinosylcytosin
7	116668662	C	T	A/A,A/A	0.93	ST7-OT4	intron,nc	8.279	Response to gemcitabine or arabinosylcytosin
10	37579117	A	C	A/A,A/A	0.94	ATP8A2P1	intron,nc	2.346	Multiple complex diseases
12	79387804	C	T	A/A,A/A	0.92	RP11-390N6.1	intron,nc	2.716	Response to taxane treatment (placlitaxel)
15	42527218	C	A	A/A,A/A	0.91	TMEM87A	intron	10.12	Multiple complex diseases

4.8 Figures

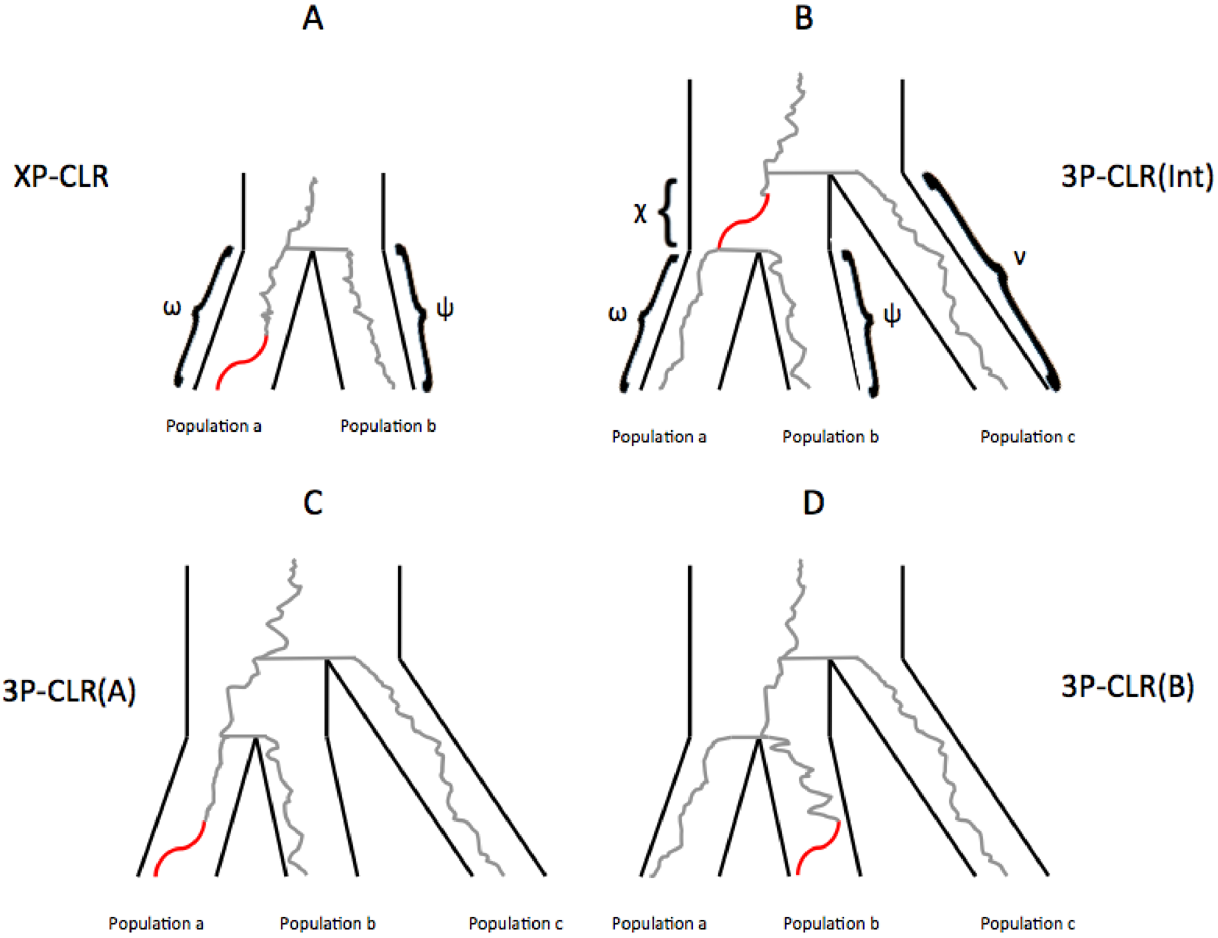


Figure 4.1. Schematic tree of selective sweeps detected by XP-CLR and 3P-CLR. While XP-CLR can only use two populations (an outgroup and a test) to detect selection (panel A), 3P-CLR can detect selection in the ancestral branch of two populations (3P-CLR(Int), panel B) or on the branches specific to each population (3P-CLR(A) and 3P-CLR(B), panels C and D, respectively). The greek letters denote the known drift times for each branch of the population tree.

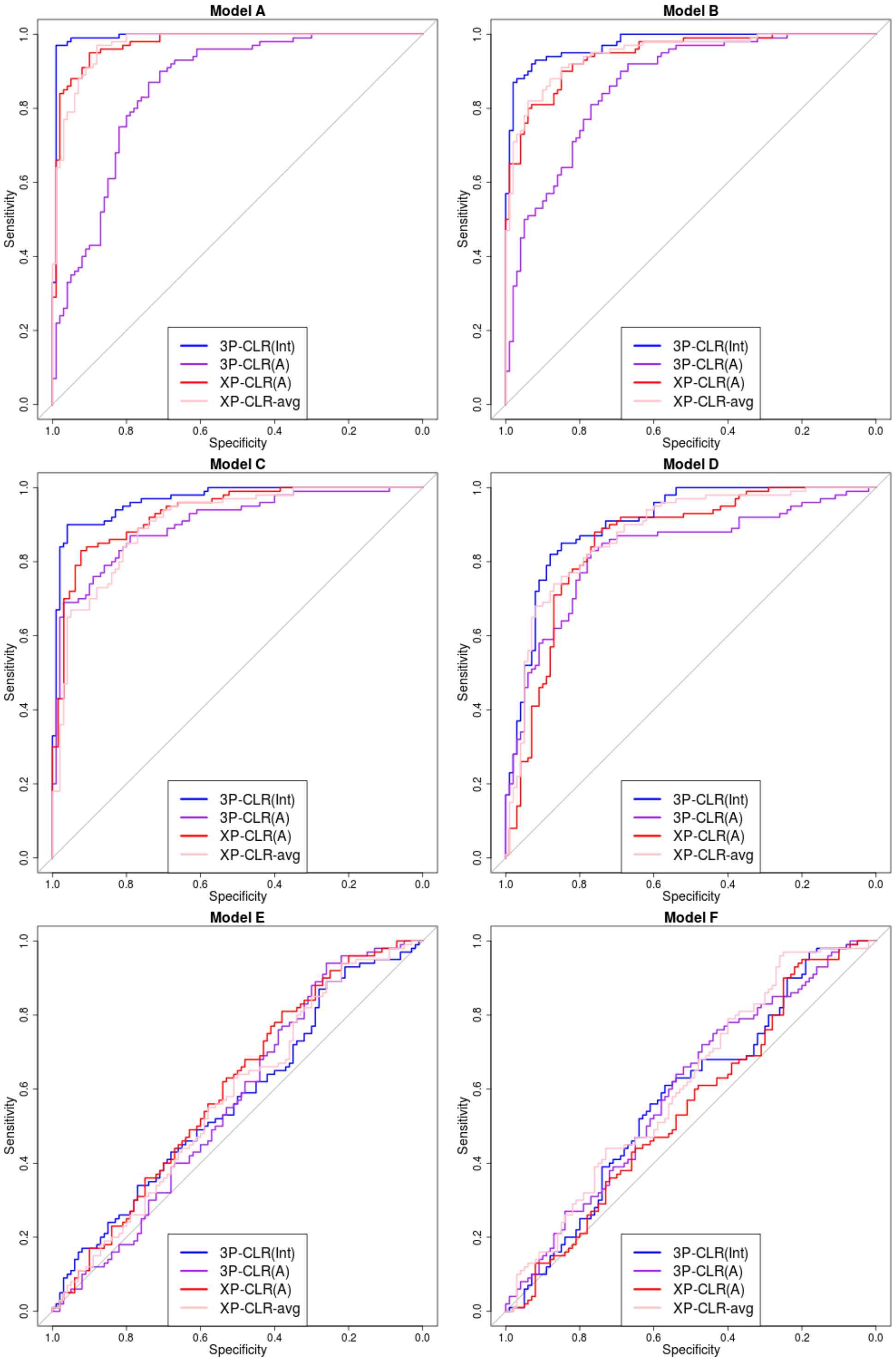


Figure 4.2. ROC curves for performance of 3P-CLR(Int), 3P-CLR(A) and two variants of XP-CLR in detecting selective sweeps that occurred before the split of two populations a and b, under different demographic models. In this case, the outgroup panel from population c contained 100 haploid genomes. The two sister population panels (from a and b) also have 100 haploid genomes each.

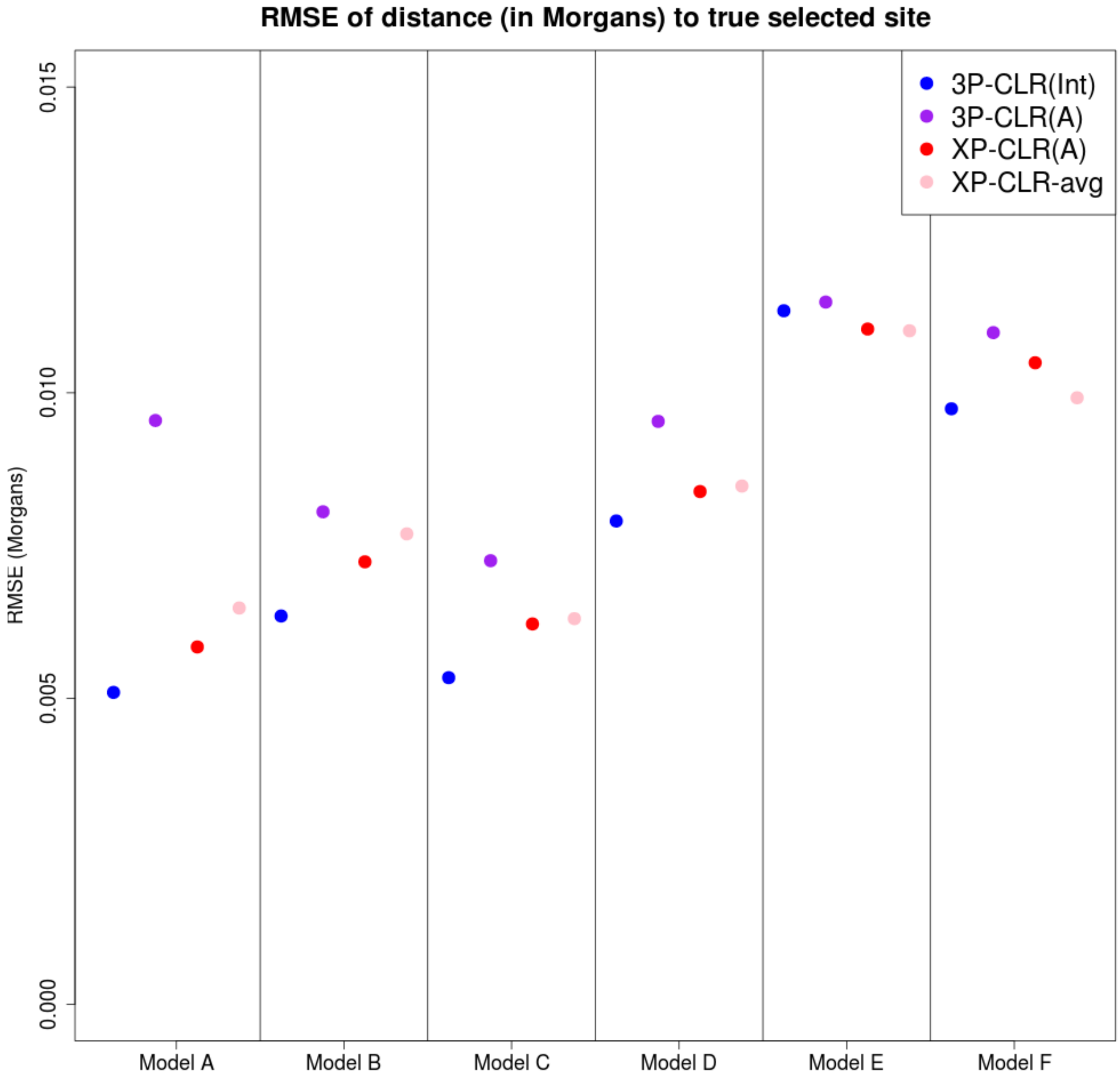


Figure 4.3. Root-mean squared error for the location of sweeps inferred by 3P-CLR(Int), 3P-CLR(A) and two variants of XP-CLR under different demographic scenarios, when the sweeps occurred before the split of populations a and b. In this case, the outgroup panel from population c contained 100 haploid genomes and the two sister population panels (from a and b) have 100 haploid genomes each.

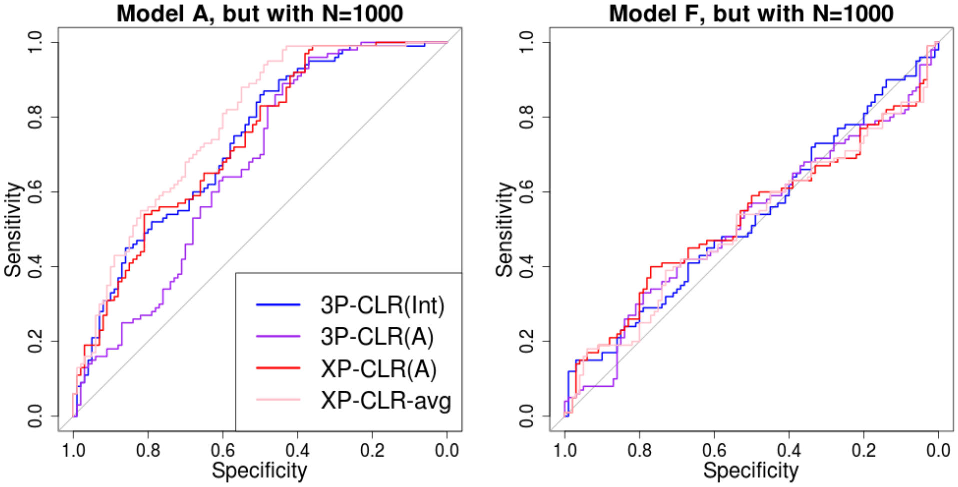


Figure 4.4. ROC curves for performance of 3P-CLR(Int), 3P-CLR(A) and two variants of XP-CLR in detecting selective sweeps that occurred before the split of two populations a and b, under two demographic models where the population size is extremely small ($N_e = 1{,}000$).

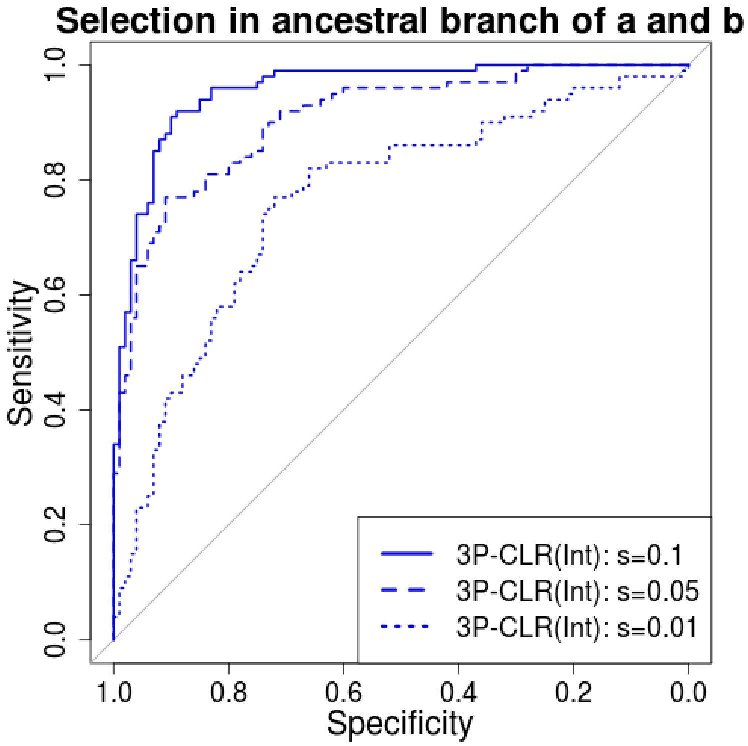


Figure 4.5. Performance of 3P-CLR(Int) for a range of selection coefficients. We used the demographic history from model B (Table 4.1) but extended the most ancient split time by 4,000 generations. The reason for this is that we wanted the internal branch to be long enough for it to be easy to sample simulations in which the beneficial allele fixed before the split of populations a and b, even for weak selection coefficients.

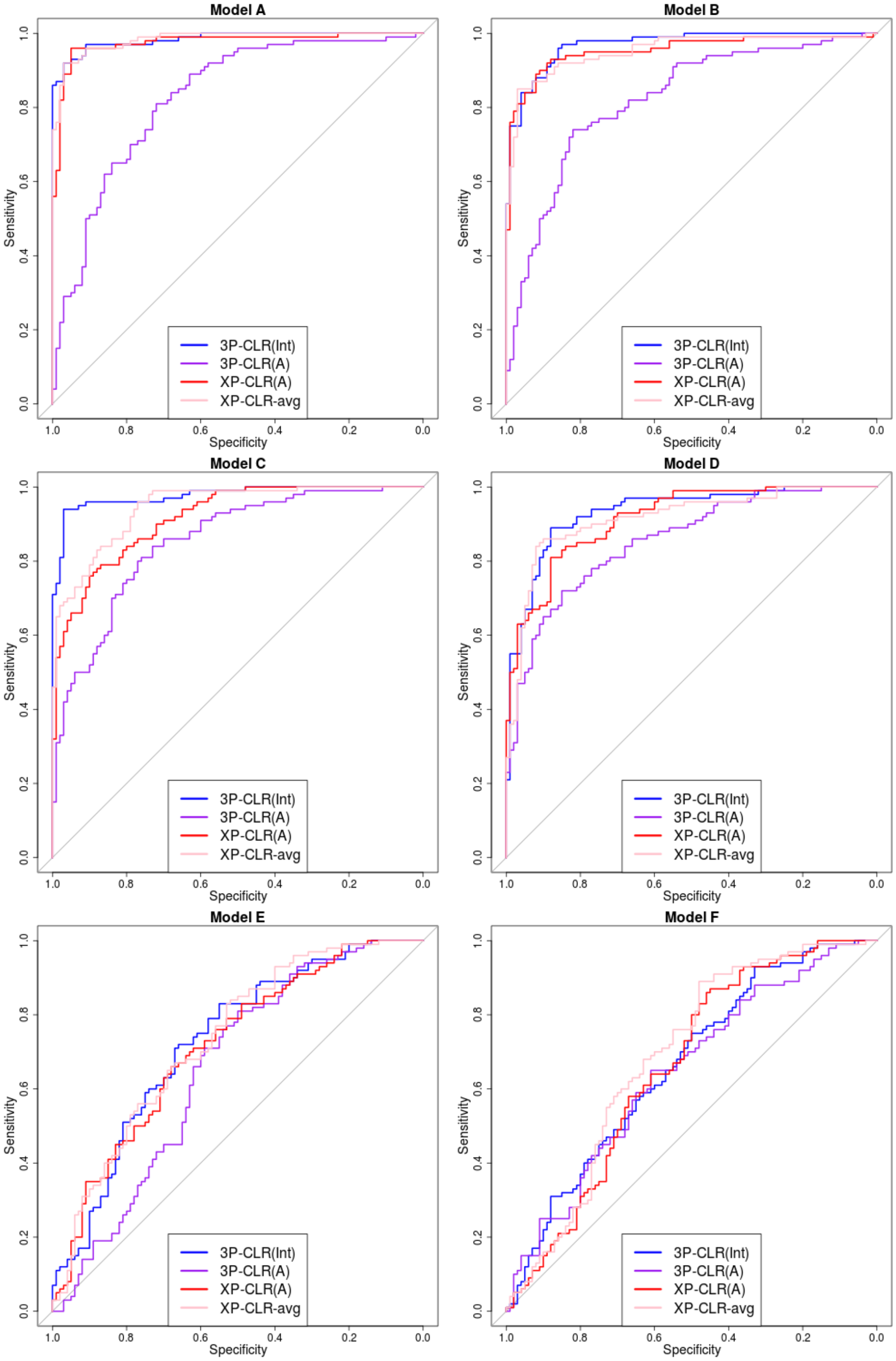


Figure 4.6. ROC curves for performance of 3P-CLR(Int), 3P-CLR(A) and two variants of XP-CLR in detecting selective sweeps that occurred before the split of two populations a and b, under different demographic models. In this case, the outgroup panel from population c contained 10 haploid genomes. The two sister population panels (from a and b) have 100 haploid genomes each.

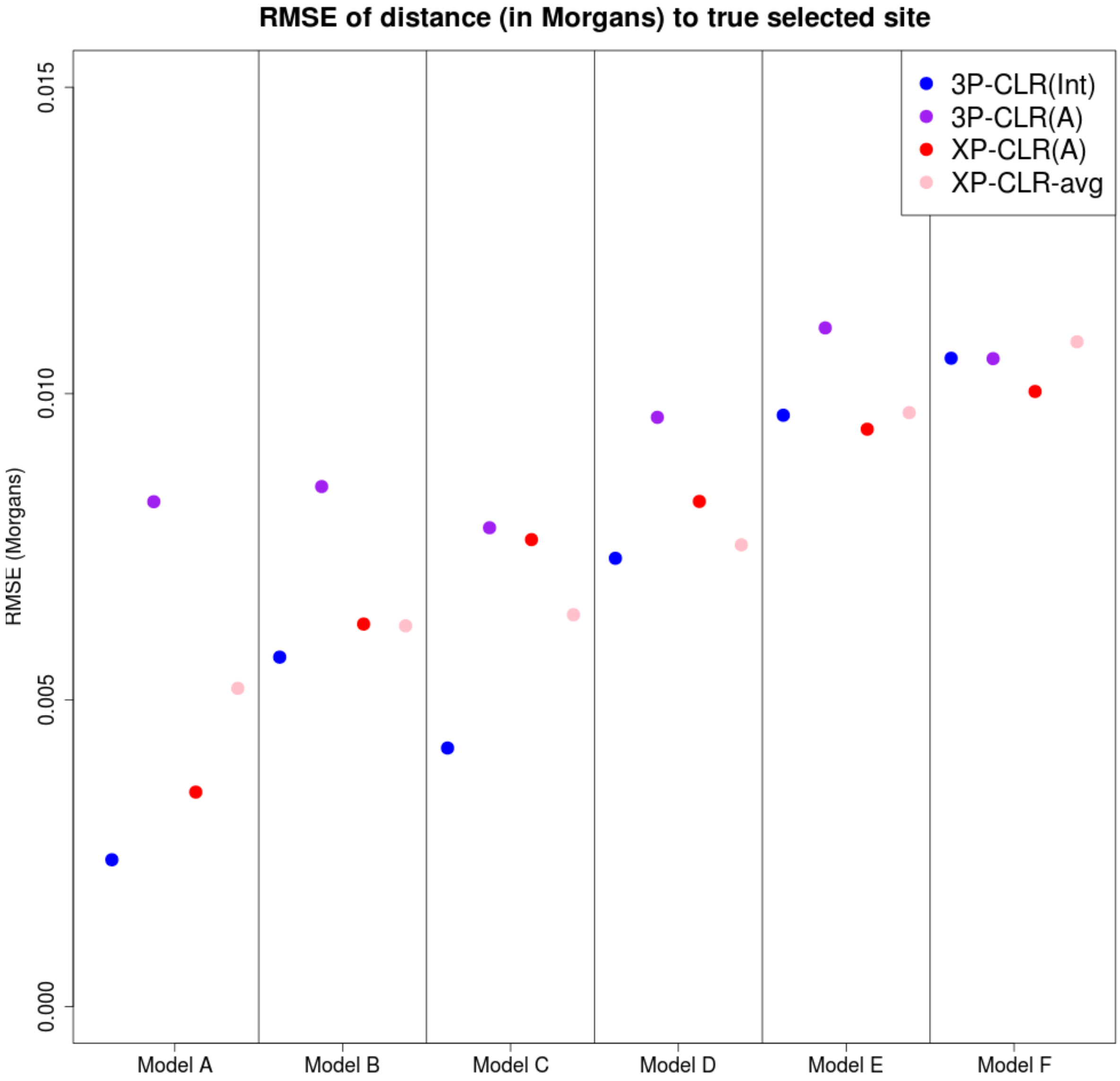


Figure 4.7. Root-mean squared error for the location of the sweep inferred by 3P-CLR(Int), 3P-CLR(A) and two variants of XP-CLR under different demographic scenarios, when the sweeps occurred before the split of populations a and b. the outgroup panel from population c contained 10 haploid genomes and the two sister population panels (from a and b) have 100 haploid genomes each.

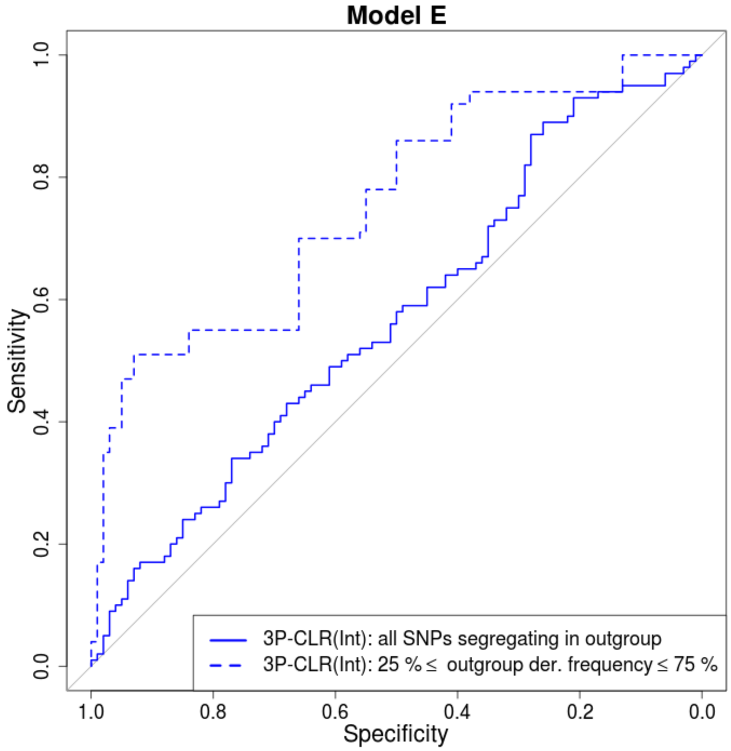


Figure 4.8. For demographic scenarios with very ancient split times, it is best to use sites segregating at intermediate frequencies in the outgroup. We compared the performance of 3P-CLR(Int) in a demographic scenario with very ancient split times (Model E) under two conditions: including all SNPs that are segregating in the outgroup, and only including SNPs segregating at intermediate frequencies in the outgroup. In both cases, the number of sampled sequences from the outgroup population was 100.

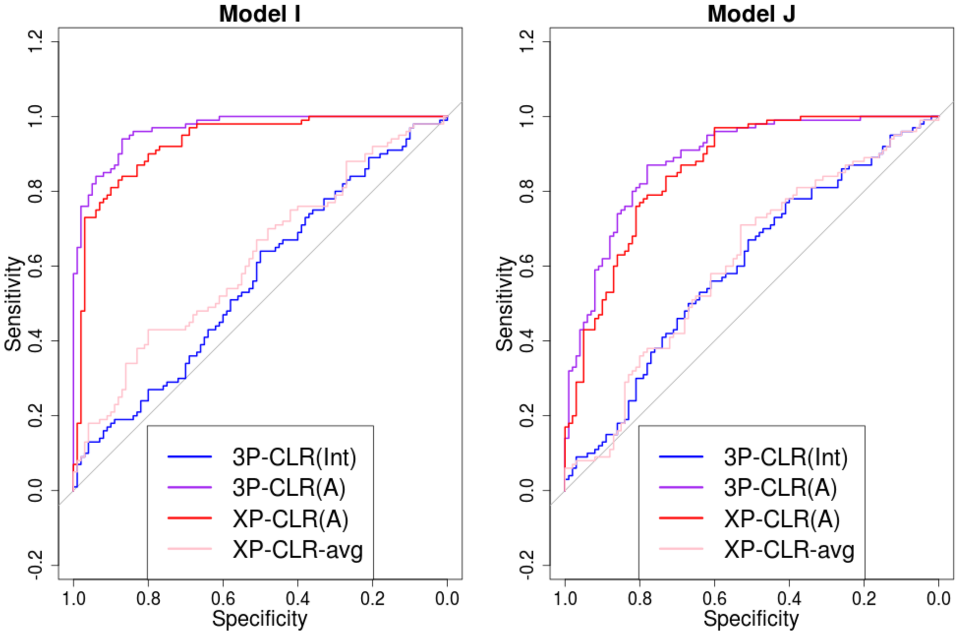


Figure 4.9. 3P-CLR(Int) is tailored to detect selective events that happened before the split t_{ab}, so it is largely insensitive to sweeps that occurred after the split. ROC curves show performance of 3P-CLR(Int) and two variants of XP-CLR for models where selection occurred in population a after its split from b.

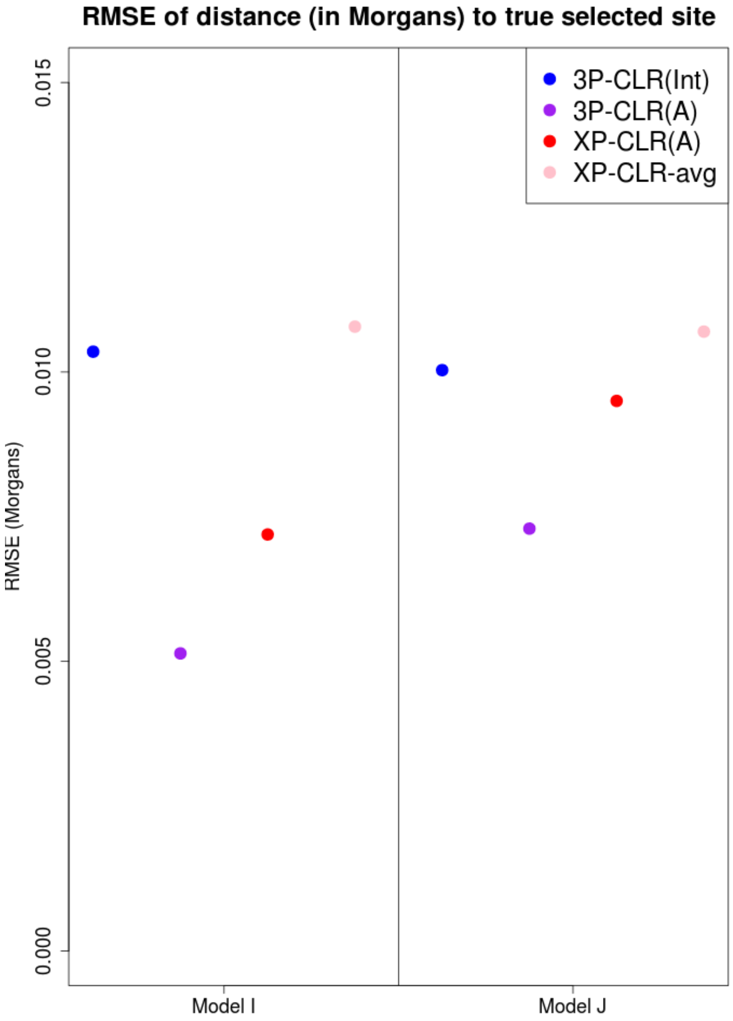


Figure 4.10. Root-mean squared error for the location of the sweep inferred by 3P-CLR(Int), 3P-CLR(A) and two variants of XP-CLR under different demographic scenarios, when the sweeps occurred in the terminal population branch leading to population a, after the split of populations a and b. In this case, the outgroup panel from population c contained 100 haploid genomes and the two sister population panels (from a and b) have 100 haploid genomes each.

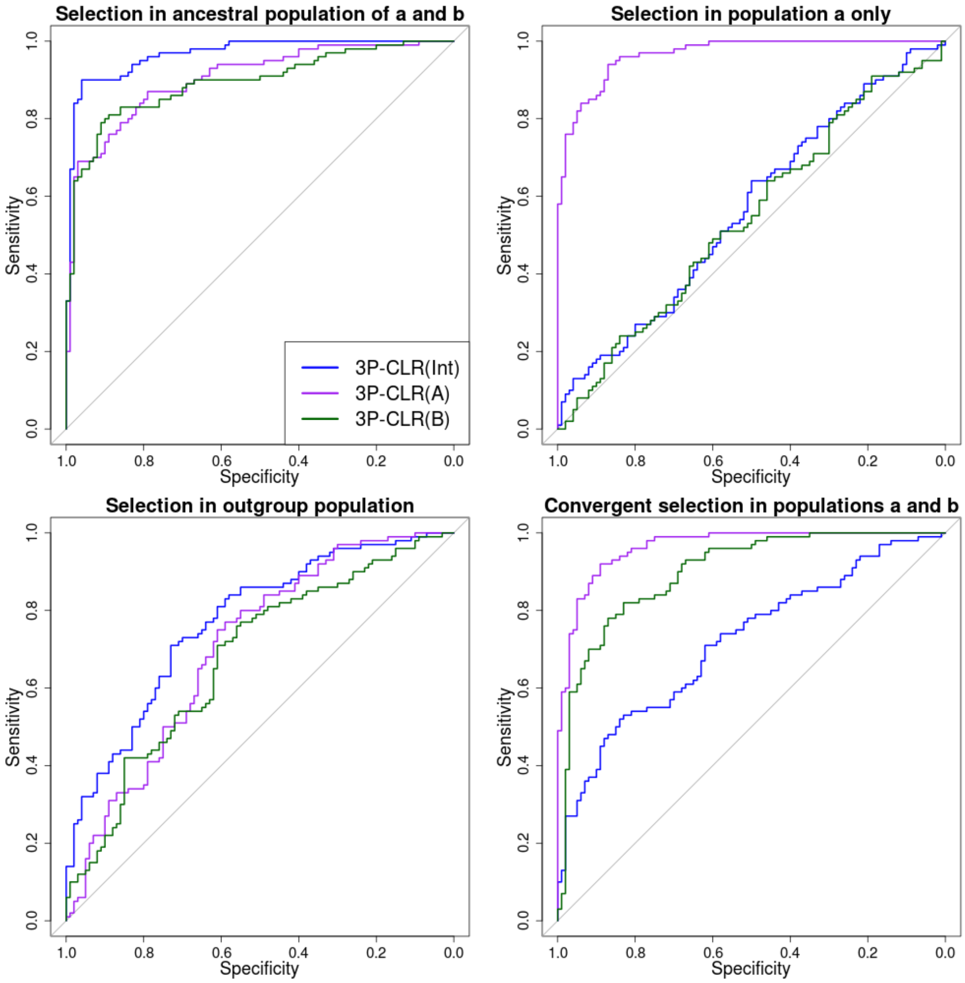


Figure 4.11. ROC curves for performance of 3P-CLR(Int), 3P-CLR(A) and 3P-CLR(B) when the selective events occur in different branches of the 3-population tree. Upper-left panel: Selection in the ancestral population of populations a and b. This is the type of events that 3P-CLR(Int) is designed to detect and, therefore, 3P-CLR(Int) is the most sensitive test in this case, though 3P-CLR(A) and 3P-CLR(B) show some sensitivity to these events too. Upper-right panel: Selection exclusive to population a. This is the type of events that 3P-CLR(A) is designed to detect, and it is therefore the best-performing statistic in that case, while 3P-CLR(B) and 3P-CLR(Int) are insensitive to selection. Lower-left panel: Selection in the outgroup population. In this case, none of the statistics seem very sensitive to the event, though 3P-CLR(Int) shows better relative sensitivity than the other two statistics. Lower-right panel: Independent selective events in populations a and b at the same locus. Here, both 3P-CLR(A) and 3P-CLR(B) perform best. In all cases, we used the split times and population sizes specified for Model C.

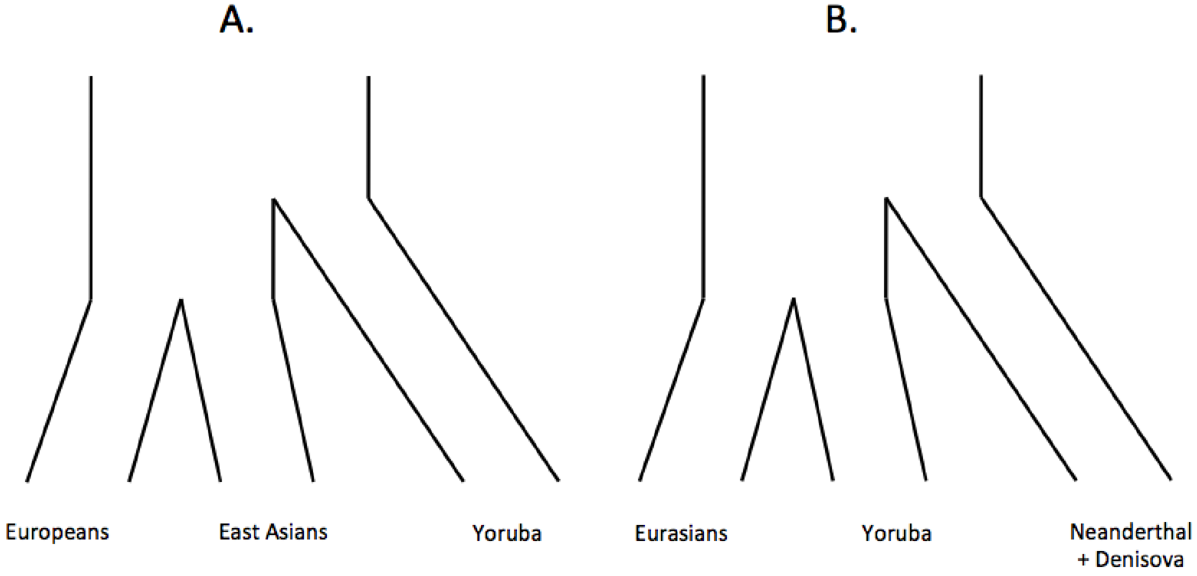


Figure 4.12. A. Three-population tree separating Europeans, East Asians and Yoruba. B. Three-population tree separating Eurasians, Yoruba and archaic humans (Neanderthal+Denisova).

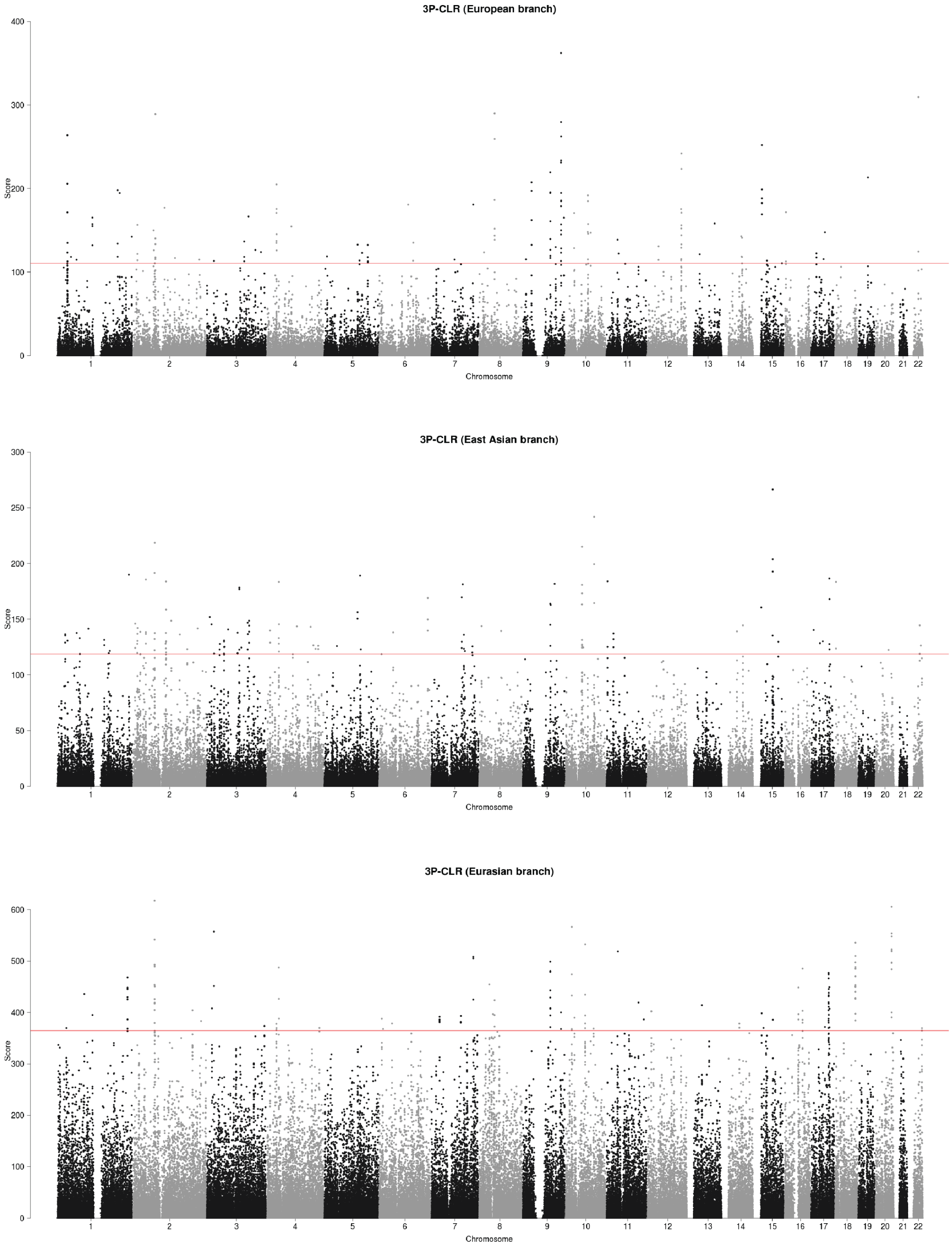


Figure 4.13. 3P-CLR scan of Europeans (upper panel), East Asians (middle panel) and the ancestral population to Europeans and East Asians (lower panel), using Yoruba as the outgroup in all 3 cases. The red line denotes the 99.9% quantile cutoff.

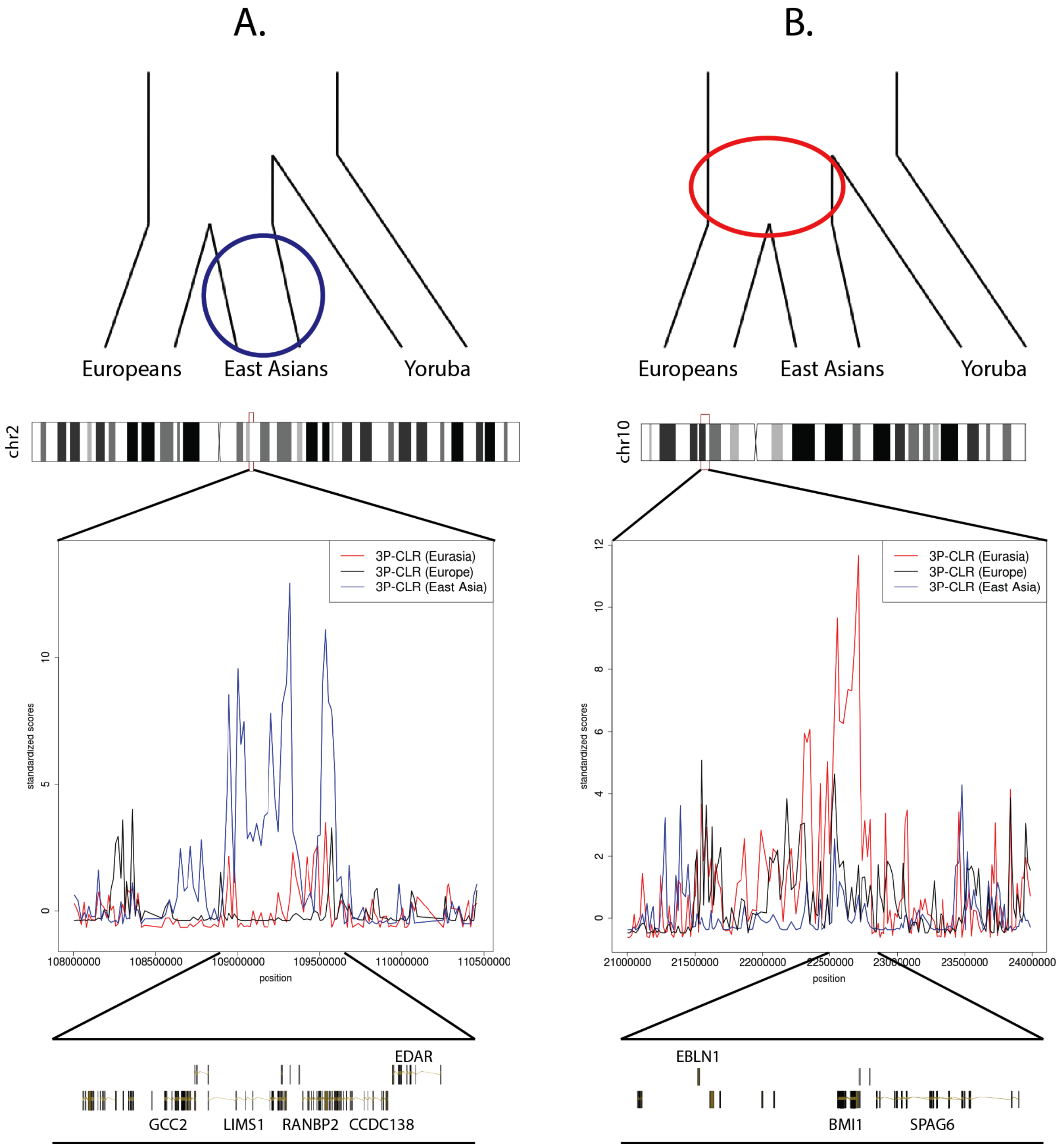


Figure 4.14. 3P-CLR scan of Europeans (black), East Asians (blue) and the ancestral Eurasian population (red) reveals regions under selection in different branches of the population tree. To make a fair comparison, all 3P-CLR scores were standardized by substracting the chromosome-wide mean from each window and dividing the resulting score by the chromosome-wide standard deviation. A) The region containing *EDAR* is a candidate for selection in the East Asian population. B) The region containing genes *SPAG6* and *BMI1* is a candidate for selection in the ancestral population of Europeans and East Asians. The image was built using the GenomeGraphs package in Bioconductor.

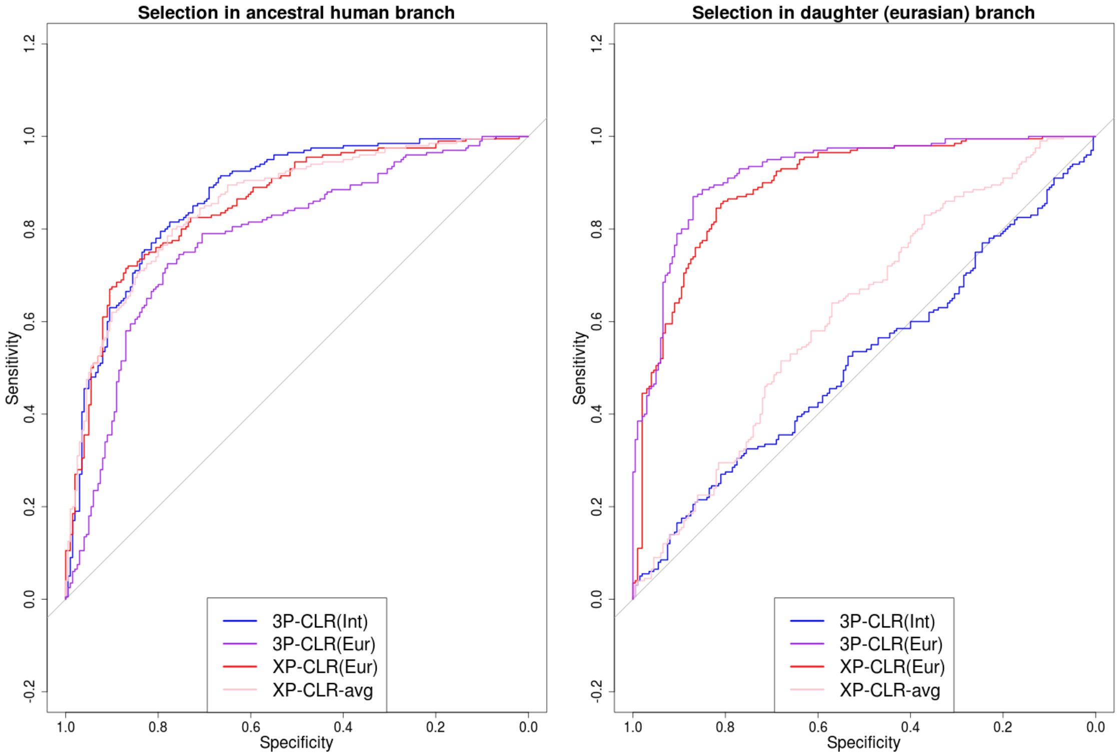


Figure 4.15. ROC curves for 3P-CLR run to detect selective events in the modern human ancestral branch, using simulations incorporating the history of population size changes and Neanderthal-to-Eurasian admixture inferred in Prüfer et al. (2014).

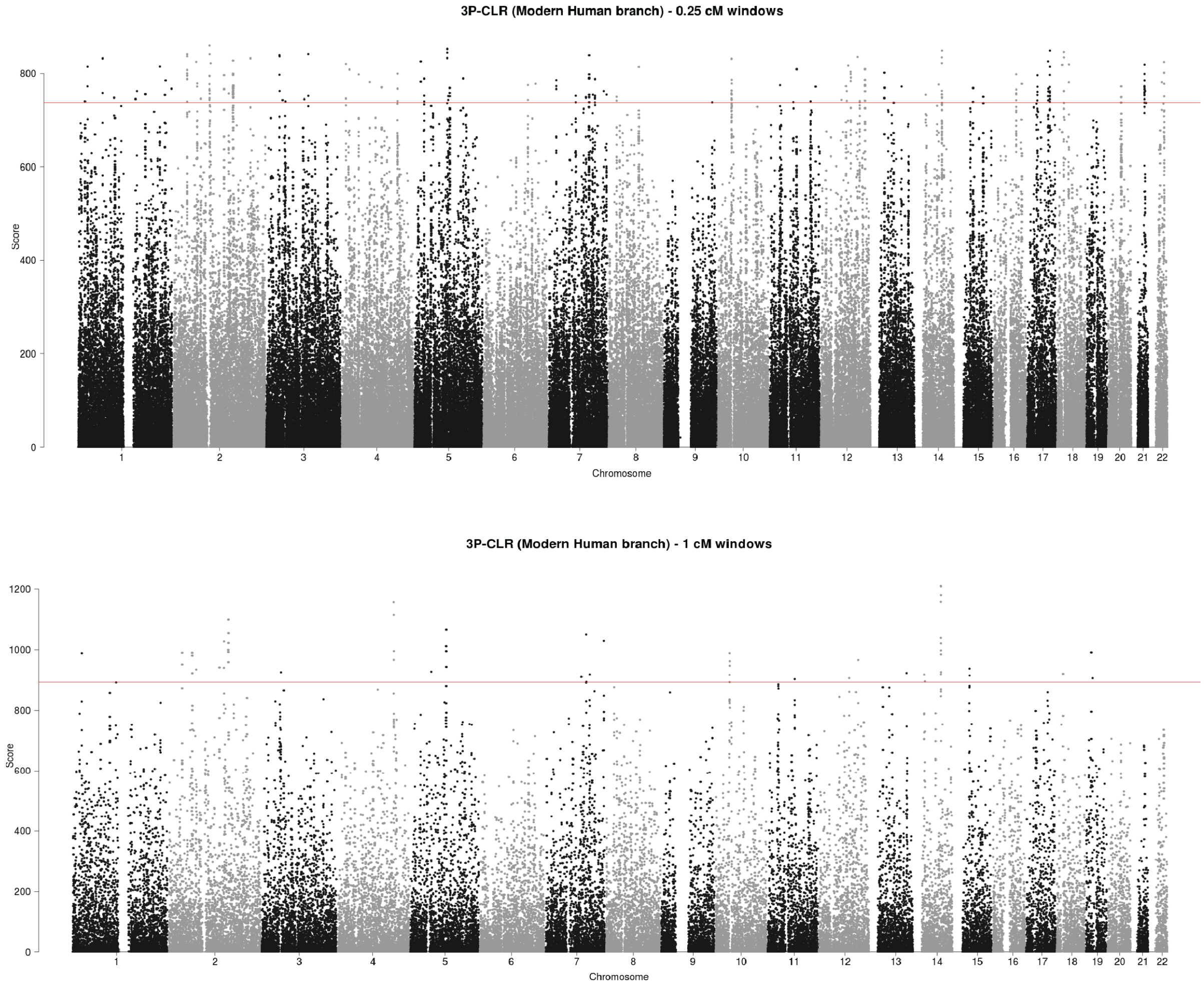


Figure 4.16. 3P-CLR scan of the ancestral branch to Yoruba and Eurasians, using the Denisovan and Neanderthal genomes as the outgroup. The red line denotes the 99.9% quantile cutoff. The top panel shows a run using 0.25 cM windows, each containing 100 SNPs, and sampling a candidate beneficial SNP every 10 SNPs. The bottom panels shows a run using 1 cM windows, each containing 200 SNPs, and sampling a candidate beneficial SNP every 40 SNPs.

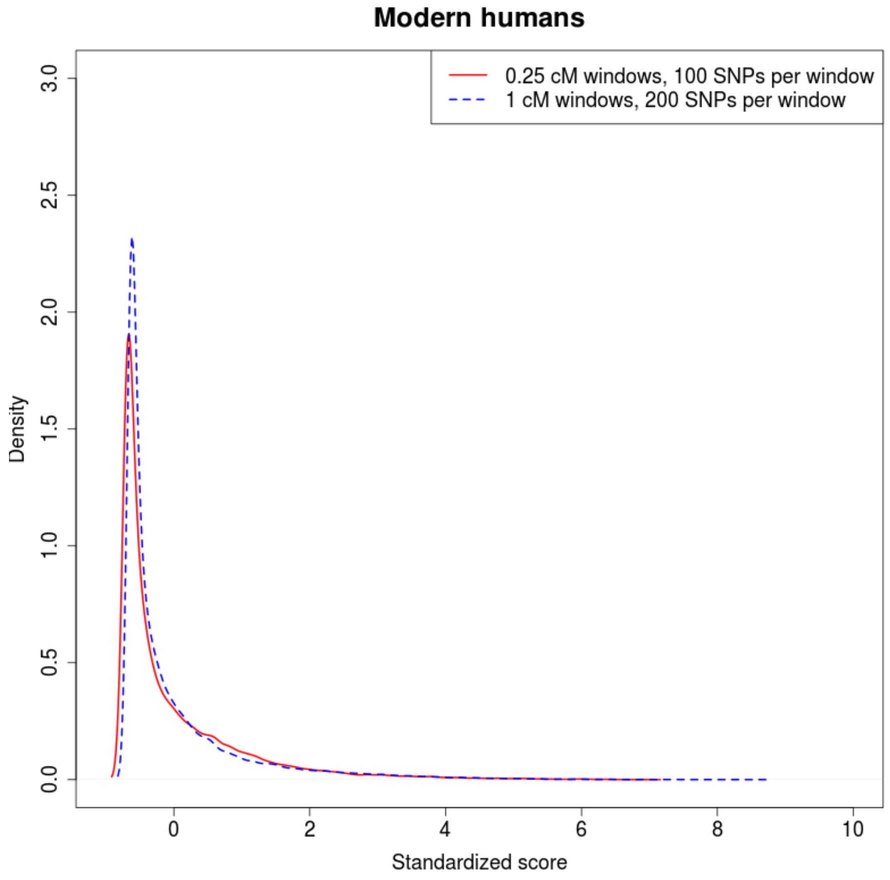


Figure 4.17. Comparison of 3P-CLR on the modern human ancestral branch under different window sizes and central SNP spacing. The red density is the density of standardized scores for 3P-CLR run using 0.25 cM windows, 100 SNPs per window and a spacing of 10 SNPs between each central SNP. The blue dashed density is the density of standardized scores for 3P-CLR run using 1 cM windows, 200 SNPs per window and a spacing of 40 SNPs between each central SNP.

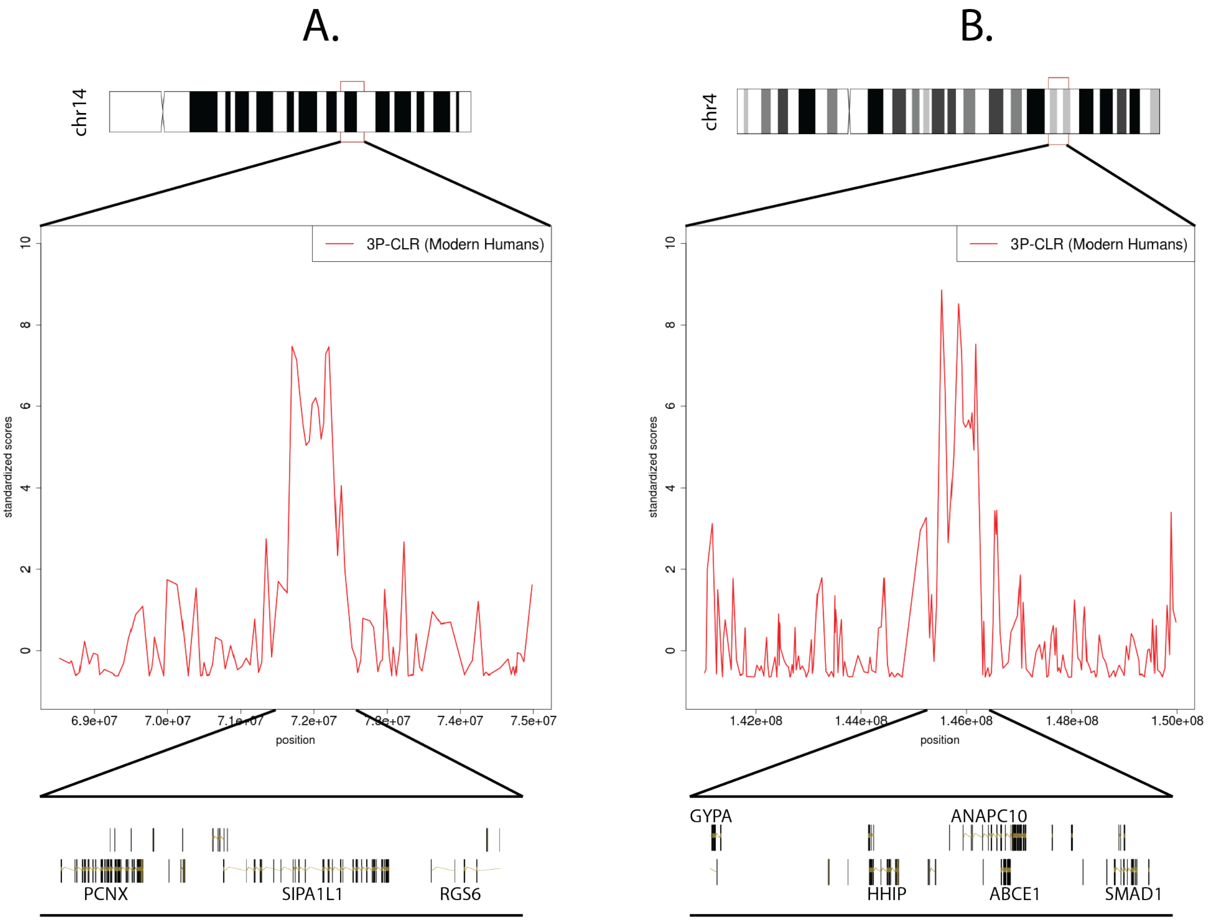


Figure 4.18. Two of the strongest candidates for selection in the modern human lineage, after the split from Neanderthal and Denisova. We show scores from the 1 cM scan, but the signals persist in the 0.25 cM scan. To make a fair comparison, all 3P-CLR scores were standardized by substracting the chromosome-wide mean from each window and dividing the resulting score by the chromosome-wide standard deviation. A) The region containing *SIPA1L1*. B) The region containing *ANAPC10*. The image was built using the GenomeGraphs package in Bioconductor.

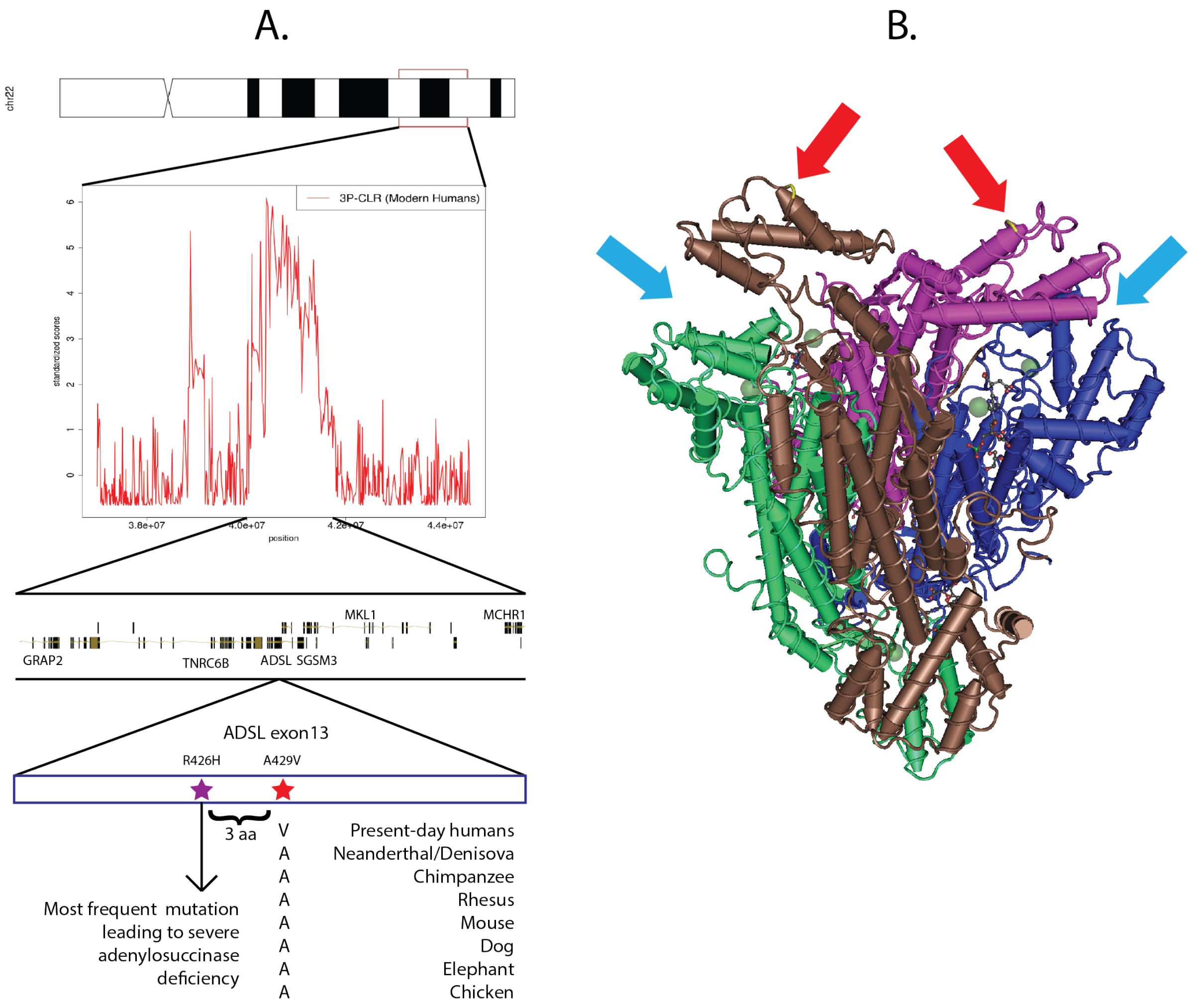


Figure 4.19. ADSL is a candidate for selection in the modern human lineage, after the split from Neanderthal and Denisova. A) One of the top-scoring regions when running 3P-CLR (0.25 cM windows) on the modern human lineage contains genes *TNRC6B*, *ADSL*, *MKL1*, *MCHR1*, *SGSM3* and *GRAP2*. The most disruptive nonsynonymous modern-human-specific change in the entire list of top regions is in an exon of *ADSL* and is fixed derived in all present-day humans but ancestral in archaic humans. It is highly conserved across tetrapods and lies only 3 residues away from the most common mutation leading to severe adenylosuccinase deficiency. B) The *ADSL* gene codes for a tetrameric protein. The mutation is in the C-terminal domain of each tetrameric unit (red arrows), which are near the active sites (light blue arrows). Scores in panel A were standardized using the chromosome-wide mean and standard deviation. Vertebrate alignments were obtained from the UCSC genome browser (Vertebrate Multiz Alignment and Conservation track) and the image was built using the GenomeGraphs package in Bioconductor and Cn3D.

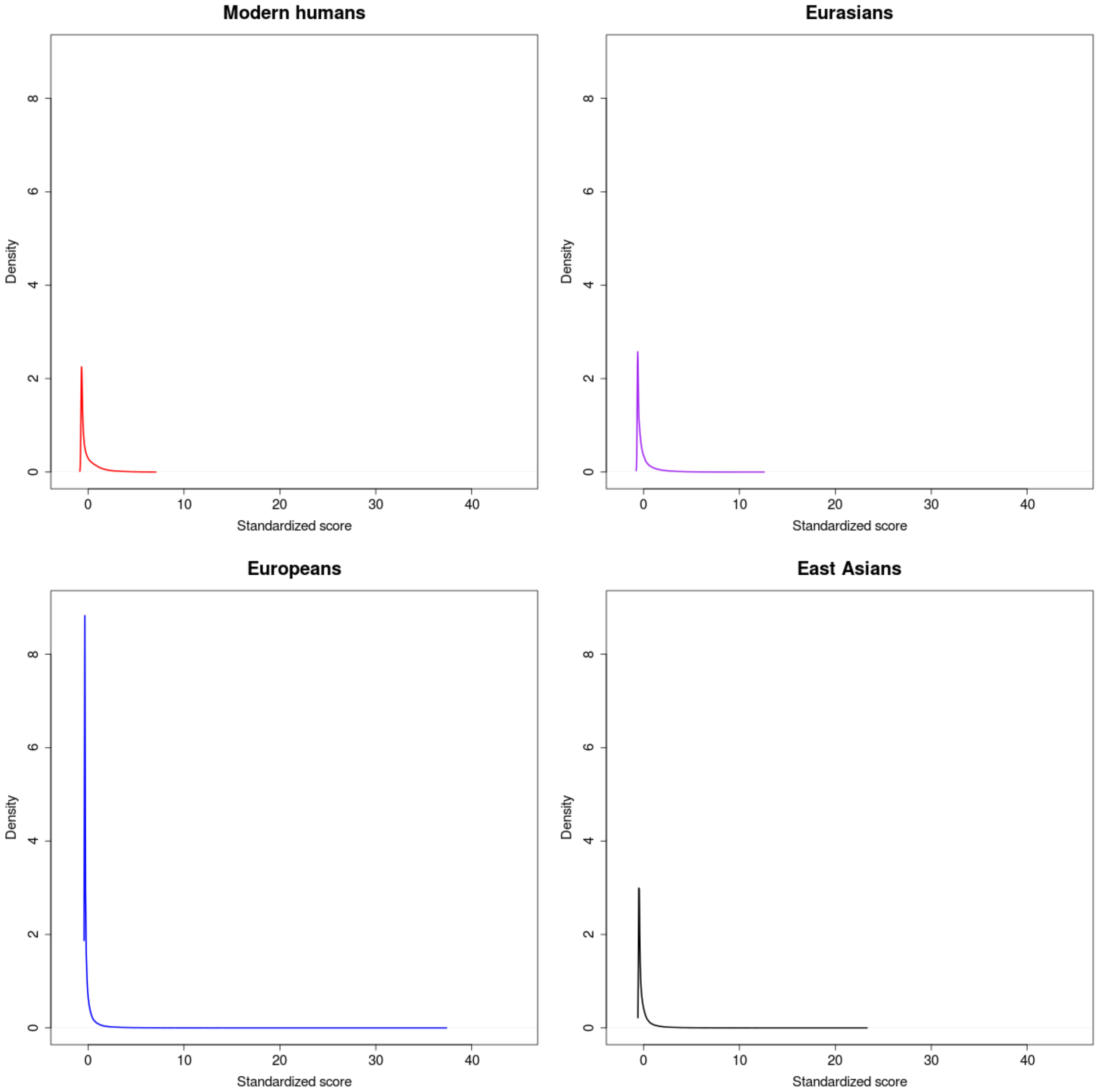


Figure 4.20. Genome-wide densities of each of the 3P-CLR scores described in this work. The distributions of scores testing for recent selection (Europeans and East Asians) have much longer tails than the distributions of scores testing for more ancient selection (Modern Humans and Eurasians). All scores were computed using 0.25 cM windows and were then standardized using their genome-wide means and standard deviations.

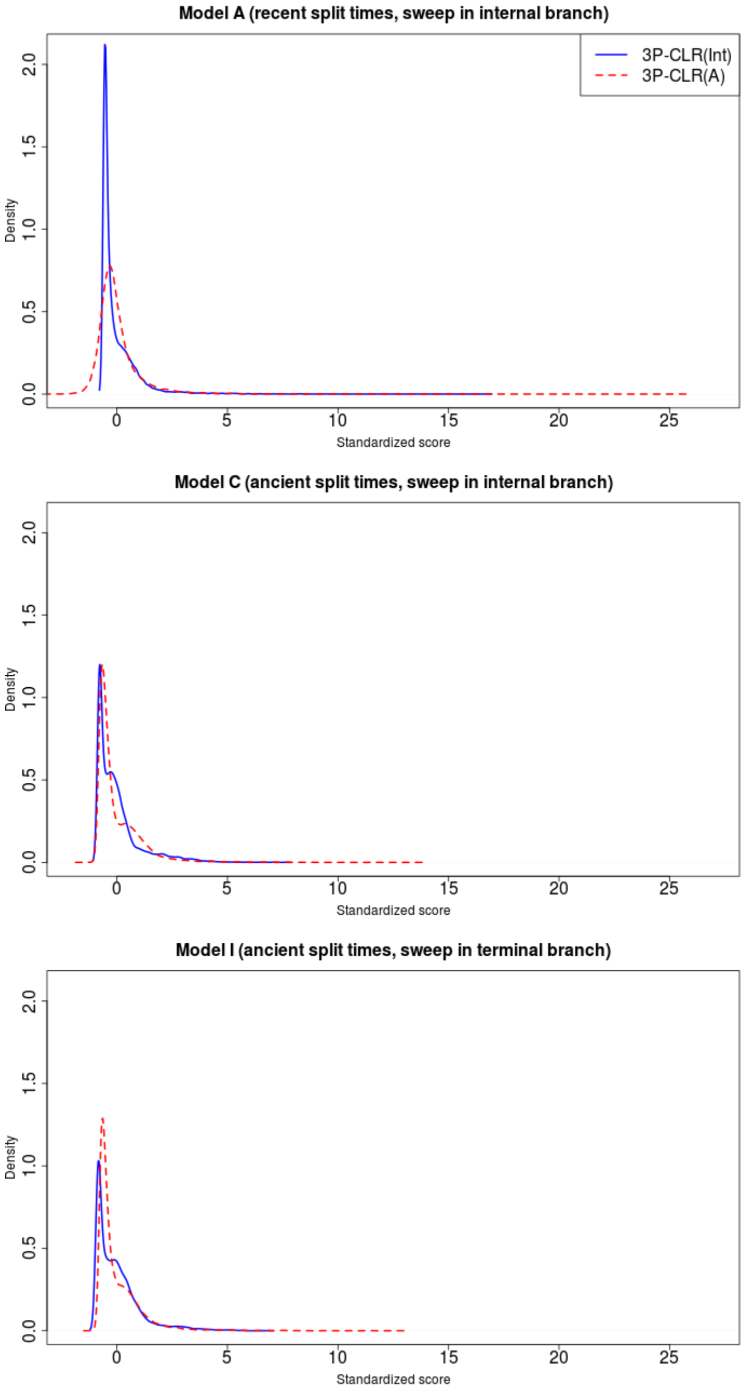


Figure 4.21. Distribution of 3P-CLR(Int) and 3P-CLR(A) scores under different demographic histories. We combined all scores obtained from 100 neutral simulations and 100 simulations with a selective sweep under different demographic and selection regimes. We then plotted the densities of the resulting scores. Top panel: Model A; Middle panel: Model C; Bottom panel: Model I. See Table 4.1 for details about each model.

5. Conclusion

The methods in this thesis are meant to serve in the advancement of paleogenomics towards a more rigorously quantitative field. The first method is geared as a first-pass analysis tool for researchers who have just sequenced and mapped an ancient hominin genome. We have implemented it in an easy-to-use C++ program that is freely downloadable and has a well-documented online manual. The other two sets of methods are meant for downstream analyses, especifically focusing on the detection of selected loci using ancient human data. These two are also applicable to non-human and non-ancient DNA data, in cases when biologists may be interested in detecting ancient selective events or loci under adaptive introgression in other organisms.

While distinct in many respects, the three methods share some similiarities. For example, the first and third methods heavily rely on diffusion theory and its approximations. In order to infer the demographic parameters, the contamination method uses either an exact formula for a multi-population site frequency spectrum [47] obtained via a Wright-Fisher diffusion equation, or a numerical approximation to such a formula (in the case of the three-population model) [49]. The third method uses an approximation to the Wright-Fisher diffusion dynamics which assumes that the variance in allele frequencies stays constant accross generations [154], which results in a Normal distribution (in the case of neutrality) or a mixture of two Normals (in the case of linked selection) [147].

Additionally, the second and third methods use previously inferred demographic parameters to better model the targeted selective events. The third method relies on a pre-computed population tree, and assumes no admixture among the populations of this tree. The second method incorporates admixture and requires a pre-computed admixture graph to be able to distinguish the outgroup, the source and the target of introgression, but relies on the use of summary statistics to distinguish selective events under different conditions.

All three of the methods could be improved by the incorporation of more complex models into their inference framework. For example, one could foresee developing a composite likelihood method like 3P-CLR that also allows for adaptive introgression between populations. Additionally, one could also expand DICE to infer additional parameters beyond admixture rates and drift times, like migration rates, bottlenecks and rates of population growth. As ancient DNA sequences were sampled in the past, they tend to display what is known as “branch shortening” - essentially missing mutations relative to present-day sequences. This phenomenon produces shorter branches when building phylogenetic trees with ancient genomes [2, 16]. In the future, it may also be possible to co-estimate the amount of branch shortening of a fossil, along with contamination and other demographic parameters. This may allow for the dating of fossils sampled beyond the temporal resolution of traditional dating methods - like radiocarbon dating.

The ongoing explosion of data in the field of paleogenomics will bring about many challenges for those tasked with interpreting and analyzing it. As the field grows, it will become imperative to use quantitative tools to extract meaningful historical and biological patterns from ancient genomes. The approaches presented in this thesis are largely based on previously developed population genetic and statistical theory. They should serve as a springboard for these quantitative approaches to become accessible and useful to the more empirically-driven field of ancient DNA. Thus, with a careful combination of theory and data, the path towards new insights about our evolutionary past will become ever easier to tread.

A. Appendix: Genotype probabilities conditional on a demography

Below we derive formulas 2.7, 2.8 and 2.9. Recall that we are interested in calculating the conditional probabilities $P[i|\mathbf{\Omega},\mathbf{O}] = P[i|y,\tau_C,\tau_A]$ for all three possibilities for the genotype in the ancient individual: $i = 0$, 1 or 2. These can be obtained from the definition of conditional probability. Let f_y^{DD} be the joint probability that a site has frequency y ($0 < y < 1$) in the contaminant panel and is homozygous for the derived allele in the ancient individual. Let f_y^{DA} be the joint probability that a site has frequency y in the contaminant panel and is heterozygous in the ancient individual. Finally, let f_y^{AA} be the joint probability that a site has frequency y in the anchor panel and is homozygous for the ancient allele in the ancient individual. Then:

$$P[\ i = 0 \mid y, \tau_C, \tau_A\ ] = \frac{f_y^{AA}}{f_y} = \frac{f_y^{AA}}{f_y^{AA} + f_y^{DA} + f_y^{DD}} \tag{A.1}$$

$$P[\ i = 1 \mid y, \tau_C, \tau_A\ ] = \frac{f_y^{DA}}{f_y} = \frac{f_y^{DA}}{f_y^{AA} + f_y^{DA} + f_y^{DD}} \tag{A.2}$$

$$P[\ i = 2 \mid y, \tau_C, \tau_A\ ] = \frac{f_y^{DD}}{f_y} = \frac{f_y^{DD}}{f_y^{AA} + f_y^{DA} + f_y^{DD}} \tag{A.3}$$

In the above expressions, the functions f depend on τ_C and τ_A, but we omit this conditioning for ease of notation. As can be seen, all we need to find is the joint probabilities f_y^{AA}, f_y^{DA} and f_y^{DD}. Here is where diffusion theory comes into play. Let $\phi(y,\tau|x,0)$ be the Kimura solution to the neutral forward diffusion equation in the absence of mutation [229], given a frequency x at time 0 and an elapsed drift time τ:

$$\phi(y,\tau|x,0) = 4x(1-x)\sum_{h=1}^{\infty} \frac{2j+1}{j(j+1)} C_{h-1}^{3/2}(1-2x) C_{h-1}^{3/2}(1-2y) e^{-j(j+1)\tau/2} \tag{A.4}$$

Here, x is the unknown population frequency of the derived allele in the ancestral population and $C_{h-1}^{(3/2)}(\bullet)$ is the Gegenbauer polynomial of order h-1 [230].

Assuming the ancestral population follows an equilibrium frequency distribution $g(x) = \theta/x$, we can write f_y^{DD} as follows:

$$f_y^{DD} = \int_0^1 \phi(y, \tau_C|x, 0)g(x) \left(\int_0^1 z^2 \phi(z, \tau_A|x, 0)dz \right) dx \tag{A.5}$$

where z is the unknown population frequency of a derived allele in the population to which the ancient individual belongs.

The expression in parentheses is the second moment of the transition density and its solution is known [144]:

$$\int_0^1 z^2 \phi(z, \tau_A|x, 0)dz = x - x(1-x)e^{-\tau_A} \tag{A.6}$$

This results in:

$$f_y^{DD} = \theta \int_0^1 \phi(y, \tau_C|x, 0)[1 - (1-x)e^{-\tau_A}]dx \tag{A.7}$$

$$f_y^{DD} = \theta \left[\int_0^1 \phi(y, \tau_C|x, 0)dx - e^{-\tau_A} \int_0^1 \phi(y, \tau_C|x, 0)dx + e^{-\tau_A} \int_0^1 x\ \phi(y, \tau_C|x, 0)dx \right] \tag{A.8}$$

The integral of the first two terms of the sum was solved in [47]:

$$\int_0^1 \phi(y, \tau_C|x, 0)dx = e^{-\tau_C} \tag{A.9}$$

The third term of the sum can be solved by noting that, though the integrand is an infinite sum (i.e. formula A.4 multiplied by x), only the integrals of the first two terms of that infinite sum are not equal to 0. This can be seen by integrating the parts of the terms of that infinite sum that depend on x:

$$\int_0^1 x^2(1-x)C_{h-1}^{(3/2)}(1-2x)dx = \begin{cases} 1/12 & h = 1 \\ -1/20 & h = 2 \\ 0 & h \geq 3 \end{cases}$$

Therefore, after integrating the first two terms of the infinite sum, we obtain:

$$\int_0^1 x\phi(y, \tau_C|x, 0)dx = \frac{1}{2}e^{-\tau_C} + \left(y - \frac{1}{2}\right) e^{-3\tau_C} \tag{A.10}$$

So we finally arrive at:

$$f_y^{DD} = \theta \left[e^{-\tau_C} - \frac{1}{2}e^{-\tau_A - \tau_C} + \left(y - \frac{1}{2}\right) e^{-\tau_A - 3\tau_C} \right] \tag{A.11}$$

We can obtain f_y^{DA} in a similar fashion:

$$f_y^{DA} = \int_0^1 \phi(y,\tau_C|x,0)g(x)\left(\int_0^1 2z(1-z)\phi(z,\tau_A|x,0)dz\right)dx \tag{A.12}$$

Solving the term in the parentheses:

$$\int_0^1 2z(1-z)\phi(z,\tau_A|x,0)dz = 2\left(\int_0^1 z\phi(z,\tau_A|x,0)dz - \int_0^1 z^2\phi(z,\tau_A|x,0)dz\right) \tag{A.13}$$

The first term of the difference is the first moment of the transition density, which is equal to x [144], while the second term is the second moment (formula A.6). Therefore:

$$f_y^{DA} = 2\theta e^{-\tau_A}\left[\int_0^1 \phi(y,\tau_C|x,0)(1-x)dx\right] \tag{A.14}$$

$$f_y^{DA} = 2\theta e^{-\tau_A}\left[\int_0^1 \phi(y,\tau_C|x,0)dx - \int_0^1 x\ \phi(y,\tau_C|x,0)\ dx\right] \tag{A.15}$$

And after using formulas A.9 and A.10, we obtain:

$$f_y^{DA} = \theta\left[e^{-\tau_A-\tau_C} + (1-2y)\,e^{-\tau_A-3\tau_C}\right] \tag{A.16}$$

To obtain f_y^{AA}, we know that, assuming the anchor population to be at equilibrium:

$$f_y = g(y) \tag{A.17}$$

And therefore:

$$f_y^{AA} + f_y^{DA} + f_y^{DD} = \frac{\theta}{y} \tag{A.18}$$

So we finally obtain:

$$f_y^{AA} = \theta\left[\frac{1}{y} - e^{-\tau_C} - \frac{1}{2}e^{-\tau_A-\tau_C} + \left(y-\frac{1}{2}\right)e^{-\tau_A-3\tau_C}\right] \tag{A.19}$$

We now have all the elements necessary to obtain the conditional probabilities from formulas A.1, A.2 and A.3, which immediately lead us to formulas 2.7, 2.8 and 2.9.

B. Appendix: Probabilistic inference using BAM files

Here, we briefly explain the way we infer fragment-specific error parameters in the optional BAM mode of DICE. Let $\mathbb{R}$ be the set of all fragments in the BAM file, and $R_j \in \mathbb{R}$ be a particular aligned fragment of length l. For fragment R_j, let $\{b_{j,1}, ..., b_{j,l}\}$ be the individuals nucleotides in the fragment. At each position of the fragment, there is a specific probability $\kappa_{j,i}$ that the base is erroneous. This probability is provided by the basecaller. Below, we will compute the likelihood of observing a base $b_{j,i} \in R_j$ under a bi-allelic model, given an error rate $\kappa_{j,i}$. Below, we focus on an individual fragment R_j and an individual position i on that fragment, so for simplicity, we drop the subscripts i and j and we let $b_{j,i} = b$ and $\kappa_{j,i} = \kappa$.

Let v be the base that was originally sampled at a given site, before deamination or mismapping. This base could be ancestral or derived. Let $P_{dam}[v \to b]$ be the probability of substitution from v to b due to post-mortem chemical damage. The probabilities of different types of damage (e.g. C→T or G→A) occurring at different positions of a fragment can be computed following [231] and [232], producing a matrix that can be provided to DICE as input. We offer the possibility of specifying different post-mortem damage matrices for the endogenous and the contaminant fragments.

Let E denote the event that a sequencing error has occurred, let D the event that chemical damage has occurred, let M be the event that R_j was correctly mapped and let $\neg$ denote the complement of an event (i.e. event has not occurred). We define the probability of observing sequenced base b given that no sequencing error has occurred at a position on a correctly mapped fragment that was originally v, by summing over two possibilities, either chemical damage occurred or it did not:

$$P[b|v, M, \neg E] = \mathbb{1}(v = b) \cdot P[\neg D] + (1 - \mathbb{1}(v = b)) \cdot P[D] \tag{B.1}$$

Here, $\mathbb{1}(v = b)$ is an indicator function that is equal to 1 if v is equal to b, and 0 otherwise. The probabilities $P[D]$ and $P[\neg D]$ are respectively equal to $P_{dam}[v \to b]$ and $1 - P_{dam}[v \to b]$.

Subsequently, we compute $P[b|v, M]$, the probability of observing b given v under the assumption that R_j was mapped at the correct genomic location. We have:

$$P[b|v, M] = (1 - \kappa) \cdot P[b|v, M, \neg E] + \kappa \cdot \frac{1}{2} \tag{B.2}$$

This is because if a sequencing error has occurred, the probability of observing b is independent of v, and therefore $P[b|v, M, E] = \frac{1}{2}$. Finally, let $P[M]$ be the probability that the fragment R_j is mapped at the correct location as given by the mapping quality. The probability of seeing b given that v was the base that was sampled before deamination is then:

$$P[b|v] = P[M] \cdot P[b|v, M] + P[\neg M] \cdot \frac{1}{2} \tag{B.3}$$

The probability of observing b given that the fragment was mismapped is independent of v, hence $P[b|v, \neg M] = \frac{1}{2}$. If either the base quality or mapping quality indicate a probability of error of 100%, $P[b|v]$ will be equal to $\frac{1}{2}$. These probabilities are used instead of the genome-wide error term ϵ in equations 2.4, 2.5 and 2.6. For instance, equation 2.4 for a specific base b in fragment R_j becomes:

$$\begin{aligned} q_2 = r_C(w \cdot P[b = der|v = der,\ contaminant] + \\ (1-w) \cdot P[b = der|v = anc,\ contaminant]) + \\ (1-r_C) \cdot P[b = der|v = der,\ ancient] \end{aligned} \tag{B.4}$$

Here, *der* is the derived base and *anc* is the ancestral base. In case different post-mortem damage matrices are provided by the user for the ancient and the contaminant fragments, the events *contaminant* and *ancient* serve to denote which damage probabilities (i.e. P_{dam}) should be used in each case.

Bibliography

[1] Adrian W Briggs et al. “Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA”. In: *Nucleic acids research* 38.6 (2010), e87–e87.

[2] Matthias Meyer et al. “A high-coverage genome sequence from an archaic Denisovan individual”. In: *Science* 338.6104 (2012), pp. 222–226.

[3] Marie-Theres Gansauge and Matthias Meyer. “Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA”. In: *Nature protocols* 8.4 (2013), pp. 737–748.

[4] Marie-Theres Gansauge and Matthias Meyer. “Selective enrichment of damaged DNA molecules for ancient genome sequencing”. In: *Genome research* 24.9 (2014), pp. 1543–1549.

[5] Jesse Dabney and Matthias Meyer. “Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries”. In: *Biotechniques* 52.2 (2012), pp. 87–94.

[6] Joseph K Pickrell and David Reich. “Toward a new history and geography of human genes informed by ancient DNA”. In: *Trends in Genetics* 30.9 (2014), pp. 377–389.

[7] Morten Rasmussen et al. “Ancient human genome sequence of an extinct Palaeo-Eskimo”. In: *Nature* 463.7282 (2010), pp. 757–762.

[8] Richard E Green et al. “A draft sequence of the Neandertal genome”. In: *Science* 328.5979 (2010), pp. 710–722.

[9] Iosif Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-day Europeans”. In: *Nature* 513.7518 (2014), pp. 409–413.

[10] Wolfgang Haak et al. “Massive migration from the steppe was a source for Indo-European languages in Europe”. In: *Nature* (2015).

[11] Iain Mathieson et al. “Genome-wide patterns of selection in 230 ancient Eurasians”. In: *Nature* 528.7583 (2015), pp. 499–503.

[12] Morten E Allentoft et al. “Population genomics of Bronze Age Eurasia”. In: *Nature* 522.7555 (2015), pp. 167–172.

[13] Michael Hofreiter et al. "DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA". In: *Nucleic acids research* 29.23 (2001), pp. 4793–4799.

[14] Richard E Green et al. "The Neandertal genome and ancient DNA authenticity". In: *The EMBO journal* 28.17 (2009), pp. 2494–2502.

[15] Susanna Sawyer et al. "Nuclear and mitochondrial DNA sequences from two Denisovan individuals". In: *Proceedings of the National Academy of Sciences* 112.51 (2015), pp. 15696–15700.

[16] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), pp. 43–49.

[17] Sriram Sankararaman et al. "The date of interbreeding between Neandertals and modern humans". In: *PLoS Genet* 8.10 (2012), e1002947.

[18] Jeffrey D Wall et al. "Higher levels of Neanderthal ancestry in East Asians than in Europeans". In: *Genetics* 194.1 (2013), pp. 199–209.

[19] Benjamin Vernot and Joshua M Akey. "Complex history of admixture between modern humans and Neandertals". In: *The American Journal of Human Genetics* 96.3 (2015), pp. 448–453.

[20] Sriram Sankararaman et al. "The genomic landscape of Neanderthal ancestry in present-day humans". In: *Nature* 507.7492 (2014), pp. 354–357.

[21] Benjamin Vernot and Joshua M Akey. "Resurrecting surviving Neandertal lineages from modern human genomes". In: *Science* 343.6174 (2014), pp. 1017–1021.

[22] Qiaomei Fu et al. "Genome sequence of a 45,000-year-old modern human from western Siberia". In: *Nature* 514.7523 (2014), pp. 445–449.

[23] Andaine Seguin-Orlando et al. "Genomic structure in Europeans dating back at least 36,200 years". In: *Science* 346.6213 (2014), pp. 1113–1118.

[24] Qiaomei Fu et al. "An early modern human from Romania with a recent Neanderthal ancestor". In: *Nature* (2015).

[25] John E Pool and Rasmus Nielsen. "Inference of historical changes in migration rate from the lengths of migrant tracts". In: *Genetics* 181.2 (2009), pp. 711–719.

[26] David Reich et al. "Genetic history of an archaic hominin group from Denisova Cave in Siberia". In: *Nature* 468.7327 (2010), pp. 1053–1060.

[27] Martin Kuhlwilm et al. "Ancient gene flow from early modern humans into Eastern Neanderthals". In: *Nature* (2016).

[28] Svante Pääbo. "The diverse origins of the human gene pool". In: *Nature Reviews Genetics* 16.6 (2015), pp. 313–314.

[29] Ivan Juric, Simon Aeschbacher, and Graham Coop. "The Strength of Selection Against Neanderthal Introgression". In: *bioRxiv* (2015), p. 030148.

[30] Kelley Harris and Rasmus Nielsen. “The genetic cost of Neanderthal introgression”. In: *bioRxiv* (2015), p. 030387.

[31] Fernando Racimo et al. “Evidence for archaic adaptive introgression in humans”. In: *Nature Reviews Genetics* 16.6 (2015), pp. 359–371.

[32] Xin Yi et al. “Sequencing of 50 human exomes reveals adaptation to high altitude”. In: *Science* 329.5987 (2010), pp. 75–78.

[33] Emilia Huerta-Sánchez et al. “Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA”. In: *Nature* 512.7513 (2014), pp. 194–197.

[34] Emilia Huerta-Sanchez and Fergal P Casey. “Archaic inheritance: supporting high altitude life in Tibet”. In: *Journal of Applied Physiology* (2015), jap–00322.

[35] Fernando L Mendez, Joseph C Watkins, and Michael F Hammer. “Neandertal origin of genetic variation at the cluster of OAS immunity genes”. In: *Molecular biology and evolution* 30.4 (2013), pp. 798–801.

[36] Qiliang Ding et al. “Neanderthal introgression at chromosome 3p21. 31 was under positive natural selection in East Asians”. In: *Molecular biology and evolution* (2013), mst260.

[37] Fernando Racimo, Martin Kuhlwilm, and Montgomery Slatkin. “A test for ancient selective sweeps and an application to candidate sites in modern humans”. In: *Molecular biology and evolution* 31.12 (2014), pp. 3344–3358.

[38] Richard E Green et al. “A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing”. In: *Cell* 134.3 (2008), pp. 416–426.

[39] Pontus Skoglund et al. “Accurate sex identification of ancient human remains using DNA shotgun sequencing”. In: *Journal of Archaeological Science* 40.12 (2013), pp. 4477–4482.

[40] Morten Rasmussen et al. “An Aboriginal Australian genome reveals separate human dispersals into Asia”. In: *Science* 334.6052 (2011), pp. 94–98.

[41] Thorfinn S Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. “ANGSD: analysis of next generation sequencing data”. In: *BMC Bioinformatics* 15.1 (2014), p. 356.

[42] Richard H Byrd et al. “A limited memory algorithm for bound constrained optimization”. In: *SIAM Journal on Scientific Computing* 16.5 (1995), pp. 1190–1208.

[43] Pontus Skoglund et al. “Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal”. In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pp. 2229–2234.

[44] Gabriel Renaud et al. “Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA”. In: *Genome biology* 16.1 (2015), pp. 1–18.

[45] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[46] Warren J Ewens. *Mathematical Population Genetics 1: I. Theoretical Introduction*. Vol. 27. Springer Science & Business Media, 2004.

[47] Hua Chen et al. "The joint allele-frequency spectrum in closely related species". In: *Genetics* 177.1 (2007), pp. 387–398.

[48] Gregory Ewing and Joachim Hermisson. "MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus". In: *Bioinformatics* 26.16 (2010), pp. 2064–2065.

[49] Ryan N Gutenkunst et al. "Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data". In: *PLoS Genetics* 5.10 (2009), e1000695.

[50] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The Annals of Applied Probability* 7.1 (1997), pp. 110–120.

[51] Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. "Context dependence, ancestral misidentification, and spurious signatures of natural selection". In: *Molecular Biology and Evolution* 24.8 (2007), pp. 1792–1800.

[52] Richard R Hudson. "Generating samples under a Wright–Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2 (2002), pp. 337–338.

[53] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760.

[54] Andrew Rambaut and Nicholas C Grass. "Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees". In: *Computer applications in the biosciences: CABIOS* 13.3 (1997), pp. 235–238.

[55] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA". In: *Journal of molecular evolution* 22.2 (1985), pp. 160–174.

[56] Gabriel Renaud et al. "freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers". In: *Bioinformatics* 29.9 (2013), pp. 1208–1209.

[57] Kate R Rosenbloom et al. "The UCSC Genome Browser database: 2015 update". In: *Nucleic acids research* 43.D1 (2015), pp. D670–D681.

[58] Mark Lipson et al. "Efficient moment-based inference of admixture parameters and sources of gene flow". In: *Molecular biology and evolution* 30.8 (2013), pp. 1788–1802.

[59] Tom Higham et al. "The timing and spatiotemporal patterning of Neanderthal disappearance". In: *Nature* 512.7514 (2014), pp. 306–309.

[60] Heng Li and Richard Durbin. “Inference of human population history from individual whole-genome sequences”. In: *Nature* 475.7357 (2011), pp. 493–496.

[61] Sergi Castellano et al. “Patterns of coding variation in the complete exomes of three Neandertals”. In: *Proceedings of the National Academy of Sciences* 111.18 (2014), pp. 6666–6671.

[62] Hua Chen. “The joint allele frequency spectrum of multiple populations: a coalescent theory approach”. In: *Theoretical Population Biology* 81.2 (2012), pp. 179–195.

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

[64] John A Kamm, Jonathan Terhorst, and Yun S Song. “Efficient computation of the joint sample frequency spectra for multiple populations”. In: *arXiv preprint arXiv:1503.01133* (2015).

[65] Morten Rasmussen et al. “The ancestry and affiliations of Kennewick Man”. In: *Nature* (2015).

[66] Maanasa Raghavan et al. “Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans”. In: *Nature* 505.7481 (2014), pp. 87–91.

[67] Cheng-Jun Hu et al. “Differential roles of hypoxia-inducible factor 1α (HIF-1α) and HIF-2α in hypoxic gene regulation”. In: *Molecular and cellular biology* 23.24 (2003), pp. 9361–9374.

[68] Eric Y Durand et al. “Testing for ancient admixture between closely related populations”. In: *Molecular biology and evolution* 28.8 (2011), pp. 2239–2252.

[69] Marcus R Kronforst et al. “Hybridization reveals the evolving genomic architecture of speciation”. In: *Cell reports* 5.3 (2013), pp. 666–677.

[70] Joel Smith and Marcus R Kronforst. “Do Heliconius butterfly species exchange mimicry alleles?” In: *Biology letters* 9.4 (2013), p. 20130503.

[71] Simon H Martin, John W Davey, and Chris D Jiggins. “Evaluating the use of ABBA–BABA statistics to locate introgressed loci”. In: *Molecular biology and evolution* 32.1 (2015), pp. 244–257.

[72] Michael DeGiorgio, Mattias Jakobsson, and Noah A Rosenberg. “Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa”. In: *Proceedings of the National Academy of Sciences* 106.38 (2009), pp. 16057–16062.

[73] Philipp W Messer. “SLiM: simulating evolution with selection and linkage”. In: *Genetics* 194.4 (2013), pp. 1037–1039.

[74] Tomaz Berisa and Joseph K Pickrell. “Approximately independent linkage disequilibrium blocks in human populations”. In: *Bioinformatics* 32.2 (2016), pp. 283–285.

[75] Nicholas H Barton. “The effect of hitch-hiking on neutral genealogies”. In: *Genetical Research* 72.02 (1998), pp. 123–133.

[76] Yuseob Kim and Wolfgang Stephan. “Detecting a local signature of genetic hitchhiking along a recombining chromosome”. In: *Genetics* 160.2 (2002), pp. 765–777.

[77] Yuseob Kim and Rasmus Nielsen. “Linkage disequilibrium as a signature of selective sweeps”. In: *Genetics* 167.3 (2004), pp. 1513–1524.

[78] Bernard Y Kim and Kirk E Lohmueller. “Selection and Reduced Population Size Cannot Explain Higher Amounts of Neandertal Ancestry in East Asian than in European Human Populations”. In: *The American Journal of Human Genetics* (2015).

[79] Pontus Skoglund and Mattias Jakobsson. “Archaic human ancestry in East Asia”. In: *Proceedings of the National Academy of Sciences* 108.45 (2011), pp. 18301–18306.

[80] Kim D Pruitt et al. “The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes”. In: *Genome research* 19.7 (2009), pp. 1316–1323.

[81] Amandine Vanhoutteghem and Philippe Djian. “Basonuclins 1 and 2, whose genes share a common origin, are proteins with widely different properties and functions”. In: *Proceedings of the National Academy of Sciences* 103.33 (2006), pp. 12423–12428.

[82] Leonie C Jacobs et al. “Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans”. In: *Human genetics* 132.2 (2013), pp. 147–158.

[83] Adriana Cabral et al. “Distinct functional interactions of human Skn-1 isoforms with Ese-1 during keratinocyte terminal differentiation”. In: *Journal of Biological Chemistry* 278.20 (2003), pp. 17792–17799.

[84] Hironobu Takemoto et al. “Relation between the expression levels of the POU transcription factors Skn-1a and Skn-1n and keratinocyte differentiation”. In: *Journal of dermatological science* 60.3 (2010), pp. 203–205.

[85] Martina Hašová et al. “Hyaluronan minimizes effects of UV irradiation on human keratinocytes”. In: *Archives of dermatological research* 303.4 (2011), pp. 277–284.

[86] Christina Spilker and Michael R Kreutz. “RapGAPs in brain: multipurpose players in neuronal Rap signalling”. In: *European Journal of Neuroscience* 32.1 (2010), pp. 1–9.

[87] Ganeshwaran H Mochida et al. “CHMP1A encodes an essential regulator of BMI1-INK4A in cerebellar development”. In: *Nature genetics* 44.11 (2012), pp. 1260–1264.

[88] Perrin M Wilson et al. “Astn2, a novel member of the astrotactin gene family, regulates the trafficking of ASTN1 during glial-guided neuronal migration”. In: *The Journal of Neuroscience* 30.25 (2010), pp. 8529–8540.

[89] Terry Vrijenhoek et al. “Recurrent CNVs disrupt three candidate genes in schizophrenia patients”. In: *The American Journal of Human Genetics* 83.4 (2008), pp. 504–510.

[90] Ke-Sheng Wang, Xue-Feng Liu, and Nagesh Aragam. "A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder". In: *Schizophrenia research* 124.1 (2010), pp. 192–199.

[91] Shuibin Lin et al. "Proteomic and functional analyses reveal the role of chromatin reader SFMBT1 in regulating epigenetic silencing and the myogenic gene program". In: *Journal of Biological Chemistry* 288.9 (2013), pp. 6238–6247.

[92] Takeo Kato et al. "Segmental copy number loss of SFMBT1 gene in elderly individuals with ventriculomegaly: a community-based study". In: *Internal Medicine* 50.4 (2011), pp. 297–303.

[93] MP Krause et al. "A novel GFP reporter mouse reveals Mustn1 expression in adult regenerating skeletal muscle, activated satellite cells and differentiating myoblasts". In: *Acta Physiologica* 208.2 (2013), pp. 180–190.

[94] Lizi Wu et al. "Transforming activity of MECT1-MAML2 fusion oncoprotein is mediated by constitutive CREB activation". In: *The EMBO journal* 24.13 (2005), pp. 2391–2402.

[95] Sey-En Lin et al. "Identification of new human mastermind proteins defines a family that consists of positive regulators for notch signaling". In: *Journal of Biological Chemistry* 277.52 (2002), pp. 50612–50620.

[96] Marta Winnes et al. "Frequent fusion of the CRTC1 and MAML2 genes in clear cell variants of cutaneous hidradenomas". In: *Genes, Chromosomes and Cancer* 46.6 (2007), pp. 559–563.

[97] Sarah Linéa Von Holstein et al. "CRTC1-MAML2 gene fusion in mucoepidermoid carcinoma of the lacrimal gland". In: *Oncology reports* 27.5 (2012), pp. 1413–1416.

[98] Emi Hamano et al. "Polymorphisms of interferon-inducible genes OAS-1 and MxA associated with SARS in the Vietnamese population". In: *Biochemical and biophysical research communications* 329.4 (2005), pp. 1234–1239.

[99] Jean K Lim et al. "Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man". In: (2009).

[100] S Knapp et al. "Polymorphisms in interferon-induced genes and the outcome of hepatitis C virus infection: roles of MxA, OAS-1 and PKR". In: *Genes and immunity* 4.6 (2003), pp. 411–419.

[101] María Fedetz et al. "OAS1 gene haplotype confers susceptibility to multiple sclerosis". In: *Tissue antigens* 68.5 (2006), pp. 446–449.

[102] Michael Dannemann, Aida M Andrés, and Janet Kelso. "Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors". In: *The American Journal of Human Genetics* 98.1 (2016), pp. 22–33.

[103] Matthieu Deschamps et al. “Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes”. In: *The American Journal of Human Genetics* 98.1 (2016), pp. 5–21.

[104] Shizuo Akira, Satoshi Uematsu, and Osamu Takeuchi. “Pathogen recognition and innate immunity”. In: *Cell* 124.4 (2006), pp. 783–801.

[105] Luis B Barreiro et al. “Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense”. In: *PLoS Genet* 5.7 (2009), e1000562.

[106] Valentina Gburcik et al. “An essential role for Tbx15 in the differentiation of brown and ï£¡ï£¡ï£¡briteï£¡ï£¡ï£¡ but not white adipocytes”. In: *American Journal of Physiology-Endocrinology and Metabolism* 303.8 (2012), E1053–E1060.

[107] Iris M Heid et al. “Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution”. In: *Nature genetics* 42.11 (2010), pp. 949–960.

[108] Ching-Ti Liu et al. “Multi-ethnic fine-mapping of 14 central adiposity loci”. In: *Human molecular genetics* (2014), ddu183.

[109] Ching-Ti Liu et al. “Genome-wide association of body fat distribution in African ancestry populations suggests new loci”. In: *PLoS Genet* 9.8 (2013), e1003681.

[110] Dmitry Shungin et al. “New genetic loci link adipose and insulin biology to body fat distribution”. In: *Nature* 518.7538 (2015), pp. 187–196.

[111] SL Candille et al. “Dorsoventral patterning of the mouse coat by Tbx15”. In: *PLoS biology* 2.1 (2004), E3–E3.

[112] Luisa F Pallares et al. “Mapping of Craniofacial Traits in Outbred Mice Identifies Major Developmental Genes Involved in Shape Determination”. In: *PLOS Genet* 11.11 (2015), e1005607.

[113] Ekkehart Lausch et al. “TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome”. In: *The American Journal of Human Genetics* 83.5 (2008), pp. 649–655.

[114] GA Curry. “Genetical and developmental studies on droopy-eared mice”. In: *Journal of embryology and experimental morphology* 7.1 (1959), pp. 39–65.

[115] Manvendra K Singh et al. “The T-box transcription factor Tbx15 is required for skeletal development”. In: *Mechanisms of development* 122.2 (2005), pp. 131–144.

[116] Matteo Fumagalli et al. “Greenlandic Inuit show genetic signatures of diet and climate adaptation”. In: *Science* 349.6254 (2015), pp. 1343–1347.

[117] Fernando Racimo et al. “Archaic adaptive introgression in TBX15/WARS2”. In: *bioRxiv* (2015), p. 033928.

[118] Trevor Woodage et al. “Characterization of the CHD family of proteins”. In: *Proceedings of The National Academy of Sciences* 94.21 (1997), pp. 11472–11477.

[119] Anita Rauch et al. “Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study”. In: *The Lancet* 380.9854 (2012), pp. 1674–1682.

[120] Gemma L Carvill et al. “Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1”. In: *Nature genetics* 45.7 (2013), pp. 825–830.

[121] Richard E Giles, Nobuyoshi Shimizu, and Frank H Ruddle. “Assignment of a human genetic locus to chromosome 5 which corrects the heat sensitive lesion associated with reduced leucyl-tRNA synthetase activity in ts025Cl Chinese hamster cells”. In: *Somatic cell genetics* 6.5 (1980), pp. 667–687.

[122] Jillian P Casey et al. “Identification of a mutation in LARS as a novel cause of infantile hepatopathy”. In: *Molecular genetics and metabolism* 106.3 (2012), pp. 351–358.

[123] Xue-Yuan Dong et al. “ATBF1 inhibits estrogen receptor (ER) function by selectively competing with AIB1 for binding to the ER in ER-positive breast cancer cells”. In: *Journal of Biological Chemistry* 285.43 (2010), pp. 32801–32809.

[124] Xiaodong Sun et al. “Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer”. In: *Nature genetics* 37.4 (2005), pp. 407–412.

[125] SIGMA Type 2 Diabetes Consortium et al. “Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico”. In: *Nature* 506.7486 (2014), pp. 97–101.

[126] H-Q Qu et al. “The association between the IFIH1 locus and type 1 diabetes”. In: *Diabetologia* 51.3 (2008), pp. 473–475.

[127] Siyang Liu et al. “IFIH1 polymorphisms are significantly associated with type 1 diabetes and IFIH1 gene expression in peripheral blood mononuclear cells”. In: *Human molecular genetics* 18.2 (2009), pp. 358–365.

[128] THOMAS G Warner et al. “Separation and characterization of the acid lipase and neutral esterases from human liver.” In: *American journal of human genetics* 32.6 (1980), p. 869.

[129] H Klima et al. “A splice junction mutation causes deletion of a 72-base exon from the mRNA for lysosomal acid lipase in a patient with cholesteryl ester storage disease.” In: *Journal of Clinical Investigation* 92.6 (1993), p. 2713.

[130] Charalampos Aslanidis et al. “Genetic and biochemical evidence that CESD and Wolman disease are distinguished by residual lysosomal acid lipase activity”. In: *Genomics* 33.1 (1996), pp. 85–93.

[131] Erik G Lund et al. “cDNA cloning of mouse and human cholesterol 25-hydroxylases, polytopic membrane proteins that synthesize a potent oxysterol regulator of lipid metabolism”. In: *Journal of Biological Chemistry* 273.51 (1998), pp. 34316–34327.

[132] Nobuto Shibata et al. “Association studies of cholesterol metabolism genes (CH25H, ABCA1 and CH24H) in Alzheimer’s disease”. In: *Neuroscience letters* 391.3 (2006), pp. 142–146.

[133] Su-Yang Liu et al. “Interferon-inducible cholesterol-25-hydroxylase broadly inhibits viral entry by production of 25-hydroxycholesterol”. In: *Immunity* 38.1 (2013), pp. 92–105.

[134] Tahereh Kamalati et al. “Brk, a breast tumor-derived non-receptor protein-tyrosine kinase, sensitizes mammary epithelial cells to epidermal growth factor”. In: *Journal of Biological Chemistry* 271.48 (1996), pp. 30956–30963.

[135] Sailesh Surapureddi et al. “Identification of a transcriptionally active peroxisome proliferator-activated receptor α-interacting cofactor complex in rat liver and characterization of PRIC285 as a coactivator”. In: *Proceedings of the National Academy of Sciences* 99.18 (2002), pp. 11836–11841.

[136] Takuya Tomaru et al. “Isolation and characterization of a transcriptional cofactor and its novel isoform that bind the deoxyribonucleic acid-binding domain of peroxisome proliferator-activated receptor-γ”. In: *Endocrinology* 147.1 (2006), pp. 377–388.

[137] Bengt Persson et al. “The SDR (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative”. In: *Chemico-biological interactions* 178.1 (2009), pp. 94–98.

[138] Donald S Wood et al. “Is Nebulin the Defective Gene Product in Duchenne Muscular Dystrophy?” In: *N Engl J Med* 1987.316 (1987), pp. 107–108.

[139] Vania Yotova et al. “An X-linked haplotype of Neandertal origin is present among all non-African populations”. In: *Molecular Biology and Evolution* 28.7 (2011), pp. 1957–1962.

[140] Benjamin F Voight et al. “A map of recent positive selection in the human genome”. In: *PLoS Biol* 4.3 (2006), e72.

[141] Joseph K Pickrell et al. “Signals of recent positive selection in a worldwide sample of human populations”. In: *Genome research* 19.5 (2009), pp. 826–837.

[142] Ekaterina E Khrameeva et al. “Neanderthal ancestry drives evolution of lipid catabolism in contemporary Europeans”. In: *Nature communications* 5 (2014).

[143] Philip W Hedrick. “Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation”. In: *Molecular ecology* 22.18 (2013), pp. 4606–4618.

[144] James F. Crow and Motoo Kimura. *An Introduction to population genetics theory.* New York, Evanston, London: Harper and Row, 1970.

[145] RC Lewontin. “The interaction of selection and linkage. I. General considerations; heterotic models”. In: *Genetics* 49.1 (1964), p. 49.

[146] WG Hill and Alan Robertson. “Linkage disequilibrium in finite populations”. In: *Theoretical and Applied Genetics* 38.6 (1968), pp. 226–231.

[147] Hua Chen, Nick Patterson, and David Reich. “Population differentiation as a test for selective sweeps”. In: *Genome research* 20.3 (2010), pp. 393–402.

[148] John Maynard Smith and John Haigh. “The hitch-hiking effect of a favourable gene”. In: *Genetical research* 23.01 (1974), pp. 23–35.

[149] RC Lewontin and Jesse Krakauer. “Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms”. In: *Genetics* 74.1 (1973), pp. 175–195.

[150] Joshua M Akey et al. “Interrogating a high-density SNP map for signatures of natural selection”. In: *Genome research* 12.12 (2002), pp. 1805–1814.

[151] Bruce S Weir et al. “Measures of human population structure show heterogeneity among genomic regions”. In: *Genome research* 15.11 (2005), pp. 1468–1476.

[152] Taras K Oleksyk et al. “Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations”. In: *PLoS One* 3.3 (2008), e1712.

[153] The 1000 Genomes Project Consortium. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), pp. 56–65.

[154] George Nicholson et al. “Assessing population differentiation and isolation from single-nucleotide polymorphism data”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4 (2002), pp. 695–715.

[155] Richard Durrett and Jason Schweinsberg. “Approximating selective sweeps”. In: *Theoretical population biology* 66.2 (2004), pp. 129–138.

[156] Justin C Fay and Chung-I Wu. “Hitchhiking under positive Darwinian selection”. In: *Genetics* 155.3 (2000), pp. 1405–1413.

[157] Bruce G Lindsay. “Composite likelihood methods”. In: *Contemporary Mathematics* 80.1 (1988), pp. 221–39.

[158] Cristiano Varin, Nancy Reid, and David Firth. “An overview of composite likelihood methods”. In: *Statistica Sinica* 21.1 (2011), pp. 5–42.

[159] Nick Patterson et al. “Ancient admixture in human history”. In: *Genetics* 192.3 (2012), pp. 1065–1093.

[160] Warren J Ewens. *Mathematical Population Genetics 1: Theoretical Introduction.* Vol. 27. Springer Science & Business Media, 2012.

[161] Joseph Felsenstein. “Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates”. In: *Evolution* (1981), pp. 1229–1242.

[162] Anjali G Hinch et al. “The landscape of recombination in African Americans”. In: *Nature* 476.7359 (2011), pp. 170–175.

[163] Akihiro Fujimoto et al. “A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness”. In: *Human Molecular Genetics* 17.6 (2008), pp. 835–843.

[164] Ryosuke Kimura et al. “A common variation in EDAR is a genetic determinant of shovel-shaped incisors”. In: *The American Journal of Human Genetics* 85.4 (2009), pp. 528–535.

[165] Pardis C Sabeti et al. “Genome-wide detection and characterization of positive selection in human populations”. In: *Nature* 449.7164 (2007), pp. 913–918.

[166] Sharon R Grossman et al. “A composite of multiple signals distinguishes causal variants in regions of positive selection”. In: *Science* 327.5967 (2010), pp. 883–886.

[167] Hifzur Rahman Siddique and Mohammad Saleem. “Role of BMI1, a stem cell factor, in cancer recurrence and chemoresistance: preclinical and clinical evidences”. In: *Stem Cells* 30.3 (2012), pp. 372–378.

[168] Rossana Sapiro et al. “Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6”. In: *Molecular and cellular biology* 22.17 (2002), pp. 6298–6305.

[169] Jay A White et al. “Identification of the human cytochrome P450, P450RAI-2, which is predominantly expressed in the adult cerebellum and is responsible for all-trans-retinoic acid metabolism”. In: *Proceedings of the National Academy of Sciences* 97.12 (2000), pp. 6403–6408.

[170] Ariel R Topletz et al. “Comparison of the function and expression of CYP26A1 and CYP26B1, the two retinoic acid hydroxylases”. In: *Biochemical pharmacology* 83.1 (2012), pp. 149–163.

[171] Hans Eiberg et al. “Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression”. In: *Human genetics* 123.2 (2008), pp. 177–187.

[172] Jiali Han et al. “A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation”. In: *PLoS genetics* 4.5 (2008), e1000074.

[173] Wojciech Branicki, Urszula Brudnik, and Anna Wojas-Pelc. “Interactions between HERC2, OCA2 and MC1R may influence human pigmentation phenotype”. In: *Annals of human genetics* 73.2 (2009), pp. 160–170.

[174] Iain Mathieson et al. “Eight thousand years of natural selection in Europe”. In: *bioRxiv* (2015), p. 016477.

[175] Ruth Halaban and Gisela Moellmann. “Murine and human b locus pigmentation genes encode a glycoprotein (gp75) with catalase activity.” In: *Proceedings of the National Academy of Sciences* 87.12 (1990), pp. 4809–4813.

[176] Eimear E Kenny et al. “Melanesian blond hair is caused by an amino acid change in TYRP1”. In: *Science* 336.6081 (2012), pp. 554–554.

[177] John A Todd et al. “Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes”. In: *Nature genetics* 39.7 (2007), pp. 857–864.

[178] Karen A Hunt et al. “Newly identified genetic risk variants for celiac disease related to the immune response”. In: *Nature genetics* 40.4 (2008), pp. 395–402.

[179] Alexandra Zhernakova et al. “Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection”. In: *The American Journal of Human Genetics* 86.6 (2010), pp. 970–977.

[180] Robert Kofler and Christian Schlötterer. “Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies”. In: *Bioinformatics* 28.15 (2012), pp. 2084–2085.

[181] Qingshen Gao et al. “The E6 oncoproteins of high-risk papillomaviruses bind to a novel putative GAP protein, E6TP1, and target it for degradation”. In: *Molecular and cellular biology* 19.1 (1999), pp. 733–744.

[182] Dimitrina D Pravtcheva and Thomas L Wise. “Disruption of Apc10/Doc1 in three alleles of oligosyndactylism”. In: *Genomics* 72.1 (2001), pp. 78–87.

[183] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[184] David Brawand et al. “The evolution of gene expression levels in mammalian organs”. In: *Nature* 478.7369 (2011), pp. 343–348.

[185] Martin Kircher et al. “A general framework for estimating the relative pathogenicity of human genetic variants”. In: *Nature genetics* 46.3 (2014), pp. 310–315.

[186] I Dunham et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pp. 57–74.

[187] Kate R Rosenbloom et al. “ENCODE whole-genome data in the UCSC Genome Browser: update 2012”. In: *Nucleic acids research* (2011), gkr1012.

[188] Peter N Robinson et al. “The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease”. In: *The American Journal of Human Genetics* 83.5 (2008), pp. 610–615.

[189] ML Van Keuren et al. “A somatic cell hybrid with a single human chromosome 22 corrects the defect in the CHO mutant (Ade–I) lacking adenylosuccinase activity”. In: *Cytogenetic and Genome Research* 44.2-3 (1987), pp. 142–147.

[190] Cyril Gitiaux et al. “Misleading behavioural phenotype with adenylosuccinate lyase deficiency”. In: *European Journal of Human Genetics* 17.1 (2009), pp. 133–136.

[191] Adam Siepel et al. “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome research* 15.8 (2005), pp. 1034–1050.

[192] Gregory M Cooper et al. “Single-nucleotide evolutionary constraint scores highlight disease-causing mutations”. In: *Nature methods* 7.4 (2010), pp. 250–251.

[193] Stanislav Kmoch et al. “Human adenylosuccinate lyase (ADSL), cloning and characterization of full-length cDNA and its isoform, gene structure and molecular basis for ADSL deficiency in six patients”. In: *Human molecular genetics* 9.10 (2000), pp. 1501–1513.

[194] PD Maaswinkel-Mooij et al. “Adenylosuccinase deficiency presenting with epilepsy in early infancy”. In: *Journal of inherited metabolic disease* 20.4 (1997), pp. 606–607.

[195] Sandrine Marie et al. “Mutation analysis in adenylosuccinate lyase deficiency: Eight novel mutations in the re-evaluated full ADSL coding sequence”. In: *Human mutation* 13.3 (1999), pp. 197–202.

[196] Valérie Race et al. “Clinical, biochemical and molecular genetic correlations in adenylosuccinate lyase deficiency”. In: *Human molecular genetics* 9.14 (2000), pp. 2159–2165.

[197] Patrick Edery et al. “Intrafamilial variability in the phenotypic expression of adenylosuccinate lyase deficiency: a report on three patients”. In: *American Journal of Medical Genetics Part A* 120.2 (2003), pp. 185–190.

[198] Gunter Meister et al. “Identification of novel argonaute-associated proteins”. In: *Current biology* 15.23 (2005), pp. 2149–2155.

[199] Kevin L Du et al. “Megakaryoblastic leukemia factor-1 transduces cytoskeletal signals and induces smooth muscle cell differentiation from undifferentiated embryonic stem cells”. In: *Journal of Biological Chemistry* 279.17 (2004), pp. 17578–17586.

[200] Thomas Mercher et al. “Involvement of a human gene related to the Drosophila spen gene in the recurrent t (1; 22) translocation of acute megakaryocytic leukemia”. In: *Proceedings of the National Academy of Sciences* 98.10 (2001), pp. 5776–5779.

[201] Meg Trahey et al. “Molecular cloning of two types of GAP complementary DNA from human placenta”. In: *Science* 242.4886 (1988), pp. 1697–1700.

[202] Eitan Friedman et al. “Nonsense mutations in the C–terminal SH2 region of the GTPase activating protein (GAP) gene in human tumours”. In: *Nature genetics* 5.3 (1993), pp. 242–247.

[203] Iiro Eerola et al. “Capillary malformation–arteriovenous malformation, a new clinical and genetic disorder caused by RASA1 mutations”. In: *The American Journal of Human Genetics* 73.6 (2003), pp. 1240–1249.

[204] D Hershkovitz et al. “RASA1 mutations may cause hereditary capillary malformations without arteriovenous malformations”. In: *British Journal of Dermatology* 158.5 (2008), pp. 1035–1040.

[205] Paul J Whiting et al. “Molecular and Functional Diversity of the Expanding GABA-A Receptor Gene Family”. In: *Annals of the New York Academy of Sciences* 868.1 (1999), pp. 645–653.

[206] Howard J Edenberg et al. “Variations in GABRA2, encoding the $\alpha 2$ subunit of the GABA A receptor, are associated with alcohol dependence and with brain oscillations”. In: *The American Journal of Human Genetics* 74.4 (2004), pp. 705–714.

[207] Julia Knabl et al. “Reversal of pathological pain through specific spinal GABAA receptor subtypes”. In: *Nature* 451.7176 (2008), pp. 330–334.

[208] Yun-Yan Xiang et al. “A GABAergic system in airway epithelium is essential for mucus overproduction in asthma”. In: *Nature medicine* 13.7 (2007), pp. 862–867.

[209] Francesca Ariani et al. “FOXG1 is responsible for the congenital variant of Rett syndrome”. In: *The American Journal of Human Genetics* 83.1 (2008), pp. 89–93.

[210] MA Mencarelli et al. “Novel FOXG1 mutations associated with the congenital variant of Rett syndrome”. In: *Journal of medical genetics* 47.1 (2010), pp. 49–53.

[211] Tetsushi Sadakata and Teiichi Furuichi. “Ca 2+-dependent activator protein for secretion 2 and autistic-like phenotypes”. In: *Neuroscience research* 67.3 (2010), pp. 197–202.

[212] Jessica L Crisci et al. “On characterizing adaptive events unique to modern humans”. In: *Genome biology and evolution* 3 (2011), pp. 791–798.

[213] Adilson Guilherme et al. “Role of EHD1 and EHBP1 in perinuclear sorting and insulin-regulated GLUT4 recycling in 3T3-L1 adipocytes”. In: *Journal of Biological Chemistry* 279.38 (2004), pp. 40062–40075.

[214] Julius Gudmundsson et al. “Common sequence variants on 2p15 and Xp11. 22 confer susceptibility to prostate cancer”. In: *Nature genetics* 40.3 (2008), pp. 281–283.

[215] Shiaoching Gong et al. “A gene expression atlas of the central nervous system based on bacterial artificial chromosomes”. In: *Nature* 425.6961 (2003), pp. 917–925.

[216] Len A Pennacchio et al. “In vivo enhancer analysis of human conserved non-coding sequences”. In: *Nature* 444.7118 (2006), pp. 499–502.

[217] Mulin Jun Li et al. “GWASdb: a database for human genetic variants identified by genome-wide association studies”. In: *Nucleic acids research* (2011), gkr1182.

[218] Danielle Welter et al. “The NHGRI GWAS Catalog, a curated resource of SNP-trait associations”. In: *Nucleic acids research* 42.D1 (2014), pp. D1001–D1006.

[219] Lavinia Paternoster et al. “Genome-wide population-based association study of extremely overweight young adults–the GOYA study”. In: *PLoS One* 6.9 (2011), e24303.

[220] Karsten Suhre et al. “A genome-wide association study of metabolic traits in human urine”. In: *Nature genetics* 43.6 (2011), pp. 565–569.

[221] Roy H Perlis et al. “Genome-wide association study of suicide attempts in mood disorder patients”. In: *Genome* 167.12 (2010).

[222] Marc Henrion et al. “Common variation at 2q22. 3 (ZEB2) influences the risk of renal cancer”. In: *Human molecular genetics* 22.4 (2013), pp. 825–831.

[223] Carina M Schlebusch et al. “Genomic variation in seven Khoe-San groups reveals adaptation and complex African history”. In: *Science* 338.6105 (2012), pp. 374–379.

[224] María Inés Fariello et al. “Detecting signatures of selection through haplotype differentiation among hierarchically structured populations”. In: *Genetics* 193.3 (2013), pp. 929–941.

[225] Pardis C Sabeti et al. “Detecting recent positive selection in the human genome from haplotype structure”. In: *Nature* 419.6909 (2002), pp. 832–837.

[226] Ryan D Hernandez et al. “Classic selective sweeps were rare in recent human evolution”. In: *science* 331.6019 (2011), pp. 920–924.

[227] Luigi Pace, Alessandra Salvan, and Nicola Sartori. “Adjusting composite likelihood ratio statistics”. In: *Statistica Sinica* 21.1 (2011), p. 129.

[228] Joseph K Pickrell and Jonathan K Pritchard. “Inference of population splits and mixtures from genome-wide allele frequency data”. In: *PLoS genetics* 8.11 (2012), e1002967.

[229] Motoo Kimura. “Solution of a process of random genetic drift with a continuous model”. In: *Proceedings of the National Academy of Sciences* 41.3 (1955), p. 144.

[230] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions*. Dover New York, 1965.

[231] Aurelien Ginolhac et al. “mapDamage: testing for damage patterns in ancient DNA sequences”. In: *Bioinformatics* 27.15 (2011), pp. 2153–2155.

[232] Hákon Jónsson et al. “mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters”. In: *Bioinformatics* 29.13 (2013), pp. 1682–1684.