Title
Probabilistic models and statistical inference in population genetics

Author
Spence, Jeffrey P.

Publication Date
2019

Peer reviewed|Thesis/dissertation

Probabilistic models and statistical inference in population genetics

by

Jeffrey P. Spence

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Abby Dernburg
Professor Steven Evans
Professor Michael Nachman
Professor Rasmus Nielsen

Spring 2019

Probabilistic models and statistical inference in population genetics

Abstract

Probabilistic models and statistical inference in population genetics

by

Jeffrey P. Spence

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Yun S. Song, Chair

Advances in sequencing and genotyping technologies have enabled data collection at unprecedented scales. This data deluge has driven demand for new, scalable methods and has allowed genomic approaches to be used to answer a broader set of questions. In this dissertation I analyze probabilistic models arising in population genetics, and I develop statistical techniques for extracting information from sequencing and genotyping data. I begin by analyzing Λ and Ξ coalescents, which arise as models of the genealogy of a sample from a population with large variance in the number of offspring per individual, such as marine species with sweepstakes-like reproduction or viruses undergoing continuous, strong selection. In particular, I show how to compute the expected value of a common summary statistic–the site frequency spectrum–under such models, and prove theorems about the identifiability of the model from the observed frequency spectrum. Such results suggest that it may be possible to learn about the underlying biological processes from observed site frequency spectra. I then present a method to find segments of Neanderthal ancestry in present-day humans, and use that method to learn about selective pressures on Neanderthal ancestry. I find that the observed patterns of Neanderthal ancestry are consistent with simple negative selection, as opposed to hybrid incompatibilities. Lastly, I develop a fast method to infer fine-scale recombination rates and apply it to 26 diverse human populations, elucidating the evolutionary dynamics and molecular modifiers of local recombination rates.

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation would not have been possible without the countless hours of assistance, advice, and support of friends and colleagues.

I am exceptionally grateful to my advisor, Yun Song. A Ph.D. advisor performs many roles–teacher, manager, advocate, and more–and Yun excelled at all of them. I learned more than I thought I ever could, and Yun's excitement about science was absolutely contagious. I always felt totally supported, allowing me to spend the time and take the risks necessary to truly grow as a person and scientist.

I am also heavily indebted to my labmates: Sanjit Batra, Anand Bhaskar, Nick Bhattacharya, Ma'ayan Bresler, Khanh Dao Duc, Dan Erdmann-Pham, Jonathan Fischer, Geno Guerra, Kelley Harris, Ethan Jewett, Michael Lim, Joyce Liu, Shishi Luo, Zvi Rosen, Neil Thomas, Miaoyan Wang, Yutong Wang, and Jane Yu. I am deeply fortunate to have been able to learn from, joke with, and talk to such intelligent, hilarious, and thoughtful people. I always loved being in lab and that was entirely a product of the people that were there. My co-authors in the lab through the years–Jeffrey Chan, Jack Kamm, Sara Mathieson, Matthias Steinrücken, and Jonathan Terhorst–played an oversized role in helping me. Jack Kamm and Matthias Steinrücken were the best mentors I could have ever hoped to work with, being both exceptionally talented scientists and abundantly kind human beings. I feel truly privileged to have had the opportunity to cross paths with such extraordinary individuals.

I had the the pleasure of being a member of the inaugural Computational Biology Ph.D. cohort. The members of my cohort made the transition to graduate school as smooth as possible, and were the best people with whom I could have imagined navigating an unfamiliar city, school, and program.

The support of my family was a bedrock on which I could always rely. My parents provided everything I have ever needed to succeed, and I owe them a debt I can never repay.

Above all I would like to thank my partner Amy, without whom life would just be less fun.

Chapter 1

Introduction

1.1 Background

The "Great Obsession" of population genetics is understanding the forces that generate, maintain, and shape genetic diversity [1]. These forces act at dramatically different scales, from the molecular, such as mutations caused by DNA damage, to the population-wide, such as natural selection acting on the genetic underpinning of traits involved in competition for mates or resources. By understanding the past and present forces shaping diversity, we can learn about the history of a population–such as population expansions, bottlenecks, or migrations–as well as about the way that genotypes affect phenotypes and how natural selection acts on those phenotypes.

While the field of population genetics has existed for about a hundred years, much of the work has been entirely theoretical and based on relatively simple models. Within the last few decades, however, the combination of ever-faster computers and rapid advances in DNA sequencing and genotyping technologies has allowed population geneticists to follow their Great Obsession as never before. With modern computers we are able to analyze more complicated, more realistic models that capture interesting facets of biology, and with modern genetic datasets we are able to test subtle differences between such models against real data. This exciting combination of advances has allowed us to probe subtle aspects of the forces that shape genetic variation, for example learning about mating patterns of many species, enabling the detection of Neanderthal ancestry in modern human individuals, and finding molecular mechanisms that modulate the rates of mutation and recombination.

Yet, even with the advances in computing technology, the most interesting population-genetic models are intractable. One approach is to apply black-box machine learning algorithms to data simulated under these complex models, e.g., [2, 3, 4, 5]. These approaches attempt to learn the inverse of the mapping from the parameters of interest in the model to observed data. This is certainly a promising approach, and is only beginning to be fully explored. The downside, however, is that these approaches are discriminative as opposed to generative, which can make them less robust, and they are fundamentally less interpretable

than model-based approaches.

Here I take the approach of analyzing simplifications of or approximations to these complicated models, while retaining enough of the complexity to learn about the phenomena of interest. This requires tools from probability theory, statistics, and optimization but allows for interpretable inference.

1.2 Outline

In this dissertation, I provide three case studies of the application of probabilistic modeling and statistical inference to population genetics.

In Chapter 2, I analyze time-inhomogeneous Λ and Ξ coalescents. Such models arise in populations where there is extreme variance in the number of offspring individuals have. For example, many marine species undergo sweepstakes-like reproduction where certain lucky individuals replace a sizable fraction of the population in a single spawn. This is in direct contrast to species where each individual has at most dozens of offspring over its lifetime. Continuous, strong positive selection, such as that experienced by pathogens, is another cause of massive variance in number offspring. These models have received relatively less attention, and thus there are open questions about even basic properties of summary statistics of data from these models. One important summary statistic is the site-frequency spectrum (SFS), and in Chapter 2, I develop a scheme to compute the expected value of the SFS. I also answer basic statistical questions about the SFS, such as whether or not certain aspects of the model are identifiable from the expected SFS. Without identifiability it is hopeless to learn about the underlying process from the data, and so understanding such limits is useful for practitioners.

In Chapter 3, I present a method, `dical-admix`, which calls tracts of Neanderthal ancestry in modern humans. Most non-African individuals have inherited 1-3% of their genome from Neanderthals, and `dical-admix` determines which specific parts of an individual's genome have been introgressed by using a simplification of the coalescent with recombination that may be cast as a hidden Markov model. I then use the patterns of these introgressed tracts to learn about the selective forces acting on segments of Neanderthal ancestry. There has been some evidence that Neanderthal ancestry has been removed from modern humans over time [6, 7, 8], but it is unclear if that is because of a higher mutational load in Neanderthals coupled with more efficient purifying selection in humans [7, 8] or if there were specific hybrid incompatibilities between Neanderthals and humans [6]. By analyzing the inferred patterns of introgression in modern humans, we find evidence in favor of the mutational load hypothesis.

Finally, in Chapter 4, I develop a method, `pyrho`, which is capable of quickly and accurately inferring fine-scale variation in recombination rates from polymorphism data. In humans, recombination rates have been known to vary across the genome by several orders of magnitude and this variation has been linked to a specific protein, PRDM9 [9, 10]. Yet, much about the causes of this fine-scale variation and the extent to which fine-scale recombination rates are conserved across populations is unknown. Previous methods for

inferring recombination maps have either required extremely large sample sizes or have been confounded by differences in demographic history, stymieing such cross-population analyses. `pyrho` explicitly accounts for differences in demographic history while still inferring accurate recombination maps from tens of individuals, disentangling the effect of demography on the inference of cross-population recombination rate variation. I apply `pyrho` to 26 diverse human populations and analyze the inferred recombination maps, finding that variation in the gene that encodes PRDM9 causes detectable differences in recombination rates across populations. Yet, by analyzing the polymorphism of PRDM9 binding sites, I find evidence that PRDM9 acts to erode its own binding motifs via gene conversion, but this effect is fairly weak. I also find that local chromatin structure–especially the repressive chromatin mark H3K27me3–plays an important role in modulating fine-scale recombination rates.

Chapter 2

The site frequency spectrum for general coalescents

This is joint work with John A. Kamm and Yun S. Song. Jere Koskela provided helpful discussion on convergence to the expected SFS. This work has been published in *Genetics* [11].

2.1 Introduction

When summarizing sequence data from n individuals, a natural and often-used statistic is the site frequency spectrum (SFS), $\hat{\boldsymbol{\tau}}_n = (\hat{\tau}_{n,1}, \ldots, \hat{\tau}_{n,n-1})^T$, where $\hat{\tau}_{n,k}$ is simply the number of sites at which k out of n individuals carry the mutant (or the derived) allele. Despite being only $n-1$ numbers, the SFS still contains a surprising amount of information about the history and structure of the population from which the individuals were sampled. Indeed, for neutrally evolving populations that are well-modeled by Kingman's coalescent [12], the expected value of the SFS was first computed for populations of constant size [13], extended to populations of variable size [14, 15, 16], and has since been used as a statistic for demographic inference in numerous studies (e.g. [17, 18, 19, 20, 21, 22, 23, 24]).

Yet, not all populations are well modeled by Kingman's coalescent. In fact, Kingman's coalescent can be viewed as a special case of a broader class of coalescent processes called Λ-coalscents [25, 26]. While Kingman's coalescent only permits pairwise mergers of lineages, Λ-coalescents allow two or more lineages to merge simultaneously in a single coalescence event. Such events arise when a single individual has many offspring [27, 28], under models of recurrent selective sweeps [29, 30], in populations undergoing continuous strong selection [31, 32], and in many other models. Λ-coalescents can further be seen as special cases of a broader class of coalescents called Ξ-coalescents [33]. In Ξ-coalescents, more than one merger event can occur simultaneously, resulting in simultaneous multiple mergers. While Ξ-coalescents have received less attention than Λ-coalescents in the literature, they still arise in certain models of selection [34], models of selective sweeps [30], models with repeated

strong bottlenecks [35], and for certain diploid mating models [36]. Also, since Ξ-coalecents generalize Λ-coalescents, any results presented about Ξ-coalescents immediately pertain to Λ-coalescents.

More formally, time-homogeneous Ξ-coalescents are governed by a measure $\Xi(d\boldsymbol{x})$ on the set $\{(x_1, x_2, \ldots) : x_1 \geq x_2 \geq \cdots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\}$. Furthermore, we will consider time-inhomogeneous Ξ-coalescents with measures that decompose into a time-independent part $\Xi(d\boldsymbol{x})$ and a strictly positive function $\zeta : \mathbb{R}_{\geq 0} \to \mathbb{R}_+$ of time, where $\zeta(t)$ represents (for historical reasons) the inverse intensity. That is, the coalescent is now governed by the measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$. For example, for Kingman's coalescent, $\Xi(d\boldsymbol{x}) = \delta_{\boldsymbol{0}}(d\boldsymbol{x})$, the point mass at zero, and $\zeta(t)$ corresponds to the scaled effective population size at time t. For other models, $\zeta(t)$ does not necessarily correspond to the population size, but has an interpretation specific to the model. For example, [31] show empirically that the rate of coalescence events in a model of continuous strong selection is a nonlinear function of the population size and the first two moments of the distribution of mutational effects. For a review of the mechanics of Λ-coalescents, see [25] and for a review of Ξ-coalescents, see [33]. For an alternative perspective, see [37] and [35] for a lookdown construction of particle systems with general reproduction mechanisms.

As mentioned above, the expected SFS for Kingman's coalescent is well understood, and can, in fact, be computed for an arbitrary ζ in $O(n^2)$ time [16]. For Λ- and Ξ-coalescents, however, the expected SFS can only be computed for constant ζ and the method for Λ-coalescents takes $O(n^4)$ time [38] and the method for Ξ-coalescents takes time exponential in n as a sum over partitions of the first n numbers must be performed [39]. Here we present a method that can compute the expected SFS for time-inhomogeneous Λ- and Ξ-coalescents with arbitrary ζ in $O(n^3)$ time. In the case where ζ is a constant function, our method can compute the expected SFS in $O(n^2)$ time given the rate matrix $\boldsymbol{Q}$ of the ancestral process, which will be defined more precisely below. We also prove some results about the sample size needed to make Λ identifiable for popular classes of Λ measures for constant ζ, as well as results about the sample size needed to make ζ identifiable for a fixed $\Xi(d\boldsymbol{x})$.

There has also been some related work on determining the asymptotic behavior of the expected SFS as $n \to \infty$. In this setting, [40, 41] derive some simple formulae for time-homogeneous Λ-coalescents that come down from infinity. For finite n, however, these asymptotic formulae can be rather inaccurate. Indeed, even for $n = 10,000$, [38] show that for some Λ-coalescents, there is a sizable discrepancy between the asymptotic formulae and the SFS obtained by simulation, highlighting the need for finite-sample calculations. Nevertheless, such asymptotic results highlight some interesting properties of Λ-coalescents and are reviewed in [42].

The remainder of this paper is organized as follows. We first present our main results about the computation of the SFS for time-inhomogeneous coalescents, and discuss the practical runtime of our implementation. We also investigate the variation in the empirical SFS and study the ability to infer the underlying model using the empirical SFS. Then, we prove some identifiability results about general coalescents. We conclude with discussion on

the implications of our results.

2.2 Main theoretical results on the expected SFS

Here we present our theoretical results on the expected SFS for a general Ξ-coalescent with a measure of the form $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$. These results lead to an $O(n^3)$-time algorithm for computing the expected SFS and can be improved to $O(n^2)$ if ζ is a constant function. Briefly, we use subsampling arguments to show that the expected SFS $\boldsymbol{\tau}_n = \mathbb{E}[\hat{\boldsymbol{\tau}}_n]$ can be computed from $\boldsymbol{a}_n := (\mathbb{E}T_2^{\mathrm{MRCA}}, \ldots, \mathbb{E}T_n^{\mathrm{MRCA}})^T$, where $\mathbb{E}T_k^{\mathrm{MRCA}}$ denotes the expected time to the most recent common ancestor for sample size $k \in \{2, \ldots, n\}$. Then, we show how to compute $\boldsymbol{a}_n$ using a spectral decomposition of the rate matrix $\boldsymbol{Q}$ of the ancestral process (also known as the block-counting process) of the time-homogeneous coalescent corresponding to $\Xi(d\boldsymbol{x})$. More specifically, $\boldsymbol{Q}$ is a lower triangular matrix where $(\boldsymbol{Q})_{ij}$ is the instantaneous rate at which i unlabeled lineages merge to form j unlabeled lineages when $\zeta \equiv 1$. For example, for Kingman's coalescent,

$$(\boldsymbol{Q})_{ij} = \begin{cases} \binom{i}{2}, & j = i-1, \\ -\binom{i}{2}, & j = i, \\ 0, & \text{otherwise.} \end{cases}$$

Using this notation, we are now ready to state our main result. The rest of the section will then provide lemmas which contain formulae for the matrices in Theorem 1, as well as a proof of those lemmas and Theorem 1.

Theorem 1. *Consider an arbitrary time-inhomogeneous Ξ-coalescent governed by a measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$, such that the expected time $c_{k,k}$ to the first coalescence for a sample of size k is finite for $k \in \{2, \ldots, n\}$. Let $\boldsymbol{c}_n = (c_{2,2}, \ldots, c_{n,n})^T$. Then, there exists a universal matrix $\boldsymbol{A} \in \mathbb{R}^{n-1 \times n-1}$ that does not depend on the measure and a matrix $\mathbf{L} \in \mathbb{R}^{n-1 \times n-1}$ that depends on Ξ but not ζ, such that*

$$\boldsymbol{\tau}_n = \frac{\theta}{2}\mathbf{A}\boldsymbol{a}_n \quad \text{and} \quad \boldsymbol{a}_n = \mathbf{L}\boldsymbol{c}_n,$$

where $\frac{\theta}{2}$ is the population-scaled mutation rate. Furthermore, this allows $\boldsymbol{\tau}_n$ to be computed in $O(n^3)$ time.

Computing the matrix $\mathbf{L}$ in Theorem 1 is costly. For time-homogeneous coalescents, it is possible to compute $\boldsymbol{a}_n$ directly, resulting in the following corollary:

Corollary 1. *In the same setting as Theorem 1, if ζ is a constant function, then $\boldsymbol{\tau}_n$ can be computed in $O(n^2)$ time.*

In what follows, Lemmas 1 and 2 provide formulae to compute the universal matrix $\boldsymbol{A}$, while Lemmas 3 and 4 provide formulae to compute $\mathbf{L}$, which is related to the spectral

decomposition of the rate matrix $\boldsymbol{Q}$. The expected first coalescence times $\boldsymbol{c}_n = (c_{2,2}, \ldots, c_{n,n})^T$ can be computed as [16, 18]

$$c_{k,k} = \int_0^\infty \mathbb{P}\left\{\text{time of first coalescence for } k \text{ individuals} > t\right\} dt = \int_0^\infty e^{(\boldsymbol{Q})_{kk} \int_0^t \frac{1}{\zeta(s)} ds} dt.$$

Note that since $\boldsymbol{A}$ and $\mathbf{L}$ do not depend on ζ, the SFS depends on time and the inhomogeneity of the coalescent process only through the first coalescence times $\boldsymbol{c}_n$.

Lemma 1. *Let $\boldsymbol{\gamma}_n := (\tau_{2,1}, \tau_{3,2}, \ldots, \tau_{n,n-1})^T$ denote the anti-singleton entries (i.e., entries where exactly one individual has the ancestral allele and all other individuals have the derived allele) of the SFS for samples of sizes $2, \ldots, n$. Then,*

$$\boldsymbol{\tau}_n = \mathbf{B}\boldsymbol{\gamma}_n,$$

where the entries of $\mathbf{B} \in \mathbb{R}^{n-1 \times n-1}$ are given by

$$(\mathbf{B})_{ij} = \begin{cases} (-1)^{i-j} \frac{1}{j+1} \binom{n-i-1}{j-i} \binom{n}{i}, & i \leq j, \\ 0, & i > j. \end{cases}$$

Proof. We use induction to show that

$$\tau_{n,i} = \sum_{j=i}^{n-1} (-1)^{i-j} \frac{1}{j+1} \binom{n-i-1}{j-i} \binom{n}{i} \tau_{j+1,j}. \tag{2.1}$$

Using exchangeability and a subsampling argument similar to that of [24, Lemma 2], we obtain, for $k > l+1$,

$$\tau_{k-1,l} = \frac{l+1}{k} \tau_{k,l+1} + \frac{k-l}{k} \tau_{k,l}, \tag{2.2}$$

which follows from removing an individual uniformly at random from a sample of size k. Now, define the *level* of $\tau_{n,i}$ as $n-i$ and note that (2.1) holds for level 1, i.e., for $\tau_{l,l-1}$ on the left

hand side. Assume that (2.1) holds for level $n-i-1$. Then,

$$\begin{aligned}
\tau_{n,i} &= \frac{n}{n-i}\tau_{n-1,i} - \frac{i+1}{n-i}\tau_{n,i+1} \\
&= \frac{n}{n-i}\left[\sum_{j=i}^{n-2}(-1)^{i-j}\frac{1}{j+1}\binom{n-i-2}{j-i}\binom{n-1}{i}\tau_{j+1,j}\right] \\
&\quad - \frac{i+1}{n-i}\left[\sum_{j=i+1}^{n-1}(-1)^{i+1-j}\frac{1}{j+1}\binom{n-i-2}{j-i-1}\binom{n}{i+1}\tau_{j+1,j}\right] \\
&= \binom{n}{i}\left\{\frac{1}{i+1}\tau_{i+1,i} + (-1)^{n-1-i}\frac{1}{n}\tau_{n,n-1}\right. \\
&\qquad \left. + \sum_{j=i+1}^{n-2}(-1)^{i-j}\frac{1}{j+1}\left[\binom{n-i-2}{j-i} + \binom{n-i-2}{j-i-1}\right]\tau_{j+1,j}\right\} \\
&= \binom{n}{i}\sum_{j=i}^{n-1}(-1)^{j-i}\frac{1}{j+1}\binom{n-i-1}{j-i}\tau_{j+1,j},
\end{aligned}$$

where the first equality holds by the recursion (2.2) and the second equality holds by the inductive hypothesis, by noting that $\tau_{n-1,i}$ and $\tau_{n,i+1}$ are both one level below $\tau_{n,i}$. □

The following lemma relates $\boldsymbol{\gamma}_n$ to $\boldsymbol{a}_n$:

Lemma 2. *Let $\boldsymbol{\gamma}_n$, $\boldsymbol{a}_n$, and θ be defined as above. Then,*

$$\boldsymbol{\gamma}_n = \frac{\theta}{2}\mathbf{C}\boldsymbol{a}_n,$$

where $\mathbf{C} \in \mathbb{R}^{n-1\times n-1}$ is bi-diagonal with $(\mathbf{C})_{k,k-1} = -(k+1)$ and $(\mathbf{C})_{kk} = k+1$ for $k \in \{2,\ldots,n-1\}$, and $(\mathbf{C})_{11} = 2$.

Proof. As in the proof of Lemma 1, we employ a subsampling argument. Consider a sample of size $k+1$. The only way that a subsample of size k can have a different time to most recent common ancestor is if the removed individual is a singleton after all of the other lineages have coalesced. The probability that we remove that singleton to form our subsample is $\frac{1}{k+1}$. Then, the expected amount of time during which there is one singleton and all of the other individuals have coalesced scaled by the mutation rate is exactly the anti-singleton entry. Thus,

$$\frac{1}{k+1}\tau_{k+1,k} = \frac{\theta}{2}(\mathbb{E}T_{k+1}^{\mathrm{MRCA}} - \mathbb{E}T_{k}^{\mathrm{MRCA}})$$

for $k > 1$. When $k = 1$, there are only 2 lineages, so the total branch length is the anti-singleton entry. Thus, $\tau_{2,1} = \frac{\theta}{2}2\mathbb{E}T_2^{\mathrm{MRCA}}$. Rewriting this as a matrix equation for $k \in \{1,\ldots,n-1\}$ completes the proof. □

By combining Lemmas 1 and 2, we obtain the universal matrix $\mathbf{A} = \mathbf{BC}$. We now show how to compute the Ξ-dependent matrix $\mathbf{L}$. First, we establish the following result on the decomposition of the rate matrix $\boldsymbol{Q}$; this result was also obtained by [43, Equation 2.3] for the Bolthausen-Sznitman coalescent.

Lemma 3. *Fix an arbitrary Ξ-coalescent with $\lambda_i \neq \lambda_j$ for $i \neq j$, where $\lambda_i := \sum_{k=1}^{i-1} (\boldsymbol{Q})_{ik} = -(\boldsymbol{Q})_{ii}$. Let $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ denote the rate matrix of the ancestral process corresponding to $\Xi(d\boldsymbol{x})$ (that is the process counting the number of extant lineages at time t). Then,*

$$\boldsymbol{Q} = \mathbf{U}\mathbf{E}\mathbf{U}^{-1},$$

where $(\mathbf{E})_{ij} = \delta_{ij}(\boldsymbol{Q})_{ii}$, with δ_{ij} being the Kronecker delta which equals 1 if $i = j$ and 0 otherwise, and

$$(\mathbf{U})_{ij} = \begin{cases} 1, & i = j, \\ \frac{1}{\lambda_i - \lambda_j} \sum_{k=j}^{i-1} (\boldsymbol{Q})_{ik} (\mathbf{U})_{kj}, & i > j, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. By the construction of $\mathbf{U}$,

$$(\mathbf{U})_{ij} (\boldsymbol{Q})_{jj} = \sum_{k=j}^{i} (\boldsymbol{Q})_{ik} (\mathbf{U})_{kj},$$

which implies that $\mathbf{UE} = \boldsymbol{Q}\mathbf{U}$. Then, since $\mathbf{U}$ is triangular and has strictly positive diagonal entries, it is invertible. Therefore, $\boldsymbol{Q} = \mathbf{UEU}^{-1}$. □

The following result relates $\boldsymbol{a}_n := (\mathbb{E}T_2^{\mathrm{MRCA}}, \ldots, \mathbb{E}T_n^{\mathrm{MRCA}})^T$ and $\boldsymbol{c}_n = (c_{2,2}, \ldots, c_{n,n})^T$:

Lemma 4. *Let $\boldsymbol{a}_n$ and $\boldsymbol{c}_n$ be defined as above. Fix an arbitrary Ξ measure and a strictly positive function ζ. Now consider a time-inhomogeneous coalescent governed by $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$. If $c_{k,k} < \infty$, for $2 \leq k \leq n$, then*

$$\boldsymbol{a}_n = -(\mathbf{UD})_{2:n,2:n} \boldsymbol{c}_n,$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is the diagonal matrix $\mathrm{diag}([\mathbf{U}^{-1}]_{\cdot,1})$, with $[\mathbf{U}^{-1}]_{\cdot,1}$ denoting the first column of $\mathbf{U}^{-1}$, and $(\mathbf{UD})_{2:n,2:n}$ denotes the submatrix of $\mathbf{UD}$ in rows and columns 2 through n.

Proof. Note that $\mathbb{E}T_k^{\mathrm{MRCA}} = \int_0^\infty \mathbb{P}\left\{T_k^{\mathrm{MRCA}} > t\right\} dt$. Therefore,

$$\begin{aligned} \mathbb{E}T_k^{\mathrm{MRCA}} = \int_0^\infty \mathbb{P}\left\{T_k^{\mathrm{MRCA}} > t\right\} dt &= \int_0^\infty \sum_{l=2}^{k} \left[e^{\boldsymbol{Q}\int_0^t \frac{1}{\zeta(s)} ds}\right]_{kl} dt \\ &= \int_0^\infty \sum_{l=2}^{n} \left[e^{\boldsymbol{Q}\int_0^t \frac{1}{\zeta(s)} ds}\right]_{kl} dt \\ &= \int_0^\infty \sum_{l=2}^{n} \left[\mathbf{U}e^{\mathbf{E}\int_0^t \frac{1}{\zeta(s)} ds}\mathbf{U}^{-1}\right]_{kl} dt, \end{aligned}$$

where the third equality follows from the fact that $\boldsymbol{Q}$ is lower triangular and hence so is its exponential. Now, since $\mathbf{U}$ is lower triangular, its inverse is as well. Therefore, we may ignore the value of $[e^{\mathbf{E}\int_0^t \frac{1}{\zeta(s)}ds}]_{1,1}$. Letting $\mathbf{F}(t) := e^{\mathbf{E}\int_0^t \frac{1}{\zeta(s)}ds}$ but with $\mathbf{F}_{1,1}(t) := 0$, note that $\int_0^\infty \mathbf{F}(t)dt = \operatorname{diag}(0, \boldsymbol{c}_n)$. Then we have

$$\mathbb{E}T_k^{\text{MRCA}} = \int_0^\infty \sum_{l=2}^n [\mathbf{U}\mathbf{F}(t)\mathbf{U}^{-1}]_{kl}dt = \sum_{l=2}^n [\mathbf{U}\operatorname{diag}(0, \boldsymbol{c}_n)\mathbf{U}^{-1}]_{kl}.$$

Now, note that $(\mathbf{U})_{i,1} = 1$ for all i by Lemma 3 and induction. This implies $\sum_{l=1}^n [\mathbf{U}^{-1}]_{il} = \delta_{i1}$, or $\sum_{l=2}^n [\mathbf{U}^{-1}]_{il} = \delta_{i1} - [\mathbf{U}^{-1}]_{i1}$. Using this identity, we can rewrite the above expression for $\mathbb{E}T_k^{\text{MRCA}}$ as

$$\mathbb{E}T_k^{\text{MRCA}} = -\sum_{j=2}^n [(\mathbf{U}\mathbf{D})_{2:n,2:n}]_{k-1,j} c_{n,j},$$

where $\mathbf{D} = \operatorname{diag}([\mathbf{U}^{-1}]_{\cdot,1})$. Collecting these equations over $k \in \{2, \ldots, n\}$ in matrix form leads to the desired result. □

Using Lemma 4, we now see that the matrix $\mathbf{L}$ from Theorem 1 is simply $-(\mathbf{U}\mathbf{D})_{2:n,2:n}$. Lemma 3 provides a recursion to compute $\mathbf{U}$, and $\mathbf{D}$ may be computed by noting that $(\mathbf{U}^{-1})_{11} = 1$ and then since $\mathbf{U}\mathbf{U}^{-1} = \mathbf{I}$ we have

$$\mathbf{U}_{i1}^{-1} = -\sum_{j=1}^{i-1} (\mathbf{U})_{ij} (\mathbf{U}^{-1})_{j1}.$$

Proof of Theorem 1. Combining Lemmas 1, 2 and 4 we obtain the equations in the theorem. For the runtime, note that each of the $O(n^2)$ entries of $\mathbf{U}$ requires $O(n)$ computations, and so computing $\mathbf{U}$ is $O(n^3)$. The matrices composing $\mathbf{A}$ are known in closed form, however, and constructing $\mathbf{D}$ only requires filling $O(n)$ entries, each requiring $O(n)$ computations for a total of $O(n^2)$. To then obtain the SFS from $\boldsymbol{c}_n$ simply requires iterated matrix vector products taking $O(n^2)$ time. The overall procedure thus requires $O(n^3)$. □

Lemma 5. *For coalescents of the form $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$ where ζ is a constant function, $\boldsymbol{a}_n$ can be computed recursively from $\boldsymbol{c}_n$ and $\boldsymbol{Q}$ as follows:*

$$\begin{aligned}
\mathbb{E}T_2^{TMRCA} &= c_{2,2} \\
\mathbb{E}T_k^{TMRCA} &= c_{k,k} + \sum_{l=2}^{k-1} \frac{(\boldsymbol{Q})_{kl}}{\lambda_k} \mathbb{E}T_l^{TMRCA}, \qquad \text{for } k > 2.
\end{aligned}$$

Proof. The formulae follow immediately from the homogeneity of the process, recursing on the number of individuals, and noting that the probability that the first coalescence event for a sample of size k results in k lineages merging down to l lineages is $\frac{(\boldsymbol{Q})_{kl}}{\lambda_k}$. □

Proof of Corollary 1. Use Lemma 5 to compute $\boldsymbol{a}_n$ in $O(n^2)$ time. Then, $\boldsymbol{\tau}_n = \mathbf{A}\boldsymbol{a}_n$ by Theorem 1, which also takes $O(n^2)$ time to compute. □

Remark 1. *Other than computing* $\mathbf{U}$*, the algorithm presented in Theorem 1 is* $O(n^2)$*. Thus, for the Bolthausen-Sznitman Coalescent [44] or Kingman's coalescent, where* $\mathbf{U}$ *is known in closed form [43], the SFS can be computed in* $O(n^2)$ *time even for non-constant* ζ*.*

Remark 2. *The above results can easily be extended to a coalescent where both* ζ *and* $\Xi(d\boldsymbol{x})$ *depend on* t*, so long as* $\Xi(d\boldsymbol{x})$ *is piecewise constant. For example, in the recent past the population may evolve according to a* Beta*-coalescent, whereas for* t *greater than some* t_0 *the population may evolve according to Kingman's coalescent. By setting* ζ *appropriately in Theorem 1, one may obtain a "truncated SFS" [24] for each different* $\Xi(d\boldsymbol{x})$*. Then, using the truncated SFS for each epoch and the same machinery as in [24] one may compute the full SFS. The same techniques also allow one to consider multiple populations, with each population perhaps evolving according to its own* Ξ *measure.*

2.3 Numerical results

We implemented Theorem 1 and Corollary 1 in Mathematica, and the notebook is available upon request. We can compute the SFS for an arbitrary coalescent for a sample of size $n = 100$ in approximately one second and a sample of size $n = 300$ in a matter of minutes on a laptop computer, which is orders of magnitude faster than the more than one hour reported for a sample size of $n = 100$ using the current state-of-the-art method [39]. Furthermore, [39] only consider specific Ξ measures where the number of simultaneous multiple mergers is restricted. Our method has the same runtime for all Ξ measures (after computing the rate matrix and the vector of first coalescence times). See Figure 2.1 for runtime versus sample size. Furthermore, as noted above, if the spectral decomposition of the rate matrix $\boldsymbol{Q}$ is known, then the algorithm is $O(n^2)$. We also present runtimes for the Bolthausen-Sznitman coalescent (which has a closed form solution for the spectral decomposition [43]) in Figure 2.1.

As long as the rate matrix $\boldsymbol{Q}$ of the ancestral process can be found exactly, our method is numerically stable. This is the case for popular Λ-coalescents such as point-mass coalescents and Beta-coalescents, as well as point mass Ξ-coalescents. If the rate matrix must be evaluated numerically, however, high precision computation may be needed to avoid potential numerical problems due to catastrophic cancellation.

Using simulations, we now investigate the variation in the empirical SFS across independent realizations of the coalescent process and study the ability to infer the underlying model using the empirical SFS. We consider three different ζ, illustrated in Figure 2.2. Due to the association with population sizes in the case of Kingman's coalescent, we refer to ζ as the history or population size history. However, we caution that depending on the finite population size model, ζ may not represent the population size, but some other biologically relevant parameter. We consider a constant size history, a bottleneck history that undergoes a temporary 10-fold size reduction, and a growth history with repeated population doublings.

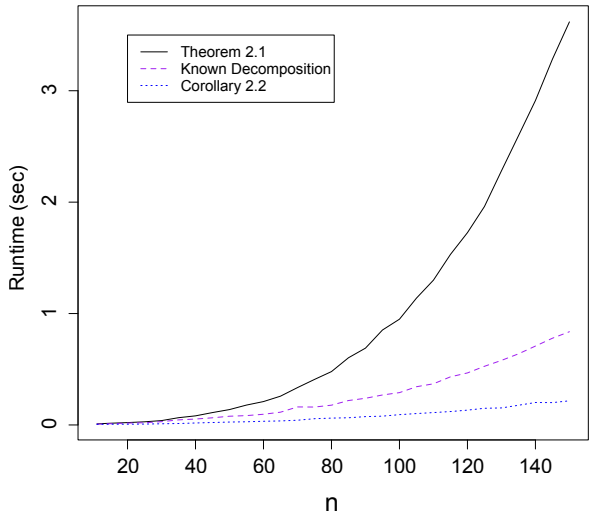


Figure 2.1: Runtime result (in seconds). Theorem 1 was used to compute the SFS for the time-homogeneous Bolthausen-Sznitman coalescent. The solid line uses Lemma 3 to compute the spectral decomposition of $\boldsymbol{Q}$ resulting in a cubic runtime. The dashed line uses the closed-form representation of the spectral decomposition of the Bolthausen-Sznitman coalescent [43, Theorem 1.1] to compute the SFS in quadratic time. The dotted line uses Corollary 1, which is also quadratic.

For each ζ, we consider Beta$(2-\alpha, \alpha)$-coalescents with $\alpha \in \{1, 1.5, 2\}$. Note that $\alpha = 1$ corresponds to the Bolthausen-Sznitman coalescent, while $\alpha = 2$ corresponds to the Kingman coalescent. For each of the nine distinct values of (ζ, α), we simulated $m = 1000$ independent trees with $n = 20$ leaves.

In Figure 2.3, we examine the observed variation in branch lengths across independent realizations of the coalescent process, from which we can deduce the variation in the observed SFS. Specifically, assume that each tree sampled from the coalescent process has the same mutation rate, and, without loss of generality, assume that time has been scaled such that the mutation rate is 1. Let $\widetilde{\tau}_{n,k}$ be the sum of branch lengths with k leaves and recall that $\hat{\tau}_{n,k}$ is the k^{th} entry of the empirical SFS on the n observed individuals. Then, $\mathbb{P}(\hat{\boldsymbol{\tau}}_n \mid \widetilde{\boldsymbol{\tau}}_n) \sim \text{Poisson}(\widetilde{\boldsymbol{\tau}}_n)$, and $\mathbb{E}[\widetilde{\boldsymbol{\tau}}_n] = \boldsymbol{\tau}_n$. In Figure 2.3, we plot $\widetilde{\tau}_{n,k}$ for each simulated tree, as well as its expected value $\mu_{n,k}$. Defining $\sigma^2_{n,k} := \text{Var}(\widetilde{\tau}_{n,k})$ for this case of $m = 1$, we also plot an estimate of the standard deviation $\hat{\sigma}_{n,k} = \sqrt{\hat{\mathbb{E}}[\widetilde{\tau}^2_{n,k}] - \tau^2_{n,k}}$, where $\hat{\mathbb{E}}$ is the empirical expectation. Now, if we sum the branch lengths and mutations over m independent trees (so then $\mathbb{E}[\widetilde{\tau}_{n,k}] = m\mu_{n,k}$, and $\text{Var}(\widetilde{\tau}_{n,k}) = m\sigma^2_{n,k}$), then $\mu_{n,k}$ and $\sigma^2_{n,k}$ describe the limiting behavior of both $\widetilde{\tau}_{n,k}$ and $\hat{\tau}_{n,k}$ as $m \to \infty$: by the Central Limit Theorem, $\frac{1}{\sqrt{m}}(\widetilde{\tau}_{n,k} - \tau_{n,k}) \to_d \mathcal{N}(0, \sigma^2_{n,k})$ and $\frac{1}{\sqrt{m}}(\hat{\tau}_{n,k} - \tau_{n,k}) \to_d \mathcal{N}(0, \sigma^2_{n,k} + \mu_{n,k})$.

A recent inconsistency result [45, Theorem 1] shows that a Λ-measure cannot be inferred from a single tree ($m = 1$), even as $n \to \infty$. Indeed, we see in Figure 2.3 that the

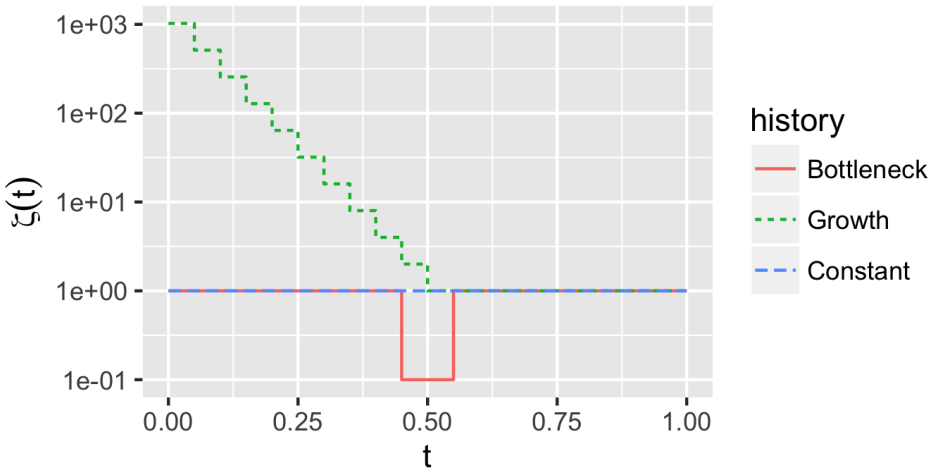


Figure 2.2: $\zeta(t)$ for three demographic scenarios: a constant size history, a bottleneck history that undergoes a temporary 10-fold size reduction, and a growth history with repeated population doublings. Note that the y-axis is stretched by $y \mapsto \log(y)$.

branch lengths $\widetilde{\tau}_{n,k}$ of a single tree can deviate substantially from $\tau_{n,k}$. For most k (say, $k \geq 5$), typically $\widetilde{\tau}_{n,k} = 0$ or $\widetilde{\tau}_{n,k} \gg \mu_{n,k}$ given a single tree. That is, for a single tree, branches subtending more than a few leaves are either not observed, or are much larger than the expected branch length. However, smaller k (especially the singletons, $k = 1$) have smaller relative standard deviation $\frac{\sigma_{n,k}}{\mu_{n,k}}$, and thus will tend to have lower relative error $\frac{\hat{\tau}_{n,k}-\tau_{n,k}}{\tau_{n,k}} \approx \mathcal{N}(0, \frac{\sigma^2_{n,k}+\mu_{n,k}}{m\mu^2_{n,k}})$ as m increases.

In the case of Kingman's coalescent, ζ is inferred by minimizing the KL-divergence between a normalized version of the empirical SFS and a normalized version of the expected SFS (e.g. [18, Equation 10]). We investigate how KL-divergence behaves as a function of the number m of independent trees simulated in the case of Λ-coalescents. Let $\boldsymbol{\tau}_n^{(\zeta,\alpha)}$ be the expected SFS under model $(\zeta(t), \alpha)$, and $\widetilde{\boldsymbol{\tau}}_n^{(\zeta,\alpha)}(m)$ the corresponding branch lengths summed over the first m simulated trees. Define $P^{(\zeta,\alpha)}(k) \propto \tau_{n,k}^{(\zeta,\alpha)}$ as the true probability distribution of derived alleles under scenario $(\zeta(t), \alpha)$, and $\widetilde{P}_m^{(\zeta,\alpha)}(k) \propto \widetilde{\tau}_{n,k}^{(\zeta,\alpha)}(m)$ as the conditional distribution of derived alleles, given the first m trees simulated under $(\zeta(t), \alpha)$. In Figure 2.4, we plot the KL-Divergence $D_{KL}\big(\widetilde{P}_m^{(\zeta_1,\alpha_1)} \big\| P^{(\zeta_2,\alpha_2)}\big)$ as a function of m, for every $(\zeta_1(t), \zeta_2(t), \alpha_1, \alpha_2)$ considered above (that is, ζ is constant, bottleneck, or growth, and α is 1, 1.5, or 2). In this case, we see that minimizing D_{KL} identifies the true scenario $(\zeta_1(t), \alpha_1(t)) = (\zeta_2(t), \alpha_2(t))$ with access to only a moderate number of independent trees (between 10 to 100).

Figure 2.4 is encouraging, as not too many independent trees are needed to distinguish between the different scenarios $(\zeta(t), \alpha)$. Unfortunately, in some cases it may be impossible to even sample two independent trees (personal communication, Jere Koskela). For example, in the model of [46], a multiple merger event happens over a single "generation", which

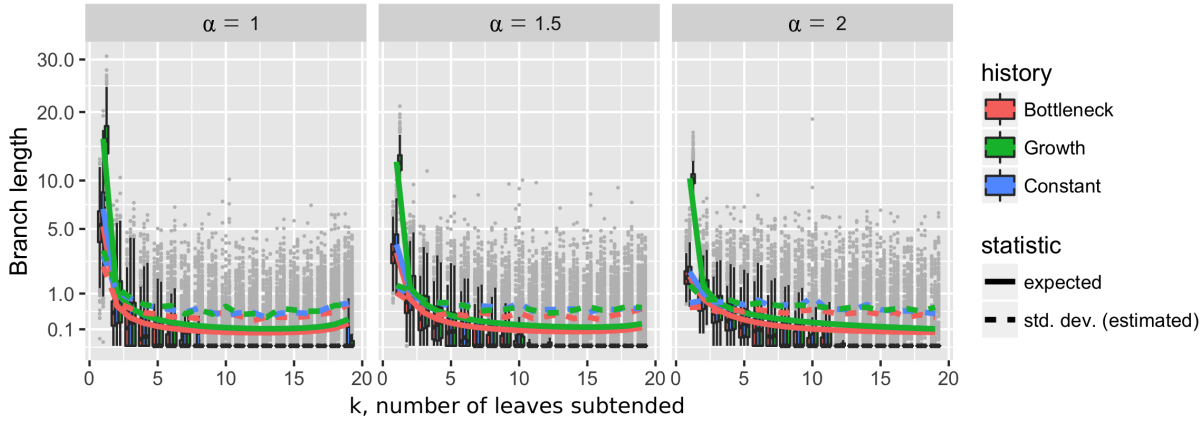


Figure 2.3: The distribution of the branch length subtending k leaves, for random trees under a Beta$(2-\alpha, \alpha)$-coalescent and $n = 20$. The solid line is the expected value from Theorem 1. We simulated 1000 independent trees per scenario and their branch length results are shown here as gray dots and box plots; the dashed line denotes the estimated standard deviation of the distribution. Note the y-axis is stretched by $y \mapsto \sqrt{y}$. The mean and standard deviation give the limiting behavior of $\hat{\tau}_{n,k}$ for many independent trees, under the Central Limit Theorem. For most k (say, $k \geq 5$), the branch length is usually 0, and has high variance relative to the mean. Thus $\hat{\tau}_{n,k}$ will tend to have higher relative accuracy for the smaller entries k.

can cause the multiple merger to affect unlinked sites, resulting in correlated coalescence times. However, in other models, multiple merger events may only affect the genome locally, and thus trees from unlinked sites are independent. For example, in the selective sweep model of [30], multiple mergers are caused by selective sweeps taking place over $O(\log(N))$ "generations", and a site experiences a multiple merger if $\frac{r_N \log(2N)}{s_N} = O(1)$, where N, s_N, r_N respectively parametrize the population size, selection strength, and recombination distance to the selected site. Thus, the independence of unlinked trees is not necessarily determined by the Λ- or Ξ-measure itself, but instead by the pre-limiting model.

2.4 Identifiability results

Before attempting to infer ζ or Ξ in practice, it is important to know whether such inference is possible using the SFS. For instance, when inferring ζ, if two different functions ζ_1 and ζ_2 produce the same SFS, then it is impossible to distinguish between the two using only the SFS. In such a case, we say that ζ is not identifiable. For Kingman's coalescent if one allows ζ to be an arbitrary positive function that produces a finite SFS, then ζ is not identifiable [47]. ζ is identifiable in the case of Kingman's coalescent, however, if one restricts ζ to be from a set of biologically realistic functions (technically, a set of functions with only a finite number

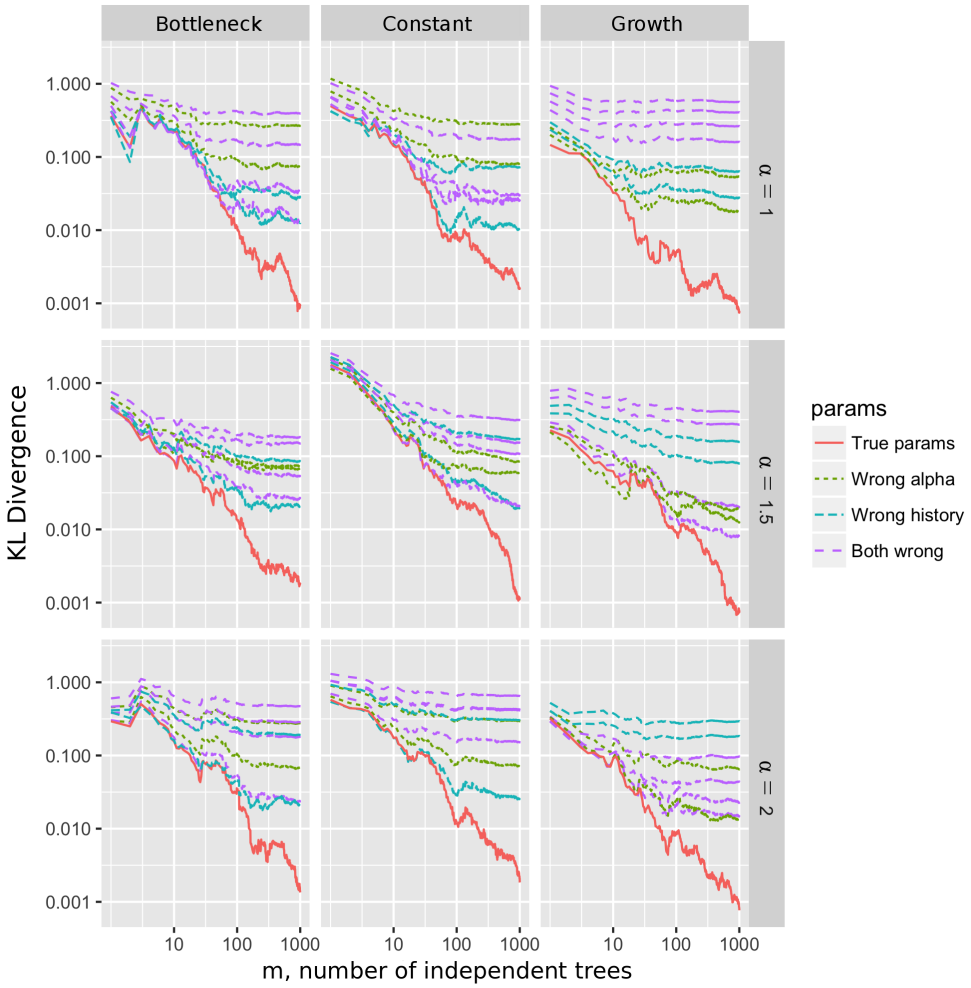


Figure 2.4: The KL-Divergence $D_{KL}\big(\widetilde{P}_m^{(\zeta_1,\alpha_1)} \big\| P^{(\zeta_2,\alpha_2)}\big)$, where $P^{(\zeta_2,\alpha_2)}(k) \propto \tau_{n,k}^{(\zeta_2,\alpha_2)}$ is the distribution of derived alleles under scenario $(\zeta_2(t), \alpha_2)$, and $\widetilde{P}_m^{(\zeta_1,\alpha_1)}(k) \propto \widetilde{\tau}_{n,k}^{(\zeta_1,\alpha_1)}(m)$ is the conditional distribution of derived alleles, given the first m trees simulated under $(\zeta_1(t), \alpha_1)$ and a mutation hitting one of those trees. For m large enough, D_{KL} is minimized by the true parameters, i.e. $(\zeta_1(t), \alpha_1(t)) = (\zeta_2(t), \alpha_2(t))$. D_{KL} can typically discriminate the true scenario for $m = 100$ trees. For $m = 10$ trees, D_{KL} is often, but not always, minimized by the true scenario.

of oscillations) [48, Theorem 11]. We show that a similar result holds for all coalescents of the form $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$ where $\Xi(d\boldsymbol{x})$ is fixed.

In general it is impossible to infer Ξ from the SFS if Ξ is not restricted. There has been some interest, however, in the case of distinguishing between a subset of Λ-coalescents [49]. We prove some results about the identifiability of the measure for various subsets of Λ-measures when ζ is a constant function. We also consider the question posed by [49] of whether or not the SFS can distinguish between exponential growth under Kingman's coalescent and a class of Λ-coalescents with constant ζ, and we show that indeed it is possible to distinguish between these cases with a surprisingly small number of samples. We note that our identifiability results require knowledge of the exact expected SFS, whereas [49] focus on the case where the expected SFS is approximated using an empirical SFS, which is what occurs in practice.

Throughout this section we assume that one has the exact expected SFS (i.e., the object computed by Theorem 1).

Identifiability of ζ for fixed Ξ measure

Before proceeding to the results and proofs, we first introduce some notation. Let $\mathcal{M}_K(\mathcal{F})$ denote the set of piecewise defined functions with at most K pieces made from some function family $\mathcal{F}$. Furthermore, let $\mathscr{S}(\mathcal{F})$ denote the sign-change complexity of $\mathcal{F}$. Informally, $\mathscr{S}(\mathcal{F})$ is the supremum of the number of times $f_1 - f_2$ crosses 0 over functions $f_1, f_2 \in \mathcal{F}$, which is related to the number of oscillations each $f \in \mathcal{F}$ is allowed to have (see [48, Definition 4] for a formal definition of $\mathscr{S}(\mathcal{F})$). We will also write ψ_n^{Ξ} for the number of 0 entries in $[\mathbf{U}^{-1}]_{\cdot,1}$ in the spectral decomposition of $\boldsymbol{Q}$ for a coalescent on n individuals governed by $\Xi(d\boldsymbol{x})$. Furthermore, denote by $\mathcal{X}$ the space of Ξ measures such that $(\boldsymbol{Q})_{k,k-1} > 0$ for all k. That is, $\mathcal{X}$ is the set of Ξ measures where for any sample size there is positive probability of a single pairwise merger. If we are only considering Λ-coalescents, then $\mathcal{X}$ contains all Λ measures except for δ_1, the star coalescent. We now present our main identifiability results and a conjectured bound on ψ_n^{Ξ}.

Our main result on the identifiability of ζ is the following theorem.

Theorem 2. *For an arbitrary Ξ-coalescent governed by the measure $\frac{\Xi(d\boldsymbol{x})}{\zeta(t)}$ where $\Xi \in \mathcal{X}$ is fixed, suppose $\mathscr{S}(\mathcal{F}) < \infty$ and $n \geq 2K + (2K-1)\mathscr{S}(\mathcal{F}) + \psi_n^{\Xi}$. Then for each expected SFS $\boldsymbol{\tau}_n$ there exists a unique $\zeta \in \mathcal{M}_K(\mathcal{F})$ consistent with $\boldsymbol{\tau}_n$.*

First, note that in the case of Kingman's coalescent, $\psi_n^{\delta_0} = 0$ for all n, and so in some sense, Kingman's coalescent is optimal in terms of the number of samples needed to ensure that a certain model space is identifiable. For the Bolthausen-Sznitman coalescent, $\psi_n^1 = 0$ for all n, which follows from the spectral decomposition [43, Theorem 1.1]. For the point mass Λ-coalescent with mass at $\frac{1}{2}$, for $n \geq 5$, all odd entries of $[\mathbf{U}^{-1}]_{\cdot,1}$ are 0 and so $\psi^{\delta_{1/2}} > 0$, thus implying that larger samples (relative to Kingman's coalescent or the Bolthausen-Sznitman coalescent) are needed for this coalescent to ensure that a given model space is identifiable.

We suspect that $\Lambda(dx) = \delta_{1/2}$ is the worst case among all Ξ-coalescents in $\mathcal{X}$ for identifiability, resulting in the following conjecture:

Conjecture 1. *For all $\Xi \in \mathcal{X}$ and $n \geq 3$, $\psi_n^{\Xi} \leq \lfloor \frac{n-1}{2} \rfloor - 1$.*

If this conjecture is true, then the bound on the sample size needed to have identifiability in Theorem 2 can be simplified to $n \geq 2[2K + (2K-1)\mathscr{S}(\mathcal{F})]$.

Identifiability of the Λ measure for a constant ζ

We also have the following results for Λ-coalescents about the identifiability of the Λ measure.

Theorem 3. *Consider the set of point-mass Λ-coalescents: $\{\delta_z : z \in [0,1]\}$. If Λ is restricted to be in this set and $n \geq 3$, then the expected SFS $\boldsymbol{\tau}_n$ uniquely determines Λ.*

Theorem 4. *Consider the set of* Beta*-coalescents: $\{\mathcal{L}(X) : X \sim \text{Beta}(2-\alpha, \alpha), \alpha \in [1,2)\}$. If Λ is restricted to be in this set and $n \geq 3$, then the expected SFS $\boldsymbol{\tau}_n$ uniquely determines Λ.*

Theorem 5. *Consider the set of coalescents:*

$$\left\{ \frac{\delta_0}{ae^{-bt}} : a, b > 0 \right\} \cup \{\delta_z : z \in [0,1]\} \cup \{\mathcal{L}(X) : X \sim \text{Beta}(2-\alpha, \alpha), \alpha \in [1,2)\},$$

that is: Kingman's coalescent with exponential growth, point-mass coalescents, or Beta*-coalescents. If Λ is restricted to be in this set and $n \geq 4$, then the expected SFS $\boldsymbol{\tau}_n$ uniquely determines Λ.*

Theorem 5 gives a positive theoretical answer to the question of whether or not the SFS can distinguish between exponential growth and multiple-merger coalescents. Using the techniques presented below, it is straightforward to obtain similar results for other subsets of Λ-coalescents.

Proofs of the identifiability results

The following Lemma will be used in proving the theorems in this section and may be of independent interest, as it shows that given the SFS for n individuals one can compute the expected time to most recent common ancestor for sample sizes $2, \ldots, n$ or vice-versa.

Lemma 6. *For all Λ- and Ξ-coalescents, there is a bijection between the expected SFS $\boldsymbol{\tau}_n$ and the expected times $\boldsymbol{a}_n$ to most recent common ancestor.*

Proof. Combine Lemmas 1 and 2 to see that $\boldsymbol{\tau}_n = \mathbf{BC}\boldsymbol{a}_n$, with $\mathbf{B}$ and $\mathbf{C}$ being universal. Then, since $\mathbf{B}$ is upper triangular and all of its diagonal entries are non-zero it is invertible. Furthermore, since $\mathbf{C}$ is bi-diagonal and the diagonal entries are all non-zero, it is also invertible. Therefore, $\mathbf{BC}$ is invertible and since $\boldsymbol{\tau}_n$ and $\boldsymbol{a}_n$ are related through an invertible matrix the transformation is bijective. □

To prove Theorem 2 we will use the following lemma.

Lemma 7. *Let $\lambda_k = -(\boldsymbol{Q}_{kk})$. For all $\Xi \in \mathcal{X}$ and all Λ other than $\Lambda(dx) = \delta_1(dx)$ (i.e., the star coalescent), the sequence $(\lambda_k)_{k\geq 2}$ is strictly increasing.*

Proof. Consider a sample of size $k+1$ and a subsample of size k. Without loss of generality, assume you remove individual $k+1$ to produce the subsample. The time to the first event is the same for both samples unless the first event only involves individual $k+1$ and one lineage from $\{1, \ldots, k\}$. That is, the total rate when there are $k+1$ lineages is equal to the total rate when there are k lineages plus k times the rate at which exactly a particular pair of individuals coalesce. Formally,

$$\lambda_{k+1} = \lambda_k + \frac{k}{\binom{k+1}{2}}(\boldsymbol{Q})_{k+1,k}.$$

By assumption, $(\boldsymbol{Q})_{k+1,k} > 0$, and so the total rates must be strictly increasing. □

We now prove Theorem 2. Our proof relies heavily on the proof of the corresponding result for Kingman's coalescent [48, Theorem 11]. We essentially show that this setting satisfies the same hypotheses as the Kingman's coalescent case and then use that result to complete our proof.

Proof of Theorem 2. By Lemma 6, the SFS is uniquely determined by $\boldsymbol{a}_n$. Then, furthermore, note that from Lemma 3 the matrix $\mathbf{U}$ is invertible since it is triangular with all non-zero entries along the diagonal. Then, by the same argument as in [48, Equation 12], we know that if the model space is not identifiable then for each k not corresponding to a zero in $\mathbf{D}$ (contributing to ψ_n^{Ξ}), λ_k must be the root of the Laplace transform of two different functions in the model space. By Lemma 7, these are all distinct, resulting in $n - \psi_n^{\Xi}$ roots. Then, by taking $n - \psi_n^{\Xi}$ sufficiently large, we obtain a contradiction via the Generalized version of Descarte's Rule of Signs [48, Theorem 4] and the theorem is proved. □

We now prove Theorems 3, 4, and 5. The idea is to explicitly calculate the $\mathbb{E}T_k^{\text{MRCA}}$ for the first few k for each allowed Λ measure and then use Lemma 6 to show that if Λ is uniquely determined by the first few $\mathbb{E}T_k^{\text{MRCA}}$, then it is uniquely determined by $\boldsymbol{\tau}_n$.

Proof of Theorem 3. $\mathbb{E}T_2^{\text{MRCA}} = 1$ for all Λ in the set of possible Λs. Consider $\Lambda = \delta_z$. Using Lemma 5 we see that:

$$\mathbb{E}T_3^{\text{MRCA}} = \frac{1}{3-2z} + \frac{3-3z}{3-2z} \cdot 1 = \frac{4-3z}{3-2z}. \tag{2.3}$$

This is a monotonically decreasing function of $z \in [0, 1]$, and so Λ is uniquely determined by $\mathbb{E}T_3^{\text{MRCA}}$. Then, appealing to Lemma 6, we see that Λ is uniquely determined by the SFS for $n \geq 3$. □

Proof of Theorem 4. A calculation similar to (2.3) gives $\mathbb{E}T_3^{\text{MRCA}} = \frac{2+3\alpha}{2+2\alpha}$ for $\Lambda = \mathcal{L}(X), X \sim \text{Beta}(2-\alpha, \alpha)$, where $\alpha \in [1, 2)$. This is a monotonically increasing function of $\alpha \in [1, 2)$ and the claim follows from the same argument as in the proof of Theorem 3. □

Proof of Theorem 5. Suppose that two distinct Λ-coalescents within the set of allowed models produce the same expected SFS for $n \geq 4$. Then, by Lemma 6, they would have the same values of $\mathbb{E}T_2^{\text{MRCA}}$, $\mathbb{E}T_3^{\text{MRCA}}$, and $\mathbb{E}T_4^{\text{MRCA}}$. By Theorems 3 and 4, we know that the Λ measures cannot both be point-mass coalescents or Beta-coalescents. From [48, Corollary 8], we also know that the Λ measures cannot both be Kingman's coalescent with different exponential growth parameters. There are thus three cases. They are all straightforward, albeit tedious.

Case 1: one Λ measure is a point-mass coalescent and the other is a Beta-coalescent. Letting $\mathbb{E}T_2^{\text{MRCA}} = 1$ (without loss of generality), we can explicitly compute $\mathbb{E}T_4^{\text{MRCA}}$ for the point-mass coalescent and the Beta-coalescent using the same recursive idea as in the proof of Theorem 3. Let $p_{m,k}$ denote the probability that when there are m lineages exactly k of them are involved in the next coalescence event. Then, by Lemma 5 $\mathbb{E}T_4^{\text{MRCA}} = c_{4,4} + p_{4,2}\mathbb{E}T_3^{\text{MRCA}} + p_{4,3}\mathbb{E}T_2^{\text{MRCA}}$. In particular, for the point-mass coalescent δ_z, this implies $\mathbb{E}T_4^{\text{MRCA}} = \frac{5}{3} + \frac{1}{2z-3} + \frac{3-2z}{18-24z+9z^2}$. Now, recalling the expression of $\mathbb{E}T_3^{\text{MRCA}}$ in (2.3) and letting

$$\mathbb{E}T_3^{\text{MRCA}} = t \tag{2.4}$$

implies $z = \frac{3t-4}{2t-3}$. Plugging this into $\mathbb{E}T_4^{\text{MRCA}}$, we see that for the point-mass coalescent,

$$\mathbb{E}T_4^{\text{MRCA}} = \frac{1}{3}\left[6t - 4 - \frac{2t-3}{6+t(3t-8)}\right]. \tag{2.5}$$

A similar calculation for the Beta-coalescent shows that

$$\mathbb{E}T_4^{\text{MRCA}} = \frac{1}{3}\left(6t - 2 + \frac{1}{t-2}\right), \tag{2.6}$$

with $t := \mathbb{E}T_3^{\text{MRCA}}$ under the Beta-coalescent. Equating (2.5) and (2.6), and solving for t results in the solution $t = 1$ or $t = \frac{4}{3}$. But, if $t = 1$, then we see that $z = 1$, the star-coalescent, which corresponds to $\alpha = 0$ for the Beta-coalescent, which is not in the set of allowed Beta-coalescents. If $t = \frac{4}{3}$, we see that $z = 0$, which corresponds to Kingman's coalescent, and $\alpha = 2$ for the Beta-coalescent, which again, is not in the set of allowed Beta-coalescents. Therefore, a point-mass coalescent and a Beta-coalescent with $\alpha \in [1, 2)$ cannot have the same $\mathbb{E}T_2^{\text{MRCA}}, \mathbb{E}T_3^{\text{MRCA}}$ and $\mathbb{E}T_4^{\text{MRCA}}$ simultaneously.

Case 2: one Λ measure is a point-mass coalescent and the other is Kingman's coalescent with exponential growth. Without loss of generality, assume that $\mathbb{E}T_2^{\text{MRCA}} = 1$ for the point-mass Λ-coalescent. The exponential-growth Kingman's coalescent model considered here has $c_{m,m} = -\frac{1}{b}e^{\binom{m}{2}/(ab)}\,\text{Ei}[-\binom{m}{2}/(ab)]$, where $\text{Ei}(x) := -\int_{-x}^{\infty}\frac{e^{-t}}{t}dt$ is the exponential integral [18, Supplemental Material Equation 5]. Then, the constraint $\mathbb{E}T_2^{\text{MRCA}} = c_{2,2} = 1$

implies $b = -e^{1/d}\,\mathrm{Ei}(-1/d)$, where $d := ab$. Furthermore, assuming this constraint and applying Theorem 1 to Kingman's coalescent, we obtain

$$\mathbb{E}T_3^{\mathrm{MRCA}} = \frac{3}{2} - \frac{c_{3,3}}{2} = \frac{3}{2} - \frac{e^{2/d}\,\mathrm{Ei}(-3/d)}{2\,\mathrm{Ei}(-1/d)}, \tag{2.7}$$

$$\mathbb{E}T_4^{\mathrm{MRCA}} = \frac{9}{5} - c_{3,3} + \frac{c_{4,4}}{5} = \frac{9}{5} - \frac{e^{2/d}\,\mathrm{Ei}(-3/d)}{\mathrm{Ei}(-1/d)} + \frac{e^{5/d}\,\mathrm{Ei}(-6/d)}{5\,\mathrm{Ei}(-1/d)}. \tag{2.8}$$

Now, in addition to $\mathbb{E}T_2^{\mathrm{MRCA}}$, if the two coalescents have the same values of $\mathbb{E}T_3^{\mathrm{MRCA}}$ and $\mathbb{E}T_4^{\mathrm{MRCA}}$, then the right hand sides of (2.4) and (2.7) must agree, while the right hand sides of (2.5) and (2.8) must agree. This implies

$$\frac{f_1(d) + e^{\frac{5}{d}}\,\mathrm{Ei}(-\frac{6}{d}) f_2(d)}{\mathrm{Ei}(-\frac{1}{d}) f_2(d)} = 0, \tag{2.9}$$

where

$$f_1(d) := 2\,\mathrm{Ei}\Big(-\frac{1}{d}\Big)\Big\{e^{\frac{4}{d}}\Big[\mathrm{Ei}\Big(-\frac{3}{d}\Big)\Big]^2 - 4e^{\frac{2}{d}}\,\mathrm{Ei}\Big(-\frac{3}{d}\Big)\,\mathrm{Ei}\Big(-\frac{1}{d}\Big) + \Big[\mathrm{Ei}\Big(-\frac{1}{d}\Big)\Big]^2\Big\},$$

$$f_2(d) := 3e^{\frac{4}{d}}\Big[\mathrm{Ei}\Big(-\frac{3}{d}\Big)\Big]^2 - 2e^{\frac{2}{d}}\,\mathrm{Ei}\Big(-\frac{3}{d}\Big)\,\mathrm{Ei}\Big(-\frac{1}{d}\Big) + 3\Big[\mathrm{Ei}\Big(-\frac{1}{d}\Big)\Big]^2.$$

However, by Lemma 8 in Section 2.6, there is no $d \in (0, \infty)$ such that (2.9) holds.

Case 3: one Λ measure is a Beta-coalescent and the other is Kingman's coalescent with exponential growth. If these two coalescents produce the same values of $\mathbb{E}T_2^{\mathrm{MRCA}}, \mathbb{E}T_3^{\mathrm{MRCA}}$ and $\mathbb{E}T_4^{\mathrm{MRCA}}$, then we must have $t = \frac{3}{2} - \frac{e^{2/d}\,\mathrm{Ei}(-3/d)}{2\,\mathrm{Ei}(-1/d)}$ in (2.6), and equating (2.6) and (2.8) implies

$$\frac{g_1(d) + 3e^{\frac{5}{d}}\,\mathrm{Ei}(-\frac{6}{d}) g_2(d)}{\mathrm{Ei}(-\frac{1}{d}) g_2(d)} = 0, \tag{2.10}$$

where

$$g_1(d) := 2\,\mathrm{Ei}\Big(-\frac{1}{d}\Big)\Big[-4e^{\frac{2}{d}}\,\mathrm{Ei}\Big(-\frac{3}{d}\Big) + \mathrm{Ei}\Big(-\frac{1}{d}\Big)\Big],$$

$$g_2(d) := e^{\frac{2}{d}}\,\mathrm{Ei}\Big(-\frac{3}{d}\Big) + \mathrm{Ei}\Big(-\frac{1}{d}\Big).$$

However, by Lemma 9 in Section 2.6, there is no $d \in (0, \infty)$ such that (2.10) holds.

Since each of the three cases results in a contradiction, we see that no such Λ measures exist, proving the identifiability claim. □

2.5 Discussion

We have presented an efficient algorithm for computing the SFS for a very general class of coalescents. While Λ- and Ξ-coalescents seem to be primarily used in practice to model the genealogies of marine species [50, 51], these coalescents also model a wide range of other phenomena including continuous strong positive selection [31], recurrent selective sweeps [29, 30], strong bottlenecks [35] and many others. Perhaps one of the reasons these coalescents are less widely used than Kingman's coalescent is because efficient inference tools have not yet been developed to the same extent.

Multiple-merger coalescents have also attracted some interest recently in the context of extremely large sample sizes [52]. In such cases the sample size is too large for the assumption of only pairwise mergers of lineages imposed by Kingman's coalescent to be biologically plausible, and indeed using Kingman's coalescent to model such populations causes biases in inference [52]. It should be possible to extend the results presented in this paper to discrete-time coalescents, such as the "exact coalescent" [53] corresponding to the coalescent arising from the discrete-time Wright-Fisher process, or any of the discrete-time random mating models considered by [28].

We also presented some encouraging identifiability results. While it is impossible in the general case to infer the inverse intensity function ζ or the measure of a Λ-coalescent from the SFS, for many biologically important cases identifiability does indeed hold. The method we presented for proving that the Λ measure is identifiable for constant ζ is powerful, but straightforward and should make it easy to prove whether or not the measure is identifiable for other sets of Λ- or Ξ-coaleascents. While we only considered the identifiability of Λ for fixed, constant ζ and the identifiability of ζ for fixed Λ or Ξ, it would be interesting to see if identifiability results can still be obtained for some model spaces while allowing both Λ and ζ to vary. It would also be interesting to extend our identifiability results for the Λ measure to some of the biologically relevant Ξ-coalescents.

Our identifiability results generally assumed access to the expected SFS. In practice, one observes a finite number of sites and so one only has a noisy estimate of the SFS. Our simulation study shows that, given a moderate number of independent trees, the empirical SFS is accurate enough to distinguish $\frac{\Lambda(dx)}{\eta(t)}$ for some simple models. However, the effect of noisy data is still largely unknown, especially in cases where convergence to the expected SFS is not guaranteed. The accuracy of inferring ζ with the empirical SFS has been studied for Kingman's coalescent [54], and it would be interesting to extend these results to general Λ-coalescents, and to the inference of the Λ-measure itself; the results presented here should make such an analysis more tractable.

2.6 Lemmas

Here we present two lemmas that are used in Theorem 5. Proofs are tedious but straightforward.

Lemma 8. *For $d \in (0, \infty)$,*

$$\frac{f_1(d) + e^{\frac{5}{d}} \operatorname{Ei}(-\frac{6}{d}) f_2(d)}{\operatorname{Ei}(-\frac{1}{d}) f_2(d)} \neq 0,$$

where

$$f_1(d) := 2 \operatorname{Ei}\Big(-\frac{1}{d}\Big)\Big\{e^{\frac{4}{d}}\Big[\operatorname{Ei}\Big(-\frac{3}{d}\Big)\Big]^2 - 4e^{\frac{2}{d}} \operatorname{Ei}\Big(-\frac{3}{d}\Big) \operatorname{Ei}\Big(-\frac{1}{d}\Big) + \Big[\operatorname{Ei}\Big(-\frac{1}{d}\Big)\Big]^2\Big\},$$
$$f_2(d) := 3e^{\frac{4}{d}}\Big[\operatorname{Ei}\Big(-\frac{3}{d}\Big)\Big]^2 - 2e^{\frac{2}{d}} \operatorname{Ei}\Big(-\frac{3}{d}\Big) \operatorname{Ei}\Big(-\frac{1}{d}\Big) + 3\Big[\operatorname{Ei}\Big(-\frac{1}{d}\Big)\Big]^2.$$

Lemma 9. *For $d \in (0, \infty)$,*

$$\frac{g_1(d) + 3e^{\frac{5}{d}} \operatorname{Ei}(-\frac{6}{d}) g_2(d)}{\operatorname{Ei}(-\frac{1}{d}) g_2(d)} \neq 0,$$

where

$$g_1(d) := 2 \operatorname{Ei}\Big(-\frac{1}{d}\Big)\Big[-4e^{\frac{2}{d}} \operatorname{Ei}\Big(-\frac{3}{d}\Big) + \operatorname{Ei}\Big(-\frac{1}{d}\Big)\Big],$$
$$g_2(d) := e^{\frac{2}{d}} \operatorname{Ei}\Big(-\frac{3}{d}\Big) + \operatorname{Ei}\Big(-\frac{1}{d}\Big).$$

In what follows, let $\mathrm{E}_1(x) := \int_x^\infty \frac{e^{-t}}{t} dt = -\operatorname{Ei}(-x)$. It is clear that $\mathrm{E}_1(x) > 0$ for all $x > 0$. Additionally,

$$e^{\frac{n}{d}} \mathrm{E}_1\left(\frac{n+1}{d}\right) = \int_{\frac{1}{d}}^\infty \frac{e^{-t}}{t + \frac{n}{d}} dt, \tag{2.11}$$

which follows from the definition of E_1 and a change of variables.

Proof of Lemma 8. First, by noting that $f_2(d) = 3[e^{\frac{2}{d}}\mathrm{E}_1(\frac{3}{d}) - \mathrm{E}_1(\frac{1}{d})]^2 + 4e^{\frac{2}{d}}\mathrm{E}_1(\frac{3}{d})\mathrm{E}_1(\frac{1}{d})$, it is easy to see that the denominator is strictly negative for $d \in (0, \infty)$. We will now show that the numerator is strictly positive for $d \in (0, \infty)$. First, by rearranging terms we see that

$$\begin{aligned} f_1(d) + e^{\frac{5}{d}} \operatorname{Ei}\Big(-\frac{6}{d}\Big) f_2(d) = 4\Big[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\Big] e^{\frac{2}{d}}\mathrm{E}_1\left(\frac{3}{d}\right) \mathrm{E}_1\left(\frac{1}{d}\right) \\ - \Big[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{2}{d}}\mathrm{E}_1\left(\frac{3}{d}\right)\Big]^2 \Big[2\mathrm{E}_1\left(\frac{1}{d}\right) + 3e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\Big]. \end{aligned} \tag{2.12}$$

Then, note

$$\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{2}{d}}\mathrm{E}_1\left(\frac{3}{d}\right) = \int_{\frac{1}{d}}^\infty \frac{e^{-t}}{t(t + \frac{2}{d})} dt < \frac{4}{d}\int_{\frac{1}{d}}^\infty \frac{e^{-t}}{t(t + \frac{5}{d})} = \frac{4}{5}\Big[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\Big].$$

Applying this inequality to the negative term on the right hand side of (2.12), we see

$$\begin{aligned}
&f_1(d) + e^{\frac{5}{d}} \operatorname{Ei}\Big(-\frac{6}{d}\Big) f_2(d) \\
&> 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\right] \\
&\quad \times \left\{e^{\frac{2}{d}}\mathrm{E}_1\left(\frac{3}{d}\right)\mathrm{E}_1\left(\frac{1}{d}\right) - \left[\frac{4}{5d}\mathrm{E}_1\left(\frac{1}{d}\right) + \frac{6}{5d}e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\right]\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t(t+\frac{2}{d})}dt\right)\right\} \\
&> 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\right]\mathrm{E}_1\left(\frac{1}{d}\right)\left[e^{\frac{2}{d}}\mathrm{E}_1\left(\frac{3}{d}\right) - \frac{4}{15}\mathrm{E}_1\left(\frac{1}{d}\right) - \frac{2}{5}e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\right] \\
&= 4\left[\mathrm{E}_1\left(\frac{1}{d}\right) - e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)\right]\mathrm{E}_1\left(\frac{1}{d}\right)\left[\int_{\frac{1}{d}}^{\infty} \frac{(\frac{1}{3}t^2 + \frac{7}{3d}t - \frac{8}{3d^2})e^{-t}}{t(t+\frac{2}{d})(t+\frac{5}{d})}\right],
\end{aligned}$$

which is greater than 0 for any $d \in (0,\infty)$ since $\mathrm{E}_1\left(\frac{1}{d}\right) > e^{\frac{5}{d}}\mathrm{E}_1\left(\frac{6}{d}\right)$ and $\frac{1}{3}t^2 + \frac{7}{3d}t - \frac{8}{3d^2} > 0$ for $t > \frac{1}{d}$. □

Proof of Lemma 9. The denominator of (2.10) is equal to $\mathrm{E}_1\left(\frac{1}{d}\right)\left[e^{\frac{2}{d}}\mathrm{E}_1\left(\frac{3}{d}\right) + \mathrm{E}_1\left(\frac{1}{d}\right)\right]$ which is strictly positive for $d \in (0,\infty)$, by definition of $\mathrm{E}_1(x)$. Furthermore, the numerator is strictly negative for $d \in (0,\infty)$ by noting the following:

$$\begin{aligned}
&g_1(d) + 3e^{\frac{5}{d}} \operatorname{Ei}\Big(-\frac{6}{d}\Big) g_2(d) \\
&= \left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right)\left[\int_{\frac{1}{d}}^{\infty} \frac{2e^{-t}}{t} + \frac{3e^{-t}}{t+\frac{5}{d}} - \frac{8e^{-t}}{t+\frac{2}{d}}dt\right] + 3\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t+\frac{5}{d}}dt\right)\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t+\frac{2}{d}}dt\right) \\
&= \left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right)\left[\int_{\frac{1}{d}}^{\infty} \frac{(-3t^2 - \frac{20}{d}t + \frac{20}{d^2})e^{-t}}{t(t+\frac{2}{d})(t+\frac{5}{d})}dt\right] \\
&\qquad + 3\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t+\frac{5}{d}}dt\right)\left[\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right) - \left(\int_{\frac{1}{d}}^{\infty} \frac{\frac{2}{d}e^{-t}}{t(t+\frac{2}{d})}dt\right)\right] \\
&= \left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right)\left[\int_{\frac{1}{d}}^{\infty} \frac{(-\frac{14}{d}t + \frac{20}{d^2})e^{-t}}{t(t+\frac{2}{d})(t+\frac{5}{d})}dt\right] - \frac{6}{d}\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t+\frac{5}{d}}dt\right)\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t(t+\frac{2}{d})}dt\right) \\
&< \left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right)\left[\int_{\frac{1}{d}}^{\infty} \frac{(-\frac{14}{d}t + \frac{20}{d^2})e^{-t}}{t(t+\frac{2}{d})(t+\frac{5}{d})}dt\right] - \frac{1}{d}\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right)\left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t(t+\frac{2}{d})}dt\right) \\
&= \left(\int_{\frac{1}{d}}^{\infty} \frac{e^{-t}}{t}dt\right)\left[\int_{\frac{1}{d}}^{\infty} \frac{(-\frac{15}{d}t + \frac{15}{d^2})e^{-t}}{t(t+\frac{2}{d})(t+\frac{5}{d})}dt\right] \\
&< 0.
\end{aligned}$$

Therefore, (2.10) holds. □

Chapter 3

Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans

This is joint work with Matthias Steinrücken, John A. Kamm, Emilia Wieczorek, and Yun S. Song. Sriram Sankararam kindly provided his introgression calls, Cathy Pfister provided helpful comments and suggestions, and we used the Global Biobank Engine resource made available to the community by the Rivas Lab. This work has been published in *Molecular Ecology* [55].

3.1 Introduction

In recent years, researchers have gathered an increasing amount of high-quality genomic sequencing data from human individuals that lived thousands of years ago [56] and individuals of extinct hominin sister groups [57, 58, 59]. These ancient samples provide unprecedented opportunities to elucidate the evolution of modern human populations and their relation to other hominins. Previous genetic evidence revealed that the ancestors of non-African humans exchanged genetic material with Neanderthal individuals after emerging out of Africa. Traces of this introgression can still be found in the genomes of modern-day humans. The emerging high-quality genomic sequence data for ancient hominins not only confirm these findings, but also allow detection of the exact location of these introgressed sequence fragments in the genomes of modern human individuals and a better understanding of their functional relevance.

Recent studies reported long regions in modern non-African individuals that are depleted of Neanderthal ancestry and enriched in genes, suggesting general negative selection against Neanderthal variants in genes [6, 60]. [7] and [8] provided further evidence that natural selection has acted to remove these introgressed segments, but there is some debate about the precise cause of this selective pressure. Dobzhansky-Müller incompatibilities (DMIs) are

a classic explanation for selection acting against introgressed alleles and play an important role in the evolution of reproductive isolation during speciation. DMIs involve alleles that have arisen separately in each population and are neutral or adaptive in isolation, but are deleterious when brought together in individuals of hybrid ancestry [61, 62]. DMIs have been observed in the hybrids of other species—e.g., *Drosophila simulans* and *D. melanogaster* [63]; *Mimulus guttatus* and *M. nasutus* [64]; and *Ambystoma californiense* and *A. tigrinum mavortium* [65]—and were hypothesized to be the cause (particular in male hybrids) of selection against Neanderthal ancestry in modern humans [6]. This hypothesis was motivated by finding significant enrichment of testes-expressed genes in regions of low Neanderthal ancestry and a substantial reduction of Neanderthal ancestry on the X-chromosome. Indeed, sex chromosomes are thought to play a special role in speciation, and it has been observed, for example in *Drosophila*, that loci contributing to reduced male fertility in hybrids are concentrated on the X-chromosome [66].

[6] and [60] reported some allelic variants associated with genetic diseases in genome-wide association studies (GWAS) that might have originated from the Neanderthal population. The studies also reported enriched Neanderthal ancestry in hair and skin related genes (keratin pathways), which suggests that these introgressed variants could have helped modern non-African populations to adapt to their local environments. All in all, the availability of Neanderthal introgression maps in modern humans has led to numerous follow-up studies investigating the various functional, evolutionary and medical implications of archaic hominin introgression into modern humans [67, 68, 69, 70, 71].

Methods to detect Neanderthal introgression tracts include a machine-learning based approach [6] that operates on suitably chosen "features" of the genetic data, and an approach based on sequence identity and divergence [60]. These methods have been extended to jointly detect Neanderthal and Denisovan introgression [72, 73]; the latter is found to be more prevalent in Oceania and Southeast-Asia.

In this article, we present a modification of the method `diCal 2.0`, previously developed by [74] for the inference of complex demographic histories. This modification, which we call `diCal-admix`, can be used to efficiently detect tracts of introgressed Neanderthal DNA. It is based on a hidden Markov model (HMM) approach that explicitly accounts for the underlying demographic history relating modern human and Neanderthal populations, including the introgression event. We first present our model-based method for detecting Neanderthal introgression and demonstrate through extensive simulations that our method is able to efficiently and accurately detect introgression in simulated data. We then apply our method to sequence data of modern humans from the 1000 Genomes Project and a high coverage genome from a Neanderthal individual from the Altai mountains [58]. Our results are in general agreement with previously obtained results, and we discuss similarities, differences, and their functional implications. In particular, we do not find evidence to support [6]'s hypothesis that DMIs played a role in shaping the pattern of introgression either genome-wide or specifically on the X-chromosome.

3.2 Materials and methods

Overview of our method

The method to detect Neanderthal introgression that we present here accounts explicitly for the underlying demographic history relating modern humans and Neanderthals. Therefore we briefly present some of the key features of this demographic model. Researchers have studied various aspects of the ancestral relations between modern African and non-African individuals using different methodologies. Furthermore, many studies have investigated the divergence of Neanderthals and modern humans. These studies resulted in several different, albeit largely consistent, estimates for the relevant demographic parameters, and here we closely follow [6] (specifically Figure SI2.1 of that paper) for consistency. However, note that [6] used a mutation rate of 2.5×10^{-8} per-site per-generation, whereas we use 1.25×10^{-8}, thus some of the numbers need to be adjusted.

The demographic model we use is depicted in Figure 3.1(a). The size of the most ancestral population and the size of the population ancestral to modern humans before the expansion are set to $N = 11,000$. The size of the population ancestral to modern humans after the expansion and the size of the African population are set to $N = 23,000$. Theses sizes are consistent with the estimates provided by [75]. The size of the non-African population after the split is set to $N = 2,000$. The size of the Neanderthal population is set to $N = 2,000$ as well, since it has been shown by [58] that the Neanderthal population size declined rapidly as the population neared extinction. The more recent small size will have a stronger impact on genetic variation than the larger ancestral size. Several studies [75] report a strong population bottleneck in European and Asian populations after the out-of-Africa event, followed by rapid exponential population growth, and [6] incorporate this into their demographic model. As we will detail later, our method considers each non-African haplotype one at a time. The genetic processes along a single ancestral lineage are not affected by the exact population size history, and thus we do not explicitly include these details in the demographic model used here.

Following [58], we set the time of divergence between modern humans and Neanderthals, T_{nean}, to 26,000 generations ago, which corresponds to 650 kya, assuming a generation time of 25 years, and the time that the population ancestral to modern humans expands is set to 12,000 generations ago [75]. The split between African and Non-African T_{div} is set to 4,000 generations ago, or 100 kya [76], and the time of the introgression or admixture event T_{admix} is set to $2,000$ generations ago [77], which corresponds to 50 kya [58]. Finally, the introgression coefficient is set as 3%, that is, a non-African individual at the time of introgression had a 3% chance that its parent was a Neanderthal individual. This is consistent with previous estimates of this quantity obtained by [78] and [8]. In Section 3.2, we will demonstrate the robustness of our method to misspecification of key parameters. It has been debated whether all non-African populations received genetic material from Neanderthals in a single pulse of introgression or several pulses. Here we assume that there has been only one pulse, and will defer disentangling such models to future work.

Before we describe our method to detect tracts of Neanderthal introgression, we provide

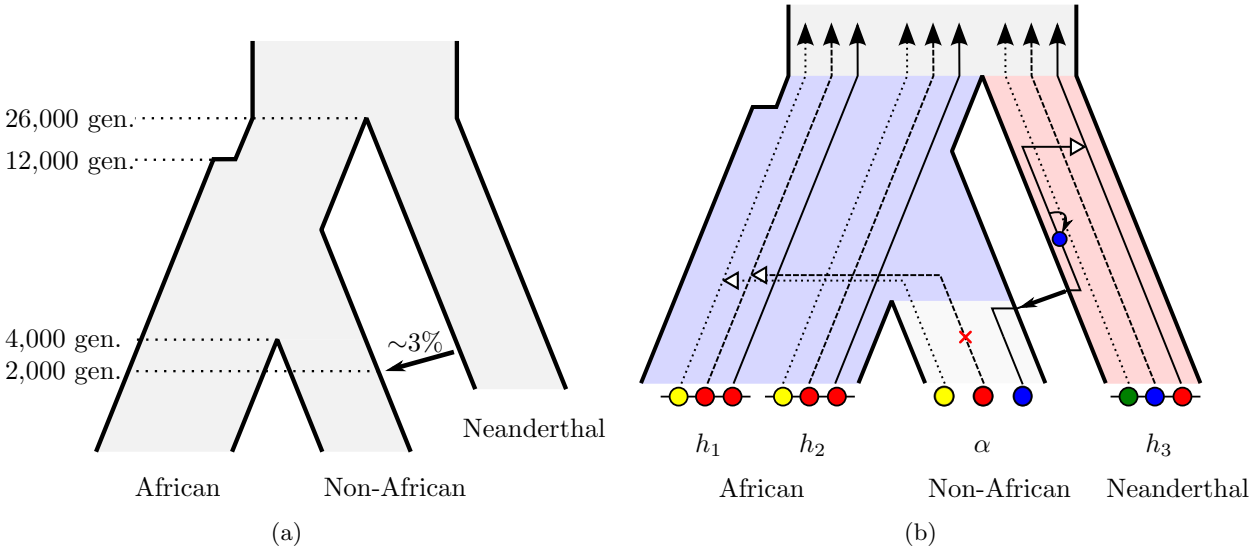


Figure 3.1: Illustration of the demographic model and the coalescent hidden Markov model. (a) A sketch of the demographic history relating African, non-African, and the Neanderthal population that contains the key features. (b) The Conditional Sampling Distribution applied to the detection of Neanderthal introgression. At each locus, the ancestral lineage of the focal Non-African (α) haplotype coalesces either with an African lineage (h_1 or h_2), or a lineage in the Neanderthal population (h_3) through the introgression event.

a brief overview of the methods developed by [6] and [60] for comparison. In [6], the authors employ a machine learning approached based on Conditional Random Fields, discriminative analogs of HMMs. To apply this framework, the authors represent the genotype data of the reference Neanderthal, the reference African population, and the focal non-African population in terms of "features" that are informative to distinguish between introgressed and non-introgressed sequence tracts. The authors chose three different classes of features: the distribution of alleles at informative SNPs; a measure of sequence divergence between the focal individual and both reference populations; and a feature to match the length distribution of the observed tracts to be consistent with the expectation from an introgression event 37–86 kya. The authors train the model on data simulated under a demographic scenario similar to Figure 3.1(a), and apply the trained model to detect introgression tracts in individuals from the 1000 Genomes dataset [79]. A modified version of this methodology was also applied in [72] to detect Denisovan ancestry in Southeast-Asians.

[60] developed a two-stage procedure to detect introgression tracts. In the first stage, the authors computed S^* statistics [80] in sliding windows along the genome. This statistic is sensitive to increased levels of diversity in high linkage disequilibrium, indicating a more ancient T_{MRCA} of a given region and also considers the tract length to detect archaic introgression. Notably, this stage does not require a reference sequence for an archaic individual. In the second stage, the authors compare the identified segments to a reference Neanderthal genome

to reliably identify Neanderthal introgression. The same two-stage approach was applied by [73] to identify Denisovan introgression.

Here, we apply a modified version of the method `diCal 2.0`, developed by [74] for inference of ancient demographies, to detect sequence tracts of Neanderthal DNA introgressed into modern humans. The method is based on the conditional sampling distribution (CSD) [81, 82, 83], which is similar to the copying model of [84], and is depicted in Figure 3.1(b). The CSD describes the distribution of sampling an additional focal genome or haplotype, conditional on having already observed a certain set of haplotypes. [74] introduce a version of the CSD that can be applied to haplotypes sampled from several subpopulations, accounting explicitly for the underlying demographic history. Under this model, the unknown genealogy relating the already observed haplotypes is approximated by a trunk genealogy of unchanging ancestral lineages extending infinitely into the past. At each locus, the ancestral lineage of the additional haplotype is absorbed into a lineage of the trunk. The dynamics of absorption depends on the underlying demographic history. In brief, lineages in different subpopulations cannot coalesce, unless continuous or point migration is possible at given rates, and the likelihood of coalescence is larger in small populations, but decreases in large populations. If an ancestral recombination event separates two loci, the haplotype of absorption and the time of absorption can change, thus different genomic segments can be copied from different haplotypes in the trunk, and the additional haplotype is realized as a mosaic of the observed haplotypes. The CSD can be implemented as an HMM along the genome of the additional haplotype, where the hidden state is the trunk haplotype that the genetic material is currently copied from, and a time of absorption. This absorption time is proportional to the likelihood that mutations can alter the genetic type at a given locus. [74] derived the emission and transition probabilities for the underlying HMM under general demographic models.

This CSD can be applied to detect introgressed tracts of Neanderthal ancestry in modern humans as follows. First, fix the underlying demography given in Figure 3.1(a) that relates modern African populations, modern non-African populations, and the Neanderthal population. The introgression event is modeled as a point migration. As depicted in Figure 3.1(b), the haplotypes sampled in the African population and a Neanderthal haplotype are used as the trunk haplotypes in their respective sub-population. Then, each non-African sample is, in turn, used as the additional haplotype in the non-African sub-population. Computing the forward and backward algorithm under the HMM for this CSD yields a marginal posterior distribution over the hidden states at each locus. Recall that these hidden states consist of both an absorption time and an absorbing haplotype. Marginalizing over the absorption time results in a posterior distribution over the trunk haplotypes, and, grouping haplotypes by sub-population, this gives a probability at each locus that the locus in the non-African haplotype is obtained from either an African (modern human) ancestor, or from an Neanderthal ancestor through introgression. Note that using an explicit time for the introgression event in the demographic model implicitly specifies a prior distribution for the length of the introgressed tracts. The software implementation of this method, `diCal-admix`, is available at `http://dical-admix.sourceforge.net/`.

Simulation study

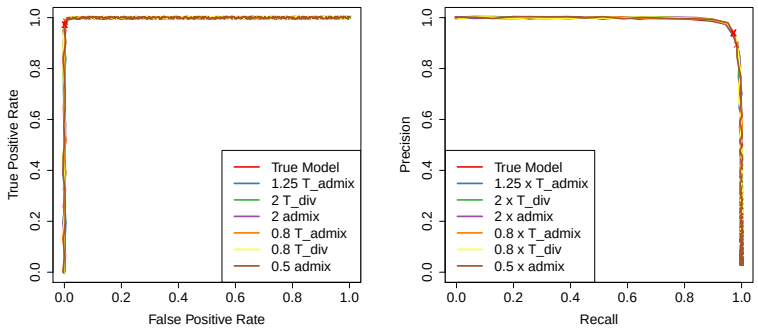
To demonstrate that `diCal-admix` can be used to accurately and efficiently identify tracts of Neanderthal introgression, we performed an extensive simulation study. To this end, we used the coalescent simulator `msprime` [85] to simulate sequence data under the demographic model given in Figure 3.1(a). Specifically, we simulated 176 haplotypes in the African population, one haplotype in the Neanderthal population, and 20 in the focal non-African population. For each such dataset we simulated 20 Mbp of sequence data, using a per generation population scaled mutation and recombination rate of 0.0005 per base (corresponding to a per-base per-generation rate of 1.25×10^{-8}). We simulated 50 replicates in each scenario, and estimated the introgression tracts on each focal haplotype.

To investigate how robust our method is to misspecification of the demographic model used for inference, our simulation study was two-fold. First, we simulated data under the demographic model given in Figure 3.1(a), and varied the demographic model used for the analysis. We kept all parameters fixed and only varied one focal parameter at a time. We varied the divergence time between modern humans and Neanderthal, using 0.5 and 2 times $T_{\text{nean}} = 26,000$ generations, and the divergence time between Africans and non-Africans, using 0.8 and 2 times $T_{\text{div}} = 4,000$ generations. Moreover, we varied the fraction of introgressed Neanderthal individuals, using 0.5 and 2 times $\mathsf{admix} = 3\%$, and the time of the introgression event, using 0.8 and 1.25 times $T_{\text{admix}} = 2,000$ generations. We also considered the effect of simulating under spatially heterogeneous recombination rates, taking 20 Mbp from the recombination maps inferred in [9]; we call this model the "*HapMap*" model. Finally, we considered four more complicated models. These models are the same as that given in Figure 3.1(a) except in the ways that we highlight below. While some of these models may not be indicative of any hypothesized demographic events, they show the robustness of `dical-admix` and highlight its potential utility in other scenarios. In the "*CEU ghost*" model, a non-African "ghost" population diverges from a basal CEU population $2,300$ generations ago, and 500 generations ago the two populations admix with 30% of their ancestry deriving from the ghost population and the remaining 70% coming from the basal CEU population. In the "*YRI ghost*" model, an ancient hominid ghost population diverges from the Neanderthal lineage $17,000$ generations ago, and then there is a 3% pulse of admixture from this ghost population into YRI $1,700$ generations ago. In the "*Nean into YRI*" model, there is a 3% pulse of admixture into both YRI and CEU $2,000$ generations ago. Lastly, in the "*Mig*" model, YRI and CEU exchange migrants at a rate of 0.001% of the population per generation.

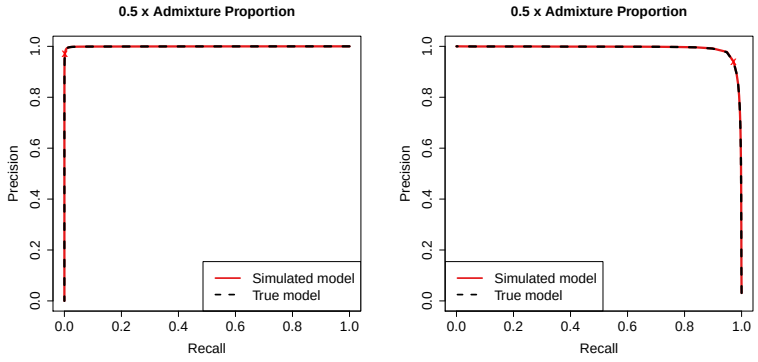
For simulated data, the true introgression status at each locus is known. Running `diCal-admix` to detect introgressed tracts in the simulated non-African individuals yields a posterior probability of introgression at each locus. Using different thresholds on this posterior probabilities for calling a locus introgressed, we can assess the true and false positive rate and the precision to generate receiver operating characteristic (ROC) and precision-recall curves for each analysis. Figure 3.2(a) shows these curves when the data simulated under the "true" model (Figure 3.1(a)) are analyzed using the different demographic models. We observe an overall good performance and robustness against misspecification of the parameters for

analyzing the data. From these plots, we determined that a threshold of 0.42 yields a good balance of the different performance metrics. We indicated this threshold along the curves by a red cross. Note that it is possible to increase the true positive rate by lowering the threshold, however, the decrease in precision would be more severe. Thus, we used this threshold in the remainder to call introgression tracts in the 1000 Genomes dataset. The ROC and precision-recall curves for the four more complicated scenarios and the *HapMap* model are provided in Appendix A.2.

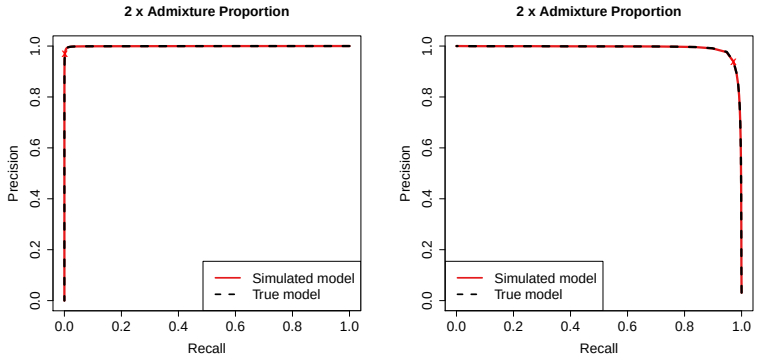
For the second part of the simulation study, we simulated data under the different demographic models. We then analyzed each dataset using the same parameters as used for the simulation on the one hand, and using the parameters of the "true" model (Figure 3.1(a)) on the other hand. The ROC and the Precision-recall curves for varying the introgression percentage are depicted in Figure 3.2(b) and Figure 3.2(c), and the curves for the remaining scenarios are given in Appendix A.2. Again, we observe that misspecifying the parameters of the analysis does not affect the performance substantially. The ROC curves demonstrate a good performance in terms of true and false positive rate in most scenarios, however, the precision-recall curves exhibit a poorer performance in some scenarios. This is to be expected, as in some scenarios, the Neanderthal and African population are genetically closer, for example, when the divergence time between Africans and non-Africans is increased, or the divergence time between Neanderthal and modern humans is decreased. If the populations are more closely related, it becomes more difficult to distinguish introgressed variation from variation shared between modern human populations. Lastly, we examine the detection power of our method with varying tract length in Appendix A.2, and observe good performance across all sizes.



(a)



(b)



(c)

Figure 3.2: Receiver operating characteristic and Precision-recall curves from the simulation study. (a) For data simulated under the "true" model, and the parameters for the analysis varied. We added a small amount of random "jitter" to the curves since they are hard to distinguish otherwise. (b) For data simulated under a model where the admixture proportion is 1.5% and analyzed using the same model and the "true" model. (c) For data simulated under a model where the admixture proportion is 6% and analyzed using the same model and the "true" model.

3.3 Results

Neanderthal introgression in the 1000 Genomes data

We applied `diCal-admix` to detect tracts of Neanderthal introgression in non-African individuals from the Phase I dataset of the 1000 Genomes Project [79], focusing on Europeans (CEU) and East Asians (CHB and CHS) in particular. These data were collected using low-coverage short-read sequencing, potentially affecting the quality of the genotype calls. However, we believe that this should not introduce substantial bias into our results. Moreover, we applied the strict mappability mask [79] to exclude from our analysis the genomic regions where no confident genotype calls were made in the 1000 Genomes dataset. We used the 88 YRI individuals (176 haplotypes) from this dataset as reference African haplotypes, assumed to have no introgressed genetic material from Neanderthals, that serve essentially as a modern human reference panel. We applied `diCal-admix` to compute the marginal posterior introgression probability along the genomic sequences of each of the 85 CEU individuals (170 haplotypes), 97 CHB individuals (194 haplotypes), and 100 CHS individuals (200 haplotypes) in turn. We used a high-coverage genomic sequence from an Altai Neanderthal individual [58] as a Neanderthal reference. [58] presented different genome alignability filters [58, SI 5b], and we used the `map35_50%`-filter, since this filter was suggested by the authors to be most appropriate for population genomic analyses.

`diCal-admix` requires that the genomic data be phased into haplotype sequences. The 1000 Genomes dataset is computationally phased, so we could use this data as provided; however, the diploid sequence of the Neanderthal individual cannot be phased using standard statistical methods. We instead used an additional pre-processing step to obtain a pseudo-haplotype sequence. As noted by [58], the Altai Neanderthal individual exhibits only a sixth of the heterozygosity of modern non-African individuals. Thus, the number of ambiguous sites that require phasing is small. We tested three different methods to obtain a haplotype allele for these remaining sites: choosing an allele uniformly at random, using the ancestral allele only, and using the derived allele only, where the ancestral states at each locus were determined using a six-primate consensus [86]. We observed little difference in our results between the different approaches, and thus we only present the results using the derived-allele approach. We used a mutation rate of 1.25×10^{-8} per-site per-generation [76] and chromosome-specific recombination rates obtained by averaging the fine-scale rates provided by [87]. Note that we made the simplifying assumption that the recombination rate is constant within each chromosome. Due to computational considerations, we did not compute the posterior at every genomic site, but rather grouped sites together into 500 bp windows; the details of this procedure are provided in [74]. Furthermore, we applied a moving average filter of length 15 kbp to the raw posterior in a post-processing step, to smooth sudden changes. We empirically observed that this filtering step improves detection.

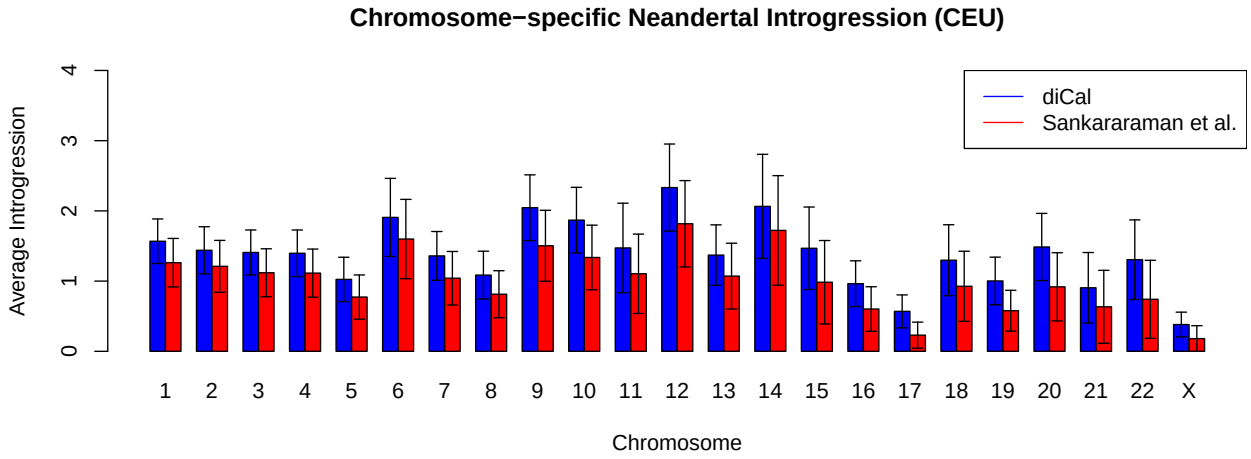
Moreover, we obtained from [6] the likelihoods of Neanderthal introgression that they computed for the same individuals. To compare these calls obtained at the SNPs to the ones obtained using `diCal-admix`, we interpolated the [6] likelihoods at the position in the middle

of the 500 bp windows employed by `diCal-admix`. As advised by [6], we used a threshold of 0.89 to call Neanderthal introgression tracts. [60] also identified tracts of Neanderthal ancestry in the individuals from the 1000 Genomes dataset, excluding the X-chromosome. We downloaded the population summaries from `http://akeylab.princeton.edu/downloads.html` and compared them to the results obtained with `diCal-admix`, when possible.

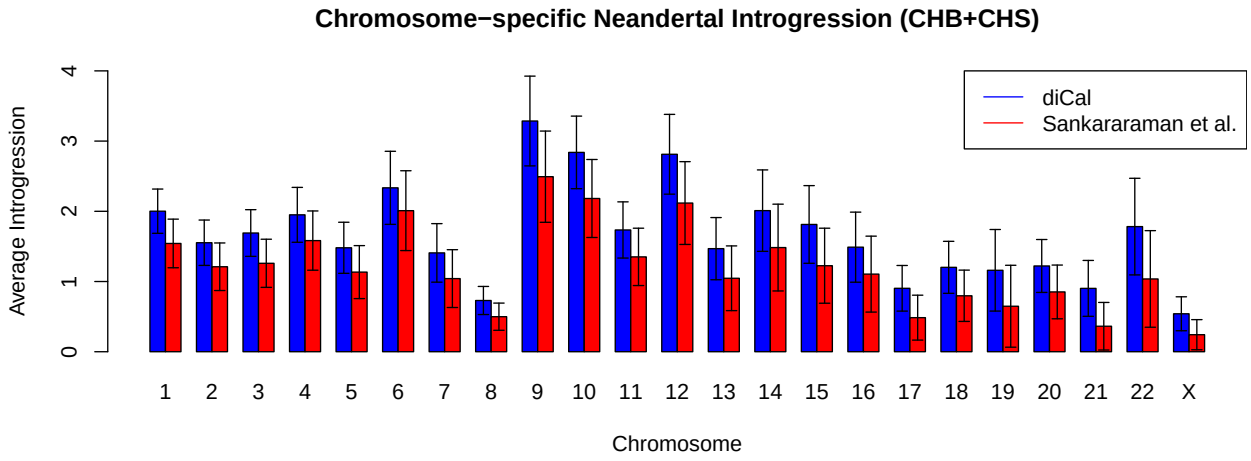
We computed the average of the marginal posterior introgression probability obtained at each locus using `diCal-admix` across each chromosome and across all CEU individuals, and separately across all CHB+CHS individuals. We performed the same averaging for the posterior probabilities obtained by [6]. Figure 3.3(a) shows the results for each chromosome in the CEU population, while Figure 3.3(b) shows the results for the CHB+CHS population. We find an average introgression probability of 1.48% in the CEU autosomes, and 1.80% in the CHB+CHS autosomes, whereas [6] report 1.13% and 1.35%, respectively. However, the average amount of introgression detected using `diCal-admix` varies, from as low as 0.57% on chromosome 17 in the CEU population, to as high as 3.29% on chromosome 9 in the CHB+CHS population. Compared to the autosomes, [6] previously reported a lower amount of introgression on the X-chromosome in CEU (0.18%) as well as in CHB+CHS (0.24%). Similarly, we observed a roughly four-fold decrease on the X-chromosome when compared to the autosomes in CEU (0.38%) and CHB+CHS (0.54%). The general patterns of introgression stratified by chromosome are in good agreement with [6]. However, `diCal-admix` detects on average 30% more introgression, which might be attributable to the fact that it detects more short tracts (see also Figure 3.7). Note that most Neanderthal ancestry proportions are lower than the 3% that was assumed in our model, which indicates that Neanderthal ancestry has been preferentially removed. In Section 3.3, we discuss in more detail possible mechanisms for this purging of Neanderthal ancestry.

The posterior distributions along the chromosomes allow for a more detailed view of Neanderthal introgression into modern humans as it varies along the genome. We determined whether a given locus is admixed on a particular haplotype by thresholding the posterior generated using `diCal-admix` at 0.42 and thresholding the posterior from [6] at 0.89. We then averaged these calls across 1 Mbp windows and across the individuals in the respective populations, and plotted the result as piece-wise constant functions. The skyline plots in Figure 3.4 show the percentage of Neanderthal introgression along chromosome 4 in CEU and the X-chromosome in the CEU population. (In Appendix A.3, we provide skyline plots for all chromosomes in the CEU and the CHB+CHS population.) In addition, we indicated the regions on the autosomes that were identified in [60] to be introgressed. As mentioned earlier, the X-chromosome was excluded in their study. We see good agreement between the calls made using `diCal-admix` and the calls from [6]. Furthermore, the regions of introgression detected by [60] cluster in regions were the skyline plots indicate introgressed genetic material.

To investigate the shared features and differences between the introgression call-sets, we generated Venn diagrams. For the `diCal-admix` posterior and the posterior from [6], we used the aforementioned thresholds to call introgression tracts in the CEU and CHB+CHS individuals. For each individual, we assessed at each locus whether either method, both methods, or no method detected Neanderthal introgression, and averaged these indicators to



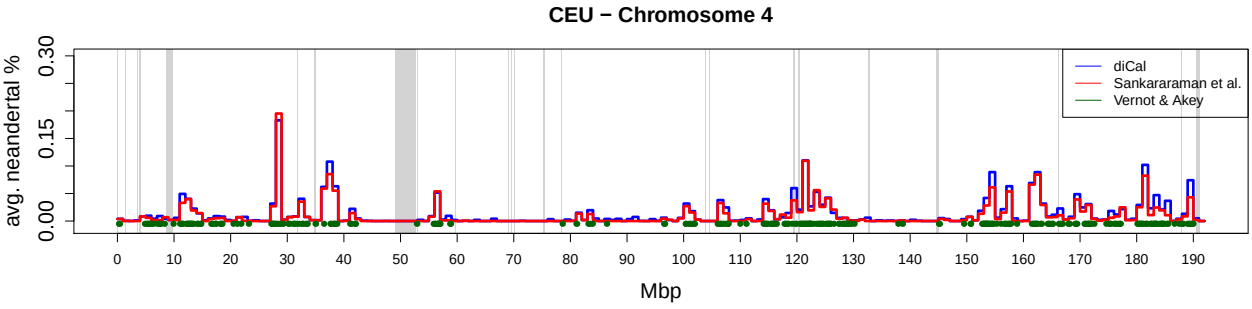
(a)



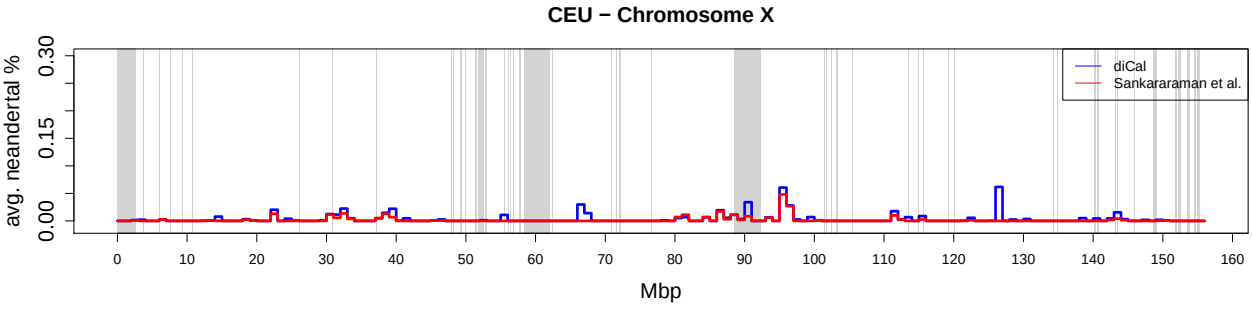
(b)

Figure 3.3: The amount of Neanderthal introgression in modern-day individuals, stratified by chromosome. The height of each bar gives the average introgression, and the whiskers indicate the standard deviation across the sample. (a) CEU. (b) CHB+CHS.

get population-wide percentages for the whole genome. Figures 3.5(a) and 3.5(b) show the Venn diagrams for the different call-sets in the CEU population, for the autosomes and the X-chromosome, respectively. Figures 3.5(c) and 3.5(d) depict the results for the CHB+CHS population, on the autosomes and the X-chromosome, respectively. We observe a large overlap between the calls based on `diCal-admix` and [6] on the autosomes, but less agreement on the X-chromosome. We also generated population-wide introgression maps in each population, called a *tiling path* by [6]. To this end, we identified those regions on the chromosomes where introgression was called for at least one individual in the respective population. We then



(a)



(b)

Figure 3.4: Skyline plot of the amount of Neanderthal introgression in the CEU population, averaged over all individuals in 1 Mbp windows. The results from `diCal-admix` are indicated in blue, and the results from [6] indicated in red. The regions reported as introgressed by [60] are indicated in green. The gray bars denote the regions were no calls were made in the 1000 Genomes dataset, which include the centromeres. (a) Chromosome 4. (b) Chromosome X.

compared these population-wide introgression maps with the population-level introgression maps published by [60]. We generated three-way Venn diagrams for the autosomes in the CEU population and the CHB+CHS population, shown in Figure 3.6, both in units of percentage of the whole autosome. Again, we observe a large overlap between `diCal-admix` and [6], but less so with [60]. This discordance might be explained to some degree by the fact that in the two-stage procedure of [60] the first step does not use sequence information from the Neanderthals as in the other methods. Thus, regions of high sequence identity between modern non-Africans and Neanderthals might be missed in this first step.

We also investigated the distribution of fragment lengths that were detected by the different methods. For all individual from a given population, we counted the number of times an introgression tract of a specific length was detected. Figure 3.7(a) and 3.7(c) depict the distributions of the absolute frequencies in the autosomes of the individuals in the CEU population and the CHB+CHS population, respectively. Figure 3.7(b) and 3.7(d) show the same distributions for the X-chromosomes. In addition to the empirical tract length distribution obtained from the 1000 Genomes individuals, we plotted the neutral expectation

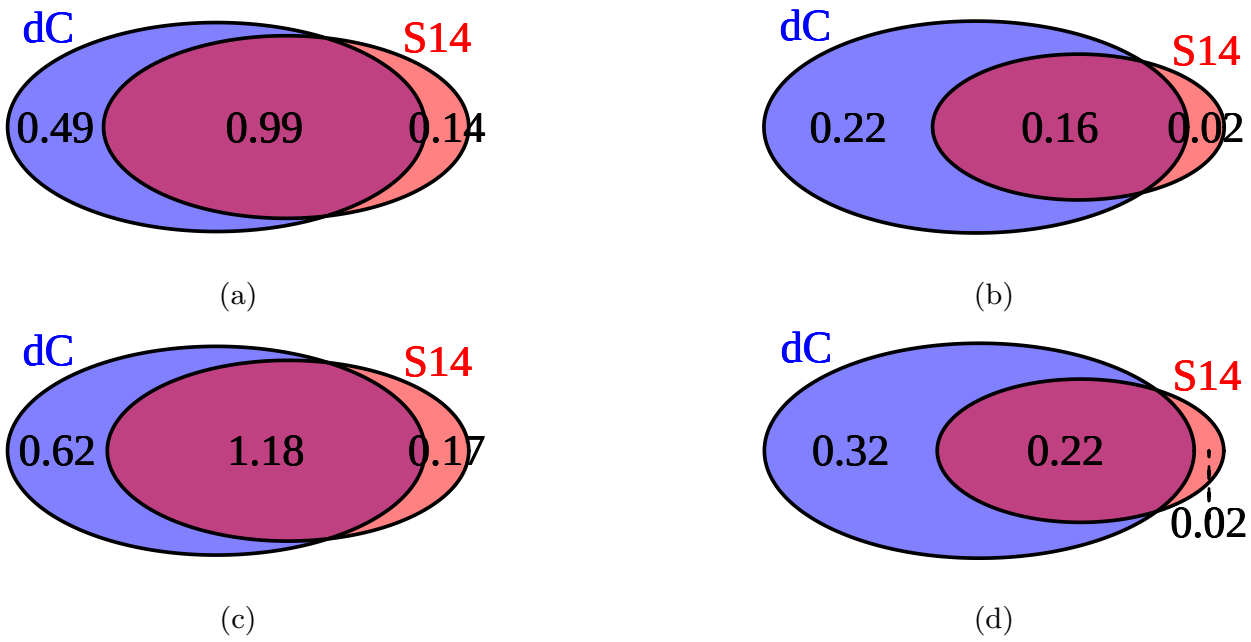
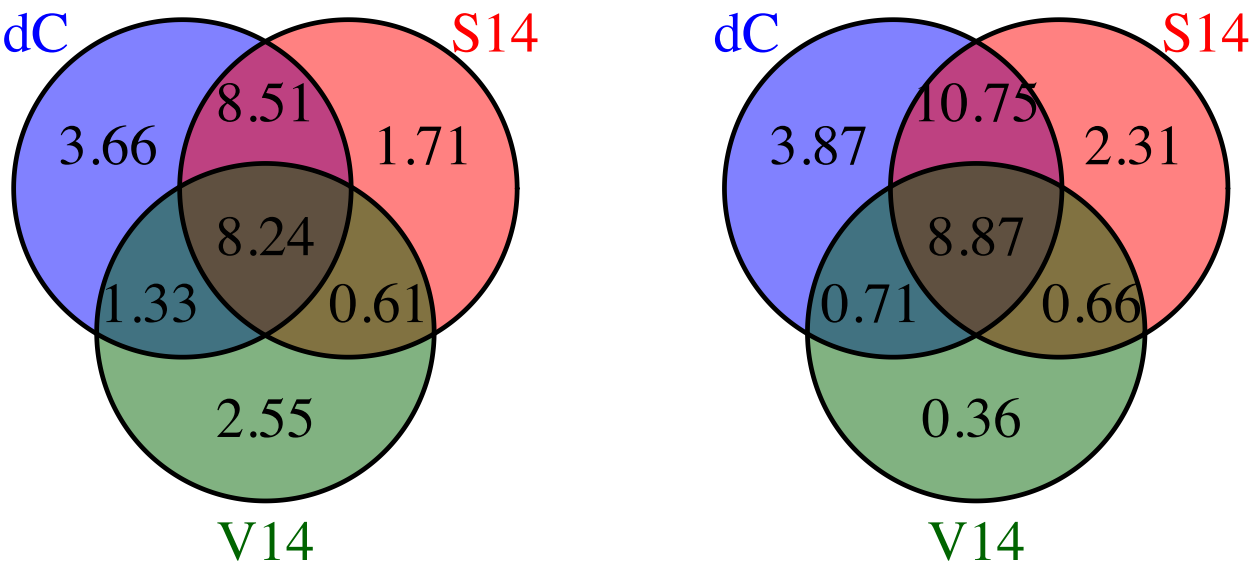


Figure 3.5: Venn diagrams of the average percentage of introgressed genetic material detected by `diCal-admix` (dC) and by [6] (S14). (a) For autosomes in the CEU individuals. (b) For the X-chromosome in the CEU individuals. (c) For autosomes in the CHB and CHS individuals. (d) For the X-chromosome in the CHB and CHS individuals.

of the absolute frequencies. This neutral expectation of the tract length distribution is computed under the following simple model. Approximating the chromosome as continuous, and considering the introgression tract in an individual at present, the distance between recombination breakpoints is exponentially distributed with parameter $g \times r$, where r is the per generation per base-pair recombination probability and g is the number of generations since the introgression event, because in each generation, there is a chance that recombination breaks down the introgression tract. Here we used $g = 2,000$ and $r = 1.19 \times 10^{-8}$ per-base per-generation for the autosomes and $r = \frac{2}{3} \times 1.18 \times 10^{-8}$ per-base per-generation for the X-chromosome. The exponential rate $g \times r$ can also be used to obtain the expected number of sequence tracts in a genome of a certain size, 3% of which are introgressed from an ancestral Neanderthal individual, which yields the expected absolute frequency.

This simple model for the neutral expectation is certainly oversimplified, but it serves as a first approximation. It is not discernible in these plots whether deviation from the neutral expectation is due to incorrect detection of the tracts, or the true underlying tracts actually being subject to non-neutral evolution. The fact that both methods deviate from the neutral expectation in qualitatively similar ways suggests both factors may be playing a role. However, it is surprising that, for the autosomes, both methods detect more long fragments than expected under the simple neutral model. This may be due to recombination rate variation along the genome or due to our slightly lower power to detect shorter fragments

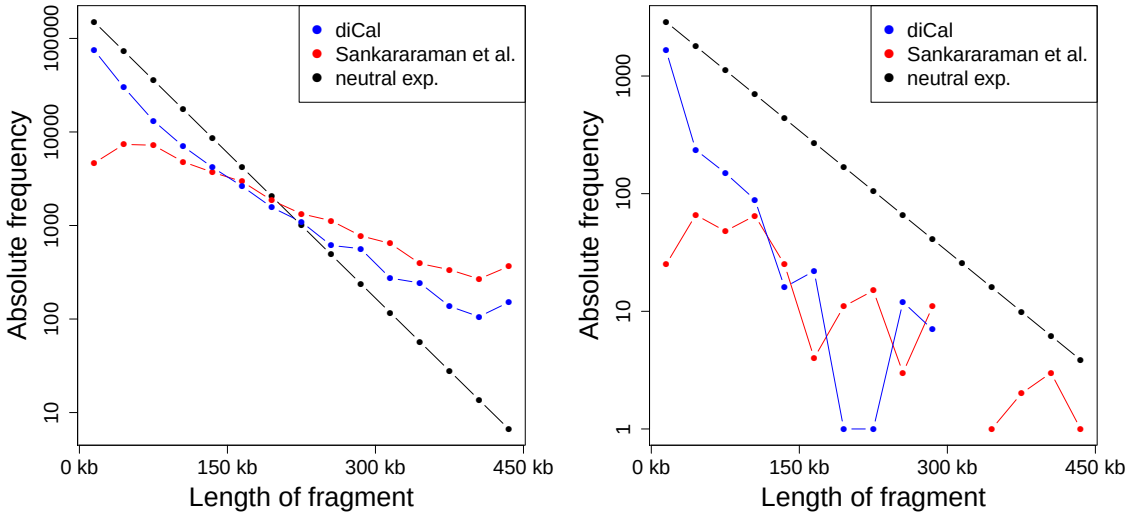


(a) (b)

Figure 3.6: Venn diagrams of the average percentage of regions on the autosomes where Neanderthal introgression was called for at least one individual in the population, by `diCal-admix` (dC), [6] (S14), and [60] (V14). (a) CEU. (b) CHB+CHS.

(see Appendix A.2). It could also suggest that either there was an additional introgression event that happened more recently than $2,000$ generations ago, or that some form of selection is acting that favors longer fragments.

In general, `dical-admix` detects more short fragments and fewer long fragments than reported by [6]. Moreover, the empirical distribution of `dical-admix` is closer to the neutral model. This and the other statistics of the empirical distribution of the introgressed Neanderthal tracts presented in this section suggest that there is merit in applying different methodologies for the detection of introgression. While all methods perform reasonably well on simulated data, they seem to be sensitive to slightly different features of the introgression tracts. Thus we suggest using the consensus of the three methods for highly confident introgression calls, and using regions unique to only some of the methods for more exploratory research.



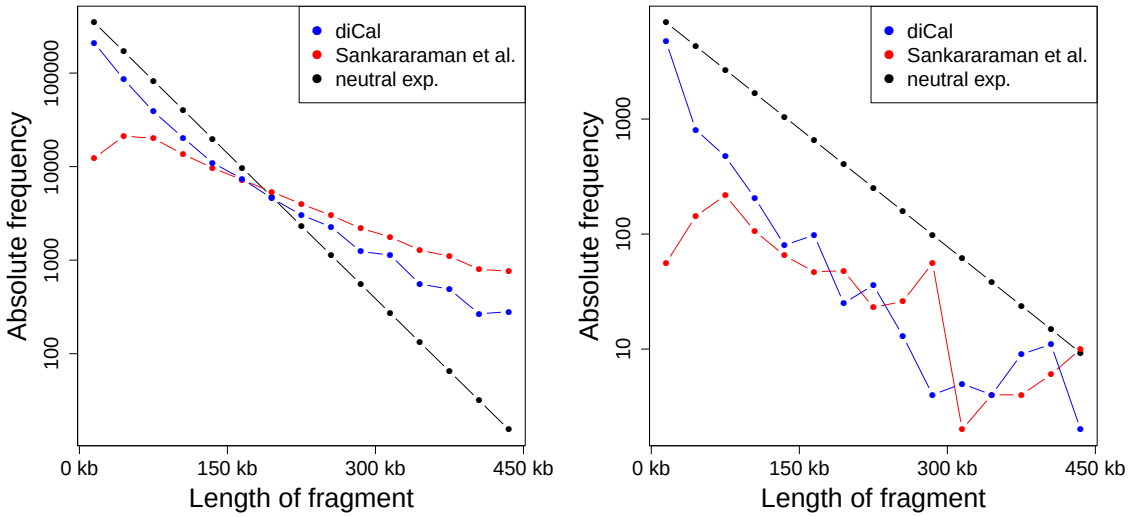


Figure 3.7: The empirical distribution of the lengths of the introgression tracts, accumulated across individuals. The different length are binned into classes of size 30 kbp. Missing values indicate unobserved classes. (a) Absolute frequency of tracts of a given length on the autosomes in the CEU population. (b) Absolute frequency of tracts of a given length on the X-chromosome in the CEU population. (c) Absolute frequency of tracts of a given length on the autosomes in the CHB+CHS population. (d) Absolute frequency of tracts of a given length on the X-chromosome in the CHB+CHS population.

Functional implications of Neanderthal introgression

To explore the functional implications of Neanderthal Introgression we performed a gene ontology (GO) analysis using GOrilla [88, 89], which looks for overrepresentation of GO terms at the top of a ranked list of genes. For each population, we ranked genes by their mean posterior probability of introgression (as determined by the `diCal-admix` posterior decoding) and looked for GO terms associated with either a lack of Neanderthal introgression or an enrichment of introgression. We restricted our analyses to the 500 bp resolution introgression calls where no more than half of the bases are masked by the 1000 Genomes strict mappability mask [79]. Furthermore, we only included genes where less than 10% of sites were masked. The results (shown in Appendix A.1) are broadly concordant between populations. Like [6], we find that genes associated with keratin are more likely to be introgressed than other genes, which hints at the possibility of adaptive introgression for these genes. Intriguingly, sensory perception, particularly olfaction, both had genes more likely to be introgressed as well as genes less likely to be introgressed. It is possible that adaptive introgression has played a role for some of these genes, for example by helping to adapt to local environments and that selection has removed introgression at other olfaction-related genes. It is also possible, however, that this is an artifact of either the introgression calls (e.g. due to high amount of polymorphism in ofactory genes [90]), or high variance in mean introgression rate (e.g. due to being smaller than other genes, or being spatially clustered in the genome).

We also investigated whether SNPs associated with particular phenotypes are more or less likely to be introgressed on average. To this end, we downloaded the results of $2,419$ GWASs that were performed on data from the UK Biobank [91, 92] and extracted all of the SNPs that were significant at a genome-wide significance level of 5.0×10^{-8} for each GWAS. We then tested whether the mean posterior probability of introgression at these SNPs was significantly higher or lower than expected using our bootstrap-like test (described below). Perhaps due to the large number of tests performed, we found only one statistically significant result: loci associated with being treated with desloratadine, a drug used to treat allergies, were significantly more likely to be introgressed (Bonferroni-corrected $p < 0.025$ in CEU and CHB+CHS, two-sided bootstrap-like test). The possibility that Neanderthal introgression may play a role in immune disorders such as allergic diseases is a tantalizing direction for future research.

Meanwhile, a number of tests were nominally significant at the $p = 0.05$ level in both populations and may be of interest for future research. In particular, we found that loci associated with sodium in urine may be less likely to be introgressed in both populations (nominal $p = 0.013$ in CEU, nominal $p = 0.0035$ in CHB+CHS), as are loci associated with the mean time to correctly identify matches, a test in which subjects attempted to quickly determine whether abstract symbols matched [93] (nominal $p = 0.0061$ in CEU, nominal $p = 0.0229$ in CHB+CHS). Some sets of loci may be more likely to be introgressed in both populations: loci associated with vaginal/uterine prolapse (nominal $p = 0.0049$ in CEU, nominal $p = 0.174$ in CHB+CHS); loci associated with being treated with glyceryl trinitrate, a drug used to treat heart disease (nominal $p = 0.0224$ in CEU, nominal $p = 0.0017$ in

CHB+CHS); loci associated with being treated with budesonide, a steroid used to treat asthma, COPD, allergies, and Crohn's disease (nominal $p = 0.0118$ in CEU, nominal $p = 0.0161$ in CHB+CHS); loci associated with being treated with cardioplen, another drug used to treat heart disease (nominal $p = 0.0205$ in CEU, nominal $p = 0.0082$ in CHB+CHS); loci associated with being treated with diclomax, a drug used to treat inflammation for example in rheumatoid arthritis (nominal $p = 0.0186$ in CEU, nominal $p = 0.0106$ in CHB+CHS); and loci associated with living in small towns (nominal $p = 0.011$ in CEU, nominal $p = 0.0233$ in CHB+CHS). We again urge caution in interpreting these results, due to the multiple testing burden making the above results not statistically significant.

Causes of selection against Neanderthal introgression

As mentioned in the Introduction, [6] hypothesized DMIs (particular in male hybrids) to be a cause of selection against Neanderthal ancestry in modern humans. This hypothesis was motivated by noting significant depletion of introgression in testes-expressed genes, as well as a substantial reduction of Neanderthal ancestry on the X-chromosome, which had been associated with DMIs in Drosophila [66]. Our results obtained using `dical-admix` also show a reduction in Neanderthal ancestry on the X-chromosome, but this reduction can be explained without appealing to DMIs, as we will discuss in more detail in Section 3.3. Here, we focus on global features of the genome and how they relate to potential DMIs.

[8] and [7] recently independently proposed that selection acts against Neanderthal ancestry due to a higher mutational load in Neanderthals rather than DMIs. In their study, [8] developed a likelihood method to explicitly infer the strength of selection against Neanderthal introgression based on the introgression maps obtained by [6]. The authors estimated a selection coefficient for deleterious exonic Neanderthal alleles around -3×10^{-4} for the autosomes. Since the estimated coefficient is on the order of the inverse of the effective population size in humans, they hypothesized that the deleterious alleles could have accumulated as a result of the small long-term effective population size in Neanderthals [58], which reduced the efficacy of selection in this population. When these deleterious alleles entered the larger human population through introgression, they were subjected to more efficient selection, and this led to the observed widespread selection against Neanderthal alleles. The authors use simulations to confirm that the population size history of Neanderthals could have indeed allowed for the accumulation of deleterious alleles on the observed order of magnitude. [7] arrived at similar conclusions while studying the strength of selection against Neanderthal introgression using forward simulations of autosomal genetic material.

In an attempt to disentangle the DMI and mutational load hypotheses, we performed a number of statistical tests. As detailed below, we found evidence that there was selection against Neanderthal ancestry, but could not find statistically significant evidence that the selection was due to DMIs. We interpret this as evidence in favor of the mutational load hypothesis, although it is possible that more data or more powerful statistical tests may show evidence in favor of the DMI hypothesis. In particular, our tests are underpowered to detect individual DMIs and would not be able to detect if there were a small number of strong

DMIs. On the other hand, if there are relatively few DMIs, then they would not be able to explain the reduction in Neanderthal ancestry across the whole genome.

We repeatedly used a test similar to bootstrapping, which we describe presently and refer to subsequently as the *bootstrap-like test*. When performing hypothesis tests, we must account for the spatial correlation of both our introgression calls and many genomic features of interest (e.g. gene locations or local recombination rates). We also would like to account for uncertainty in the introgression calls themselves. To this end, we left the genomic features of interest in place, and then sampled new introgression calls for each chromosome by drawing, with replacement, non-overlapping 5 Mb segments of our original introgression calls from the same chromosome. We then recalculated our test statistic using this resampled set of introgression calls. Repeating this resampling procedure many times provided an approximate empirical distribution of our test-statistic under the null hypothesis of no association between our introgression calls and the genomic feature of interest. We then used this distribution to compute approximate p-values. For all of the tests presented below, we again restricted our analyses to the 500 bp windows where no more than half of the bases were masked by the 1000 Genomes strict mappability mask [79].

To begin, we looked into whether selection against Neanderthal ancestry has occurred. First, note that the admixture proportion has previously been estimated as 3% [78, 8]. If there was no subsequent "dilution" of Neanderthal ancestry ([72], but see [94], and see [95] for a comprehensive review), then under neutrality we would expect about 3% ancestry on average in present-day populations, and we would expect about half of chromosomes to have more than 3% Neanderthal introgression on average and about half of chromosomes to have less than 3% introgression. Yet, no chromosome in either CEU or CHB+CHS has, on average, more than 3.29% introgression, and most chromosomes have less than 1.5% average Neanderthal ancestry. Thus, under this simple null model we can reject neutrality ($p = 2.4 \times 10^{-7}$ in CEU, $p = 5.7 \times 10^{-6}$ in CHB+CHS, two-sided sign test, $n = 23$) in both CEU and CHB+CHS. While the above test assumed that the admixture proportion was 3%, we would would be able to reject the null hypothesis of neutrality for any admixture proportion greater than 1.48% in CEU or 1.96% in CHB+CHS, both of which are much lower than the findings of the previous studies discussed above. Indeed, [8] report a confidence a 95% confidence interval of $[3.22\%, 3.52\%]$ for the admixture proportion in CEU and $[3.45\%, 3.86\%]$ for the admixture proportion in CHB+CHS [8]. Using the estimates of [8] to infer that selection has occurred is a somewhat circular argument because [8] used the assumption of selection against Neanderthal ancestry to infer their admixture proportions. Yet, an ancient European individual has been found with greater than 6% Neanderthal ancestry, and a number of other ancient Europeans have been found with greater than 3% Neanderthal ancestry without the assumption of selection against Neanderthal ancestry [96]. These ancient genomes together with our findings suggest that there has been a significant reduction in Neanderthal ancestry.

To explore if this reduction in Neanderthal ancestry is more pronounced in genic regions, we compared the mean (across individuals and loci) frequency of introgression in regions marked as exons in the RefSeq annotation [97] to the chromosome-wide mean. The results were largely concordant when we considered transcripts or coding sequences instead of exons,

so we present below only results based on exons. For CEU, we found that in 13 out of the 23 chromosomes there is less introgression in genic regions than in the rest of the chromosome, which is not statistically significant ($p = 0.678$, two-sided sign test, $n = 23$). Likewise, for CHB+CHS, we again found that 13 chromosomes have less introgression in genic regions, which is also not statistically significant ($p = 0.678$, two-sided sign test, $n = 23$). Furthermore, only 9 chromosomes showing a reduction in introgression at exons were shared between the two populations, which is not statistically significant ($p = 0.09$, two-sided permutation test). We also performed our bootstrap-like test to see if there was any significant decrease in genic regions on any chromosome, but we did not find any significant results (unadjusted $p > 0.05$) on any chromosome in either population, except for the X chromosome in CHB+CHS (unadjusted $p = 0.02$) which can be explained simply by the multiple testing burden. We interpret these results as indicating that either selection against Neanderthal introgression is fairly weak if it is acting on most or all genes, or has only acted on a subset of genes. It is also possible that selection against Neanderthal introgression has acted on genomic elements other than exons, such as regulatory elements.

Meanwhile, we found evidence that a measure of conservation, phastCONS [98], was significantly negatively correlated with mean introgression at a given locus (Spearman's $\rho = -0.029$, $p = 0.003$ in CEU and $\rho = -0.024$, $p = 0.043$ in CHB+CHS, bootstrap-like test), which indicates that selection was more likely to remove Neanderthal ancestry at highly conserved loci. We also tested if proportion of Neanderthal ancestry was positively correlated with local population-scaled recombination rate (as inferred by [9]), which would be suggestive of selection against Neanderthal ancestry because regions of high recombination would be more likely to separate neutral regions of Neanderthal ancestry from linked deleterious regions, an idea recently explored elegantly and in more detail by [71]. Similar to [71], we found a positive association between local population-scaled recombination rate and frequency of introgression (Spearman's $\rho = 0.055$, $p < 0.001$ in CEU and $\rho = 0.054$, $p < 0.001$ in CHB+CHS, bootstrap-like test) lending further credence to the hypothesis that selection is acting against certain regions of Neanderthal ancestry.

If DMIs were the cause of selection against introgression, then we would expect that genes that code for proteins with more binding partners would be less likely to be introgressed; each protein-protein interaction (PPI) can be thought of as a possible DMI. To test this hypothesis, we used the PICKLE2.0 PPI network [99, 100], and associated a number of binding partners to each gene by counting the number of PPIs in which the protein coded by that gene participates. We found an insignificant correlation between number of binding partners and mean frequency of introgression in both populations (Spearman's $\rho = -0.016$, $p = 0.314$ in CEU, $\rho = -0.003$, $p = 0.858$ in CHB+CHS, two-sided bootstrap-like test), which provides weak indirect evidence against the DMI hypothesis.

As a more direct test of the DMI hypothesis, we tested whether proteins that interact (according to the PICKLE2.0 PPI network) are more likely to be co-introgressed. In particular, for each gene in the PPI network, we say that that gene is introgressed if any part of any of its exons is in a called introgression tract. For each individual we then assign a weight to each edge in the PPI network as follows. Let gene A and gene B be the genes that code for

the proteins involved in the interaction corresponding to the edge of interest. If each copy of gene A and gene B (i.e. on autosomes, we assume there are two copies of each gene and on the X-chromosome males have only one copy) has the same ancestry, the edge is assigned a weight of one – in this individual, this interaction is always between proteins from genes of the same ancestry. Meanwhile, if all of the copies of gene A are of one type of ancestry and all of the copies of gene B are of the other type of ancestry, then the edge is assigned a weight of zero – this interaction is never between proteins of the same ancestry. Finally, if either gene has mixed ancestry (i.e. one copy from one ancestry and the other copy from the other ancestry) the edge is assigned a weight of 0.5 – in this case it can be shown that if one randomly selects a copy of gene A and a copy of gene B, then the proteins produced by those copies will have the same ancestry 50% of the time. Thus, these edge weights are the probabilities that compatible proteins interact, assuming that both ancestry types at each locus produce the same amount of protein and the probability that a given protein is involved in a particular interaction does not depend on its ancestry. We then averaged these weights across individuals and across edges in the PPI network to obtain a test statistic. Using this test, we failed to obtain a significant result in either population ($p = 0.117$, CEU, $p = 0.647$, CHBS, one-sided permutation test), which again provides some evidence against the DMI hypothesis. If DMIs are not widespread, then this test would likely not have power to detect them, so we are unable to rule out the possibility that there are a small number of DMIs. Yet, a small number of DMIs would be unable to explain the overall reduction in Neanderthal ancestry across the genome.

Taken as a whole, we find that while there has been selection against Neanderthal introgression, for example at highly conserved loci, it seems that the negative selection is likely not due to widespread DMIs, which lends more credence to the mutational load hypothesis. We also note that in contrast to the broad findings presented here, a small number of specific loci have been found to have experienced positive selection for archaic introgression [101, 102, 103].

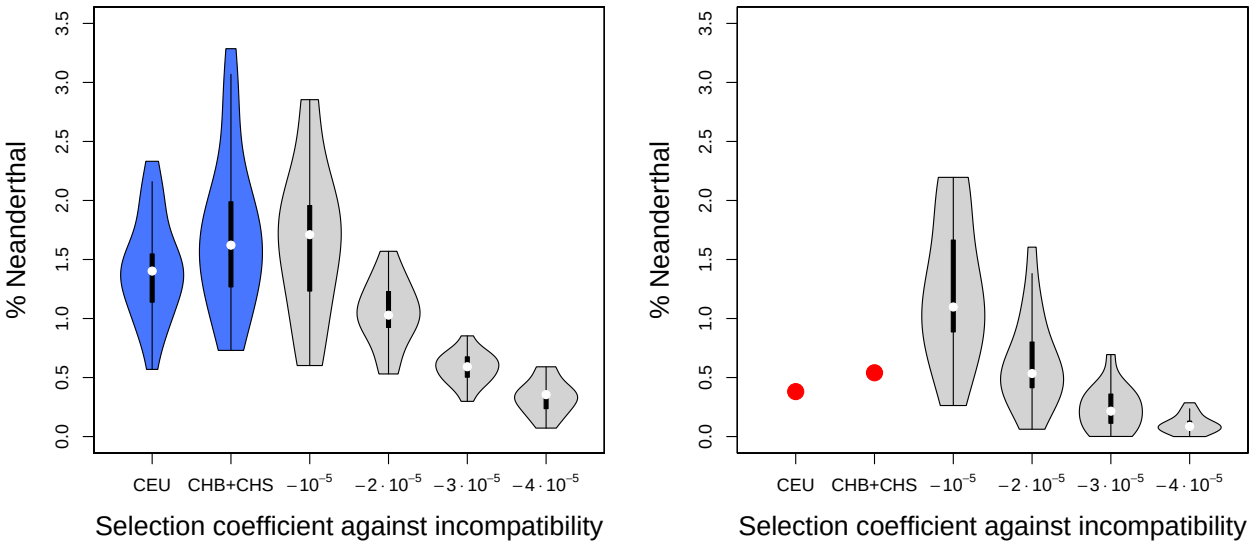
Patterns of introgression on the X-chromosome

To further explore whether selection against introgressed Neanderthal variation differed between autosomes and the X-chromosome, we performed forward simulations in a Wright-Fisher model, focusing on the dynamics of Neanderthal alleles in the modern human population after the introgression event $2,000$ generations ago. For the autosomes, we modeled each diploid individual to be comprised of two chromosomes. Each chromosome consisted of 5,000 loci, and recombination could act between these loci. The recombination rate was calibrated such that this corresponds to a 150 Mbp chromosome with a recombination rate of 1.25×10^{-8} per generation per base-pair. Similar to [7], the population size was set to $N = 1,860$ for the first 900 generations, followed by an instantaneous decrease to $N = 1,032$ with subsequent exponential growth at 0.38% per generation. For computational reasons, we limited the population size to $N = 10,000$, which is reached roughly 400 generations before present. In the generation immediately after the introgression event, the chromosomes

of 97% of the individuals in the population carry modern human alleles at all 5,000 loci on both chromosomes, and 3% carry Neanderthal alleles, representing the introgressed individuals. In the subsequent generations, the fitness of an individual is $(1-s)^D$, with selection coefficient s, and D denoting the number of Neanderthal alleles that a diploid individual carries. Figure 3.8(a) depicts the amount of Neanderthal introgression measured in the autosomes in the CEU and the CHB+CHS population, as well as the amount of Neanderthal introgression in a sample of individuals at present from populations simulated with different values for s, repeated 16 times. These simulations suggest that a selection coefficient on the order of $s = -2 \times 10^{-5}$ is more than sufficient to explain the observed reduction in Neanderthal ancestry from the initial proportion of 3% on the autosomes. These results are largely consistent with the estimates of selection against introgression provided in [8, Table 1], where -3×10^{-8} is estimated for the effective selection strength per exonic site. In our simulations, each locus corresponds to 30,000 sites. According to the annotation from the UCSC genome browser (`https://genome.ucsc.edu/`), the genome-wide density of exonic sites is 2.8%, and thus a simulated locus contains roughly 840 exonic sites. The simulated selection strength of $s = -2 \times 10^{-5}$ then corresponds to a selection strength of -2.4×10^{-8} per exonic site.

We also performed simulations for the X-chromosome to see if the strength of selection is similar on the autosomes and X-chromosome. In our simulation, the fitness of males was determined solely by their single X-chromosome and their Y-chromosome was modeled as selectively neutral. In females, to model X-inactivation, we chose one chromosome randomly to determine the fitness. There are several methods to calibrate the selection coefficient, s. Here we chose to calibrate s such that, in both females and males, carrying only Neanderthal variants at a certain locus has the same affect on fitness for the X-chromosome as for an autosome. Consequently, to determine the exponent for the fitness in females, the number of Neanderthal alleles on the active chromosome is multiplied by two. This calibration effectively corresponds to an additive selection model in females. In males, the number of Neanderthal alleles on the single X-chromosome is also multiplied by two to obtain the exponent. Note that [8] used a different calibration, where selection in males is half as strong. Figure 3.8(b) shows the Neanderthal proportions on the X-chromosome in both populations, and the results of the simulations for different values of s. We observe that using this calibration for the selection coefficient, for the same strength of selection, the amount of Neanderthal ancestry in the population at present is reduced on the X-chromosome compared to the autosomes. While this might seem at odds with the reduced effective population size for the X-chromosome, and hence a lower efficacy of selection, it can be explained by the fact that genetic variants in males have a stronger impact than they would in a diploid autosomal population of reduced effective size. Moreover, we observe in Figure 3.8(b) that a selection coefficient not much stronger than $s = -2 \times 10^{-5}$ can result in the reduction of Neanderthal introgression observed on the X-chromosomes in the 1000 Genomes data.

To explore the possibility of hybrid male infertility, we modified the simulations for the X-chromosome as follows. In addition to the global selection with coefficient s against Neanderthal alleles at all loci, we designated 0.5% of the 5000 loci to be *incompatibility*-loci.



(a) (b)

Figure 3.8: Distribution of amount of Neanderthal introgression on the different chromosomes. Given empirically in CEU and CHB+CHS, and sampled from the present day generation under the Wright-Fisher simulations for different selection coefficients, in 16 replicates. (a) The distributions on the autosomes. (b) The distributions on the X-chromosomes.

These *incompatibility* loci only affect fitness in male individuals. The fitness is multiplied by $(1 - s_I)^{4\frac{C}{M}(M-C)}$, where s_I is the selection coefficient against incompatibility, M is the total number of *incompatibility* loci and C is the number of Neanderthal alleles a male individual carries at these loci. The exponent is proportional to the number of incompatible pairs. Thus, if an individual carries only modern human or only Neanderthal alleles at these loci, the exponent is zero, and the fitness is not affected. The exponent equals its maximal value of $\frac{M}{2}$ when $C = \frac{M}{2}$, that is, half of the *incompatibility* loci carry the Neanderthal allele, and the other half carries the human allele. Figure 3.9 depicts the results of the simulations for $s = -10^{-5}$, and different *incompatibility* selection coefficients s_I. Note that the order of magnitude of s_I is higher then s, because it is acting on fewer loci. The simulations show that a mechanism like this type of hybrid incompatibility could indeed decrease the introgression further than global weak selection against Neanderthal variants by itself, although such an explanation is not necessary to fit the observed levels of introgression.

3.4 Discussion

In this paper, we introduced a modification of the method `diCal 2.0`, which was developed by [74] to infer complex demographic histories from full-genomic sequence data. We applied this modification (`diCal-admix`) to detect tracts of genetic material in modern non-African

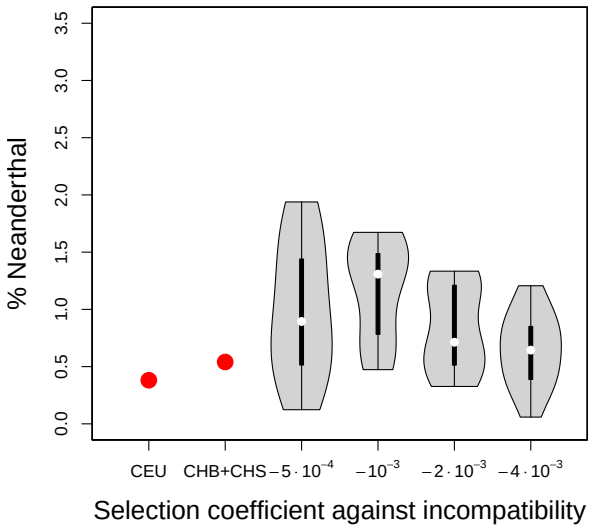


Figure 3.9: Distribution of amount of Neanderthal introgression on the X-chromosome. Given empirically in CEU and CHB+CHS, and sampled from the present day generation under the Wright-Fisher simulations for different models of male hybrid infertility, in 16 replicates.

individuals that introgressed into the population when non-Africans and Neanderthals exchanged genetic material about $2,000$ generations ago. We demonstrated in an extensive simulation study that `diCal-admix` can accurately and efficiently detect tracts of Neanderthal introgression. Furthermore, we applied `diCal-admix` to detect introgression in the individuals sampled from the CEU, CHB, and CHS populations as part of the 1000 Genomes Project [79]. We exhibited some of the methodological and empirical differences between `diCal-admix` and previous results reported by [6] and by [60]. While they are generally in good agreement, we observed some differences. This highlights the importance of the development of different methodologies to generate a consensus. We also reported some of the functional implications of introgression, which confirms previously reported findings of wide-spread selection against introgression, enrichment of Neanderthal introgression in certain classes of genes, but a general signal of depleted Neanderthal introgression in conserved regions of the genome.

However, the role of the X-chromosome remains intriguing. As in previous studies, we observe a substantially lower amount of Neanderthal introgression on the X-chromosome compared to the autosomes. [6] hypothesized that this reduction is due to DMIs reducing male fertility, further supported by significantly reduced Neanderthal introgression in genes expressed in testes. However, we do not find significant evidence for DMIs, and the GO-term enrichment analyses did not reveal any patterns of enrichment or depletion of Neanderthal ancestry in genes related to spermatogenesis, the testes, or infertility more broadly.

Additionally, compared to modern humans and Neanderthals (separated by tens of thousands of generations), the species in which DMIs have been observed are substantially more diverged—e.g., about 20 million generations between *D. simulans* and *D. melanogaster* [104]; 200 to 500 thousand generations between *M. guttatis* and *M. nasutus* [105]; and 750 thousand to 5 million generations between *A. californiense* and *A. tigrinum mavortium* [106, 107]. The divergence time separating modern humans and Neanderthals is only about a factor of two older than the divergence time between the most diverged human populations (e.g., as inferred by [108] and [109] using an updated mutation rate as discussed in [76]) and no DMIs are known to occur in admixtures of modern human populations, raising the question of how such incompatibilities between modern humans and Neanderthals could have arisen so quickly.

Other studies [8, 7] and our simulations suggest that only a moderate strength of selection is required to explain the observed reduction on the X-chromosome. However, the evidence that has been collected to date does not seem to be sufficient to fully characterize the importance of the X-chromosome. Resolving these questions will require a more comprehensive analysis of larger samples of contemporary genetic data like the Simons Genome Diversity Project [110] and the individuals from Phase III of the 1000 Genomes Project [79]. Moreover, additional high-quality data for hominin sister groups [57, 59] will improve the detection of introgression. Detecting introgression in genetic samples of ancient modern humans [56] will also allow resolution of the evolutionary trajectory of introgressed genetic material over time. Incorporating the distribution of tracts on an individual level and different models of diploid selection into the inference frameworks will improve the inference of the strength of selection and allow reliable testing of different models. Additionally, a better understanding of the evolution of incompatibilities, and more careful investigation of the gene content on the X-chromosome will help shed light on the role of the X-chromosome in the Neanderthal introgression landscape. In general, maps of introgressed Neanderthal and Denisovan ancestry will facilitate the interpretation of patterns of human genomic variation and further the understanding of how archaic introgression influenced the trajectory of human evolution.

Chapter 4

Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations

This is joint work with Yun S. Song. Jane A. Yu downloaded and processed chromatin data and Ethan Jewett tested our software. This work is available as a preprint on the *bioRxiv* [111].

4.1 Introduction

Meiotic recombination is a fundamental genetic process and a critical evolutionary force which generates haplotypic diversity in sexually reproducing species. In many species, including humans, a zinc finger-containing protein, PRDM9, directs recombination, resulting in hotspots of recombination at its binding sites [112]. Yet, PRDM9 binds ubiquitously throughout the genome, including at promoters, and only a subset of these correspond to recombination hotspots, suggesting that PRDM9 binding may be necessary but not sufficient [113]. PRDM9 is capable of trimethylating H3K4 and H3K36 [114], and in species that lack a functional copy of *PRDM9*, recombination is concentrated at promoters [115], indicating that chromatin structure plays a role in recombination [116].

PRDM9-directed recombination has fundamental consequences: recombination hotspots partition the genome into blocks with low inter-block linkage but high intra-block linkage, shaping patterns of linked selection [71]. Additionally, an excess of sites where PRDM9 binds one chromosome but not its homolog can lead to male sterility [117, 118]. Such asymmetric binding sites are common in inter-species hybrids, providing a mechanism for the long-known phenomenon of *PRDM9* acting as a speciation gene [119]. Furthermore, asymmetric binding followed by the introduction of a double-strand break and subsequent homology-directed

repair results in meiotic drive against the PRDM9-binding allele, which is equivalent to genic selection at the population level [120]. Over evolutionary timescales, this meiotic drive erodes the binding sites of PRDM9, generating strong positive selection on *PRDM9* mutants with new binding sites [121, 10], explaining why *PRDM9* is one of the fastest evolving genes [122]. These evolutionary dynamics have been studied theoretically [120, 123] and between species [10], but previous empirical investigations have been primarily qualitative rather than quantitative.

We developed a new method, called `pyrho`, to infer fine-scale recombination rates while taking population demography into account and applied it to 26 diverse human populations from phase 3 of the 1000 Genomes Project (1KG) [79]. We then used the resulting accurate, high-resolution maps to investigate the determinants, impacts, and dynamics of recombination rate variation. Software implementing our method and the inferred recombination maps are available at `https://github.com/popgenmethods/pyrho`.

4.2 Fast, accurate inference of fine-scale recombination rates

Our method uses polymorphism data from unrelated individuals to infer fine-scale recombination maps and can be applied to either phased or unphased data. We make use of a composite likelihood approach [124, 125, 126] that has been shown to have favorable statistical properties [127], but unlike previous methods we avoid computationally expensive Markov chain Monte Carlo (MCMC) by using a penalized likelihood framework and gradient-based optimization [128, 129]. Increasing computational efficiency by moving from a Bayesian formulation to a frequentist formulation is a common approach (e.g. [130]). Our approach is between 10 and 450 times faster than `LDhat` [125], a popular MCMC-based method, while improving accuracy (Section 4.6, Figure 4.1A, Figure 4.6, and Table 4.2). We also make use of our recent work on computing two-locus likelihoods [131]: this allows us to scale to hundreds of individuals whereas `LDhat` can accommodate at most 100 diploid individuals, and, importantly, enables us to account for non-equilibrium demographic histories. Failing to account for past fluctuations in population size has been shown to significantly impact the accuracy of inferred fine-scale recombination rates [132, 131, 133]. The details of our method are presented in Section 4.6.

Using samples of unrelated individuals, we are able to produce more accurate, higher resolution maps from tens to hundreds of individuals than admixture-based [134, 135] or trio-based methods [87], which require data from thousands or tens of thousands of individuals, making our method applicable to a broader set of species and populations, including unadmixed populations and populations with few sequenced individuals. Indeed, many recent studies have used approaches similar to ours in a number of non-model organisms (e.g., flycatchers [136], monkey flowers [137], house mice [138], and sticklebacks [139]) and would benefit from properly accounting for the demographic histories of these species.

A major difference between our method and trio-based or admixture-based methods is the sex and temporal resolution of our inferred recombination rates. Trio-based study designs are capable of inferring sex-specific fine-scale recombination rates and measure the present-day recombination rate. Our method and admixture-based methods infer time-averaged and sex-averaged recombination maps because they implicitly average over many generations. In the case of admixture-based maps, the inferred recombination maps are an average since the time of admixture, whereas in our case the averaging is over much longer time-scales in a way that depends on sample size but, in humans, is on the order of hundreds of thousands of years. Lager sample sizes will cause the inferred maps to depend more on recent recombination rates, but the exact temporal dependence of such methods depending on sample size is an open theoretical question.

To explore variation in fine-scale recombination rates across human populations, we inferred population size histories for each of the 26 populations in 1KG [79] using `smc++` [140] (Figure 4.2A) and used these size histories to infer population-specific fine-scale recombination maps. Our maps provide a significantly better fit of the observed r^2, a commonly used measure of linkage disequilibrium, especially at finer scales (mean square error between empirical and theoretical quantiles: $p < 1 \times 10^{-5}$ for each population considered–CEU (Utah residents with northern and western European ancestry), CHB (Han Chinese in Beijing, China), and YRI (Yoruba in Ibadan, Nigeria)–for all comparisons between our maps and those inferred in [134, 9, 87, 79]; two-sided permutation test; Section 4.6, and Figures 4.1B and 4.7). This improvement is particularly pronounced in non-European populations, such as YRI, and could be due to unrealistic assumptions of equilibrium demography made by other methods, a mismatch between the populations used to compute the other maps (e.g., the recombination maps from DECODE [87] are inferred using Icelanders), or to previous methods having hyperparameters tuned to European-like demographies.

4.3 Recombination maps reflect demographic history

Our inferred recombination maps are largely concordant between populations, with high correlation between all maps, even at the single base pair resolution (Spearman's $\rho > 0.70$ for all pairs), but some differences remain. As seen in Figure 4.2B, the correlation between recombination maps largely recapitulates known demographic history, clustering continental-level super populations, and at a finer resolution separating northern and southern European populations, and to a lesser extent separating the eastern African Luhya in Webuye, Kenya (LWK) from west African and primarily west African-descended populations. Admixed American populations show similarity to both African and European populations, particularly the Iberian population in Spain (IBS), especially in Puerto Ricans (PUR), providing evidence that the trans-Atlantic slave trade and European colonization, respectively, may have impacted the recombination rates of present-day admixed American populations.

While such correlations in fine-scale recombination rates could be due to increased sharing of recombinations in the genealogy of individuals from more closely related populations, they

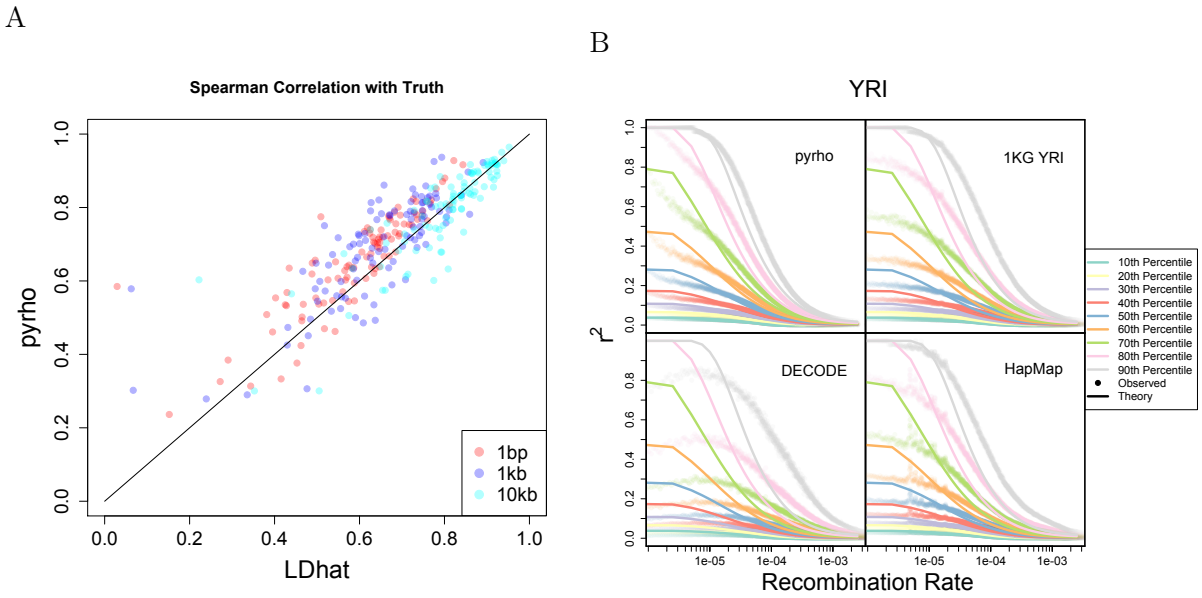


Figure 4.1: Accuracy of `pyrho` on simulated and real data. (**A**) Spearman correlation between inferred and true maps for 100 simulations, each 1 Mb long, for both `pyrho` and `LDhat` with our method showing improved performance especially at finer scales. (**B**) Our inferred recombination maps provide a better fit to observed patterns of linkage disequilibrium as measured by r^2. For a pair of SNPs, r^2, is a random quantity and depends on the rate of recombination between the SNPs. Solid lines show theoretical deciles of this distribution for pairs of sites separated by different recombination distances with minor allele frequency > 0.1 at both sites . Shaded points are the deciles of the empirical distribution obtained by considering pairs of sites with minor allele frequency > 0.1 binned by the recombination rate separating them according to the different recombination maps.

could reflect population-level differences in the determinants of fine-scale recombination rate, such as differences in local chromatin structure, PRDM9 binding site locations, or *PRDM9* alleles. Indeed, there are multiple *PRDM9* alleles that bind different motifs in humans [141], and while the *PRDM9*-A allele predominates in all non-African populations, both the the *PRDM9*-A and *PRDM9*-C alleles are common in African populations, suggesting that African populations may have additional recombination hotspots. This is borne out in our inferred maps, with computationally predicted PRDM9-A binding motifs showing elevated recombination rates in all populations but computationally predicted PRDM9-C binding motifs only showing elevated rates in African populations (Figure 4.2C). While computational prediction of binding motifs for PRDM9 is difficult [142] imperfect predictions should not result in the population-specific elevation of recombination rates within predicted PRDM9-C binding motifs.

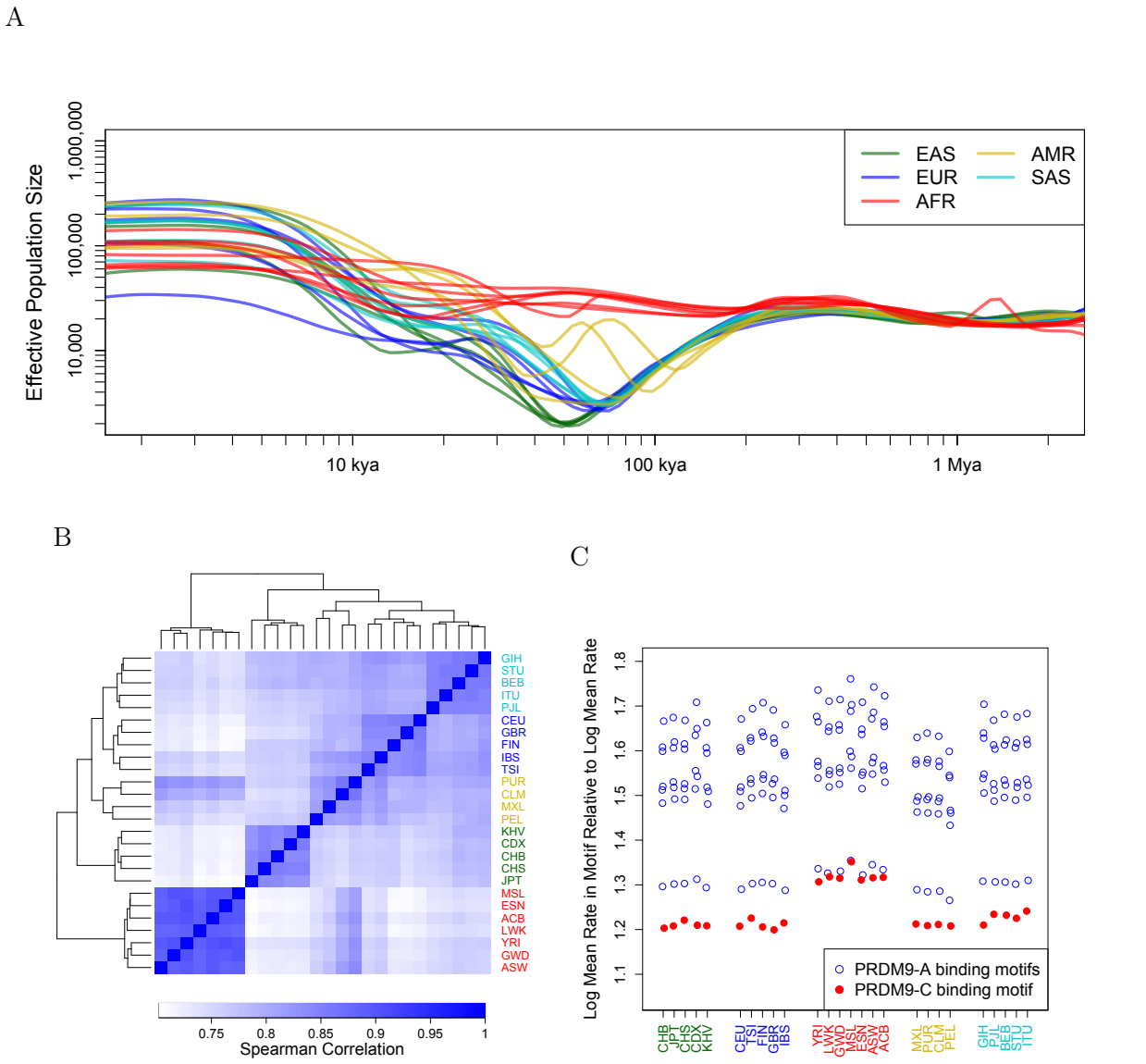


Figure 4.2: Interplay of demographic history and fine-scale recombination rates. (**A**) Population sizes as inferred by `smc++`. All non-African populations show an out-of-Africa bottleneck, which is deepest in east Asian populations. (**B**) Heatmap of the Spearman correlation between the inferred recombination maps at single base pair resolution. All maps show a high degree of correlation, yet the relative correlations agree with continental levels of population differentiation. (**C**) Recombination rates at different PRDM9 binding motifs in each population, normalized by the average recombination rate in that population. All PRDM9-A binding motifs show elevated recombination rates across all populations, while PRDM9-C binding motifs have elevated rates in African populations. Three letter population codes are defined in Table 4.1.

Population Code	Population	Super Population Code
ACB	African Caribbeans in Barbados	AFR
ASW	Americans of African Ancestry in SW USA	AFR
BEB	Bengali from Bangladesh	SAS
CDX	Chinese Dai in Xishuangbanna, China	EAS
CEU	Utah Residents (CEPH) with Northern and Western European Ancestry	EUR
CHB	Han Chinese in Beijing, China	EAS
CHS	Southern Han Chinese	EAS
CLM	Colombians from Medllin, Colombia	AMR
ESN	Esan in Nigeria	AFR
FIN	Finnish in Finland	EUR
GBR	British in England and Scotland	EUR
GIH	Gujarati Indian from Houston, Texas	SAS
GWD	Gambian in Western Divisions in the Gambia	AFR
IBS	Iberian Population in Spain	EUR
ITU	Indian Telugu from the UK	SAS
JPT	Japanese in Tokyo, Japan	EAS
KHV	Kinh in Ho Chi Minh City, Vietnam	EAS
LWK	Luhya in Webuye, Kenya	AFR
MSL	Mende in Sierra Leone	AFR
MXL	Mexican Ancestry from Los Angeles USA	AMR
PEL	Peruvians from Lima, Peru	AMR
PJL	Punjabi from Lahore, Pakistan	SAS
PUR	Puerto Ricans from Puerto Rico	AMR
STU	Sri Lankan Tamil from the UK	SAS
TSI	Toscani in Italia	EUR
YRI	Yoruba in Ibadan, Nigeria	AFR

Table 4.1: Populations in the 1KG dataset [79]. The super populations are **AFR**: African, **AMR**: admixed American, **EAS**: East Asian, **EUR**: European, **SAS**: South Asian.

4.4 Rate of erosion of PRDM9 binding sites

An important consequence of PRDM9-driven recombination is meiotic drive against PRDM9 binding alleles, resulting from homology directed repair of double-strand breaks initiated at the binding motif. While this process has been examined using the divergence between humans and closely related species [10, 143], the magnitude of the effect has not been quantified. As meiotic drive is equivalent to genic selection on evolutionary timescales [120], we may summarize its strength in terms of an effective selection coefficient, s, acting against PRDM9 binding alleles. This selection must be strong enough to explain the substantial divergence between humans and closely related species at PRDM9 binding sites [10, 143], but not so strong as to drive population level differences within humans: male hybrids from species of mice with substantial differences in the locations of PRDM9 binding sites are infertile [118, 117], whereas such incompatibilities obviously do not exist in humans.

To estimate the selection coefficient s, we computationally predicted genomic regions bind PRDM9-A across the autosomes for each haplotype in 1KG and constructed a diallelic sample frequency spectrum (SFS) for each population by treating sequences that can putatively bind PRDM9-A as one allele and sequences that cannot as the alternative allele (Section 4.6). Because PRDM9 is predicted to bind ubiquitously and not all PRDM9 binding sites are recombination hotspots, we subdivided each SFS by local recombination rate. We then used each SFS to infer s while controlling for background selection and misspecification of the demography (Section 4.6). For low to moderate recombination rates, we inferred selection coefficients close to zero, consistent with these PRDM9 binding sites not being "true" recombination hotspots, while for the highest recombination rates, we inferred weak but non-zero selection against the PRDM9 binding allele ($s \approx 5 - 15 \times 10^{-5}$, Figure 4.3A).

The above analysis implicitly assumes that the strength of selection has been temporally constant, which is certainly violated as the motif-determining zinc finger array of *PRDM9* evolves extremely rapidly (e.g., archaic hominins likely do not have the *PRDM9*-A allele) [144]. To address this issue, we constructed an SFS of PRDM9 non-binding alleles that are private to Europeans (or private to East Asians) that have most likely arisen since the divergence of Europeans and East Asians (Section 4.6), and compared the proportion of rare alleles to that seen in a putatively neutral SFS of private SNPs. The observed deficit of singletons and excess of alleles present in more copies (doubletons, tripletons, and quadrupletons) are consistent with an s between 0.5×10^{-4} and 3×10^{-4}, suggesting that our previous estimate of s is likely a lower bound, but of the correct order of magnitude (Figure 4.3B).

Overall, this indicates that the meiotic drive acting against PRDM9 binding sites is equivalent to selection on the order of the inverse of the effective population size, meaning that it is a fairly weak evolutionary force. This is in contradiction to previous assumptions that PRDM9 rapidly erodes its own binding sites [123] and calls into question the hypothesis that this erosion causes the rapid evolution of *PRDM9*. A more plausible explanation for the rapid evolution of *PRDM9* is that a small number of frequently-used hotspots are crucial for the proper segregation of chromosomes during meiosis, and that the strength of meiotic drive at these hotspots is much stronger. PRDM9 bindings sites on short chromosomes–especially

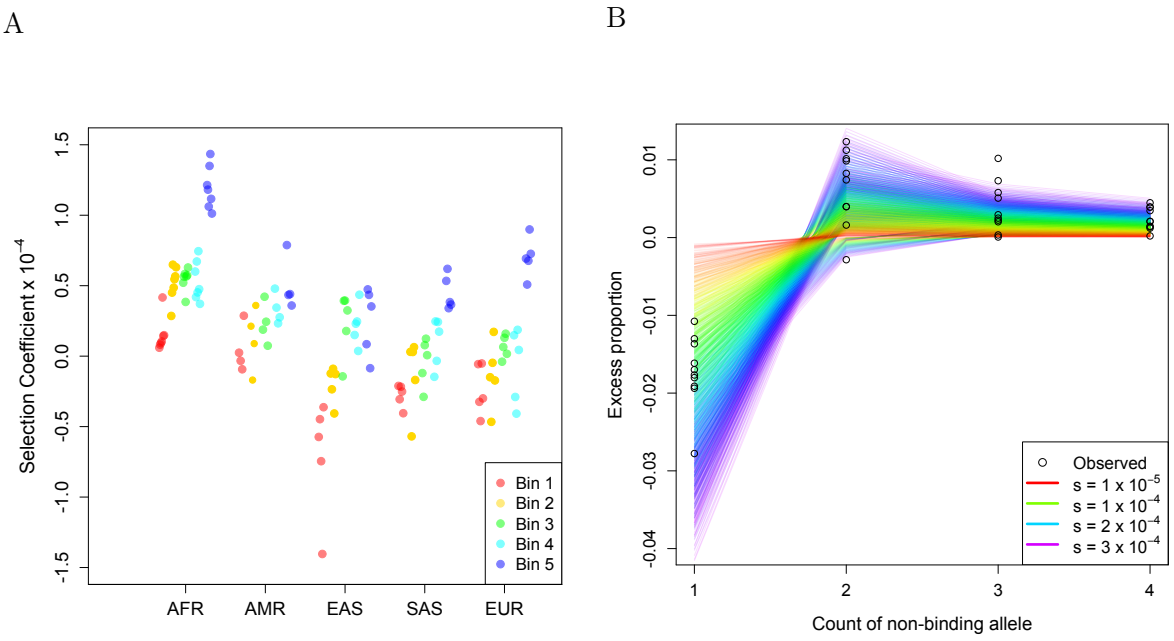


Figure 4.3: Gene conversion acts like weak selection to remove PRDM9 binding sites. (**A**) Strength of selection acting against PRDM9-A binding alleles for different populations in different bins of recombination rate. Bin 1 is for per-generation rates, $r \in [0, 1.45 \times 10^{-9})$, bin 2 is $r \in [1.45 \times 10^{-9}, 2.78 \times 10^{-9})$, bin 3 is $r \in [2, 78 \times 10^{-9}, 5.25 \times 10^{-9})$, bin 4 is $r \in [5.25 \times 10^{-9}, 1.19 \times 10^{-8})$, and bin 5 is $r \in [1.19 \times 10^{-8}, \infty)$. Bins were chosen such that approximately the same number of polymorphic PRDM9 binding sites fall within each bin. Selection is stronger at bins with higher recombination rates. (**B**) The difference in the frequency of singletons, doubletons, tripletons, and quadrupletons between all PRDM9 non-binding alleles private to European or east Asian populations and SNPs matched by recombination rate. Colored lines are the theoretical expectations under varying effective selection coefficients assuming that SNPs evolve neutrally. Different lines of the same color correspond to expectations under different demographic scenarios. Points are the observed differences in the 10 East Asian and European populations. We see a depletion of singletons and excess of more common variants consistent with a selection coefficient between 0.5×10^{-4} and 3×10^{-4}.

in the pseudo-autosomal region in males–are promising candidates because recombination is necessary for proper chromosomal segregation and there are likely only a handful of potential PRDM9 binding sites in such small regions [120, 145]. This would be consistent with our findings since we infer only an average strength of meiotic drive on autosomes, which does not preclude that a small number of sites on the autosomes or sites on the sex chromosomes might be experiencing extremely strong meiotic drive.

4.5 Chromatin affects fine-scale recombination rates

Because fine-scale recombination rates vary substantially even outside of PRDM9-driven hotspots, we also searched for modulators of fine-scale recombination rates beyond PRDM9, finding a role for chromosome length, distance to the telomere, and chromatin state. Specifically, there is a nearly linear relationship between total physical and total genetic length across chromosomes, with a significantly positive slope and intercept (Figure 4.8A; slope, $p = 7.76 \times 10^{-13}$; intercept $p = 1.30 \times 10^{-7}$). The positive intercept confirms that chromosomes require some minimum number of crossovers during meiosis, while the positive slope indicates that longer chromosomes can and do have more crossovers. Furthermore, recombination rates are elevated in subtelomeric regions (Figure 4.8B), likely due to the geometry of the chromosomes during meiosis [146].

We also found a significant role for chromatin structure in shaping fine-scale recombination rates. We used annotations from chromHMM [147] called on 127 ENCODE epigenomes [148]; because this dataset does not contain calls in gametic cells, we used the most common chromatin state across the 127 cell types as the label for each locus. As a result our chromatin state labels are at best a proxy for the true chromatin state in pre-meiotic cells and there may be significant differences between such cells and the 127 cell types in the ENCODE dataset [149]. Furthermore, as mentioned above, our recombination estimates represent a historical average, whereas the ENCODE dataset measures modern chromatin structure. As both recombination rate and chromatin structure are likely changing over time, there is a mismatch of time-scales. With these caveats in mind, we found that recombination rate varies significantly across chromatin states (Figure 4.4; ANOVA $p < 2.2 \times 10^{-16}$), and that this effect is not driven by differences in background selection (Figure 4.10 and Section 4.6). Repetitive regions of the genome have the lowest recombination rates, consistent with a previous finding that a motif present in THE1B repeats is associated with lower recombination rates [113], and suggesting that recombination suppression in repetitive regions is a broader phenomenon. We also found lower recombination rates in transcribed regions, providing support for the hypothesis that PRDM9 evolved to direct recombination away from functionally important regions [150]. Furthermore, recombination rates are low in "closed" heterochromatic or quiescent regions perhaps because these regions preclude access to the recombination machinery.

We found that chromatin states partially characterized by H3K27me3, especially those called as being repressed by Polycomb group proteins (PcGPs), have the highest recombination rates, suggesting a role for H3K27me3 and PcGPs in meiotic recombination. This connection has been noted before, with PcGPs being recruited to double-strand breaks [151] and disruption of the PcGP repression pathway leading to improper chromosomal segregation [152]. This improper segregation in PcGP mutants may be due to a reduced number of successful crossover events in the absence of the H3K27me3 marks deposited by PcGPs. We also note that the substantial impact of chromatin on local recombination rates, along with differences between chromatin structure in male and female gametic progenitor cells, could explain previously observed sex-specific differences in fine-scale recombination rates [153]. While this

manuscript was in preparation, a pedigree-based analysis of crossover recombinations in a large number of Icelandic parent-offspring pairs also found that H3K27me3 and PcGPs are associated with higher local recombination rates [154], corroborating our finding based on population genetics analysis.

The distribution of PRDM9 binding sites across chromatin states is non-uniform (Figure 4.8D; χ^2 test $p < 2.2 \times 10^{-16}$) and putative PRDM9 binding is associated with a 49% increase in recombination rate (Figure 4.8C, t-test $p < 2.2 \times 10^{-16}$), but the variation in recombination rate across chromatin state cannot be explained by differences in PRDM9 binding ($p < 2.2 \times 10^{-16}$ when controlling for PRDM9 binding status). "Bivalent" chromatin states characterized by active H3K4me1 marks and repressive H3K27me3 marks are particularly enriched for putative PRDM9 binding sites, with over 90% of loci in such states being within 100 bp of a putative PRDM9 binding site. This enrichment cannot be explained by the methyltransferase activity of PRDM9, which trimethylates H3K4 and H3K36 [114], leaving the cause of this enrichment unknown.

To investigate the interplay of PRDM9 and chromatin state, we compared a model where PRDM9 affects recombination rate in a chromatin-independent fashion (independent effects model) with a model where PRDM9 can have different effects in different chromatin contexts (dependent effects model), and found that the dependent effects model fits better (F-test $p < 2.2 \times 10^{-16}$). In spite of favoring the dependent effects model, we found that in most chromatin states, the predicted mean recombination rate is similar to that in the independent effects model (Figure 4.4), indicating that PRDM9 and chromatin state usually act independently. A notable exception is at transcription start sites, where PRDM9 binding is found to have an attenuated effect on recombination rate. This could indicate that the recently discovered ability of PRDM9 to act as a transcription factor may be antagonistic to its role in directing recombination or that PRDM9-independent mechanisms act to suppress recombination at transcriptions start sites [113].

4.6 Materials and methods

Gradient-based estimation of fine-scale recombination rates

Penalized composite-likelihood method for phased data

To infer a fine-scale recombination map using n haplotypes with L SNPs, an obvious first approach would be to attempt to either maximize the likelihood

$$\max_{\rho_1,\ldots,\rho_{L-1}} \mathbb{P}\left[(h_{i\ell})_{(i:1\ldots n),(\ell:1\ldots L)} | \rho_1, \ldots, \rho_{L-1}\right]$$

or obtain a posterior

$$\mathbb{P}\left[\rho_1, \ldots, \rho_{L-1} | (h_{i\ell})_{(i:1\ldots n),(\ell:1\ldots L)}\right] \propto \mathbb{P}\left[(h_{i\ell})_{(i:1\ldots n),(\ell:1\ldots L)} | \rho_1, \ldots, \rho_{L-1}\right] \mathbb{P}\left[\rho_1, \ldots, \rho_{L-1}\right]$$

where $\rho_1, \ldots, \rho_{L-1}$ are the recombination rates between each pair of adjacent SNPs, and $h_{i,\ell}$ is the allele of haplotype i at position ℓ. Unfortunately, the full-likelihood of the data is

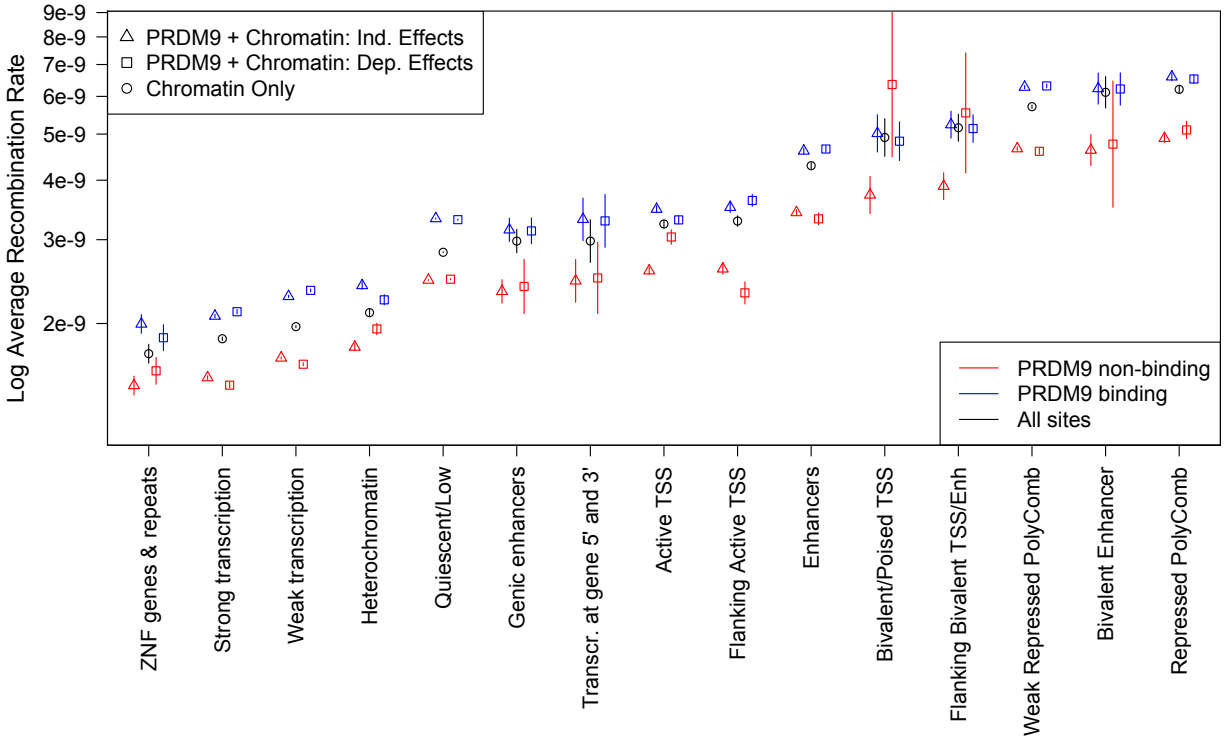


Figure 4.4: PRDM9 and chromatin structure shape fine-scale recombination rates. Different chromatin states have substantially different average recombination rates as determined by fitting a model using only chromatin state (Chromatin only), a model with independent chromatin state and PRDM9 binding effects (PRDM9 + Chromatin: Ind. Effects), and a model where PRDM9 binding may have a different effect in different chromatin states (PRDM9 + Chromatin: Dep. Effects). Sites characterized by H3K27me3 marks (bivalent states and regions repressed by Polycomb) have the highest recombination rates, while repetitive regions, transcribed regions, and heterochromatic or quiescent regions all have depressed recombination rates. ZNF: Zinc Finger Genes, TSS: Transcription Start Site

intractable. Many methods then make the following approximation [124] (but see [155, 156, 3, 157], which use machine learning or regression approaches to infer recombination rates based on simulations, and [84, 158] which use hidden Markov models)

$$\mathbb{P}\left[(h_{i\ell})_{(i:1\ldots n),(\ell:1\ldots L)}|\rho_1,\ldots,\rho_{L-1}\right] \approx \prod_{\ell,k:|\ell-k|<w} \mathbb{P}\left[(h_{i,\ell})_{i:1\ldots n},(h_{i,k})_{i:1\ldots n}|\rho_\ell+\cdots+\rho_{k-1}\right]$$

with some window size w, which works well in practice and has attractive theoretical properties [127]. This also has a justification from the composite-likelihood literature [159]. Importantly, this pairwise likelihood only depends on the total recombination rate separating the two points, which suggests that one could precompute these likelihoods for each possible two-locus haplotype configuration at a grid of recombination rates. Recent work [131] has enabled the computation of these likelihoods for sample sizes in the hundreds.

A drawback of this composite-likelihood approach is that it tends to produce extremely variable estimates. To reduce the variance in the estimate, past approaches have included a prior over recombination maps that explicitly enforces smoothness and have used Markov chain Monte Carlo (MCMC) to obtain samples from the composite posterior over recombination maps [125, 160, 126]. Note that these samples are from a "composite posterior" and not a true posterior, because we have replaced the true likelihood with the composite likelihood. Thus, although these methods provide some sense of the uncertainty in the estimated recombination map, the estimated uncertainty is likely to be inaccurate [160]. A downside of MCMC is that it is slow due to the need to repeatedly evaluate the composite likelihood.

We circumvent MCMC by performing penalized composite-likelihood inference. To enforce that recombination maps are smooth, but to allow for some large jumps (e.g., at hotspots) we add an ℓ_1 penalty to the difference of the log of adjacent recombination rates, which is referred to in other settings as the fused-LASSO [128]. Specifically, we seek to solve the following optimization problem:

$$\max_{\rho_1,\ldots,\rho_{L-1}} \left\{ \sum_{\ell,k:|\ell-k|<w} \log \mathbb{P}\Big[(h_{i,\ell})_{i:1\ldots n},(h_{i,k})_{i:1\ldots n}\Big|\rho_\ell+\ldots+\rho_{k-1}\Big] - \lambda \sum_{\ell=1}^{L-2} |\log(\rho_{\ell+1})-\log(\rho_\ell)| \right\}.$$

Note that this is a high-dimensional optimization problem, making derivative-free optimization methods prohibitively slow. We therefore seek to compute gradients of the likelihood with respect to $(\rho_1,\ldots,\rho_{L-1})$, which is problematic because we have replaced exact evaluation of the pairwise log-likelihoods by looking up entries in a precomputed table. To sidestep this issue, we linearly interpolate between the precomputed log-likelihoods, which makes computing gradients an elementary exercise in linear algebra. Note that, due to using linear interpolation, there are non-differentiable points of the likelihood function, but we circumvent this issue by arbitrarily using the slope of the line infinitesimally to the right of any non-differentiable point. We find that this does not affect the results dramatically. Furthermore, for values of the recombination rate that lie outside of the ranges precomputed in the lookup table, we use the closest entry in the lookup table (either the maximum or

minimum recombination rate in the table) and treat the derivative as zero. We utilize these gradients in a proximal gradient descent method for fused-LASSO problems [129]. We found that this optimization scheme usually converges within tens of evaluations of the objective function, making it highly efficient.

One further subtlety is that this optimization problem is non-convex, implying that there may be local optima in which our optimization scheme could get stuck. One could initialize the optimization at a number of random points and then take the best result, but we take an alternate approach. We first perform a univariate minimization, treating the region as having a single, constant recombination rate. We then use this estimate as our initialization, which should further regularize the optima we find toward being "close" to the constant recombination map.

To further speed up inference, we divide the genome into windows that contain 4001 SNPs that overlap by 100 SNPs and optimize each window independently. We then trim the inferred recombination rates corresponding to the first and last 50 SNPs from each window, and combine the resulting estimates to obtain a recombination map. This process of windowing the genome allows us to run many optimizations in parallel.

Our method is implemented in python and makes extensive use of `numba` [161], a just-in-time LLVM compiler for python, to optimize numerical routines. We also make use of `cyvcf2` [162] to enable the rapid parsing of VCF, bgzipped VCF, and BCF file formats.

Handling unphased data

Our method can also handle unphased data for genotypes from diploid organisms. In principle, one would want to maximize

$$\max_{\rho_1,\ldots,\rho_{L-1}} \sum_{h \text{ consistent with } g} \mathbb{P}\left[(h_{i\ell})_{(i:1\ldots n),(\ell:1\ldots L)} | \rho_1, \ldots, \rho_{L-1}\right],$$

where g is the observed unphased data and "h consistent with g" would be the set of phased haplotypes that are equivalent to g when unphased. We could then apply our composite-likelihood approximation to obtain

$$\max_{\rho_1,\ldots,\rho_{L-1}} \sum_{h \text{ consistent with } g} \prod_{\ell,k:|\ell-k|<w} \mathbb{P}\left[(h_{i,\ell})_{i:1\ldots n}, (h_{i,k})_{i:1\ldots n} | \rho_\ell + \cdots + \rho_{k-1}\right],$$

but unfortunately the outer sum is intractable as it requires phasing all sites simultaneously. Furthermore, it would be difficult to compute gradients under this formulation due to the product. Instead, we make a further approximation by swapping the sum and product to obtain

$$\prod_{\ell,k:|\ell-k|<w} \sum_{h \text{ consistent with } g} \mathbb{P}\left[(h_{i,\ell})_{i:1\ldots n}, (h_{i,k})_{i:1\ldots n} | \rho_\ell + \cdots + \rho_{k-1}\right].$$

While this approximation is admittedly dubious, we may arrive at the same result by first using the composite likelihood approximation, and then summing over consistent haplotypes

as follows:

$$\begin{aligned}
&\mathbb{P}\left[(g_{i\ell})_{(i:1\ldots n),(\ell:1\ldots L)}|\rho_1,\ldots,\rho_{L-1}\right] \\
&\approx \prod_{\ell,k:|\ell-k|<w} \mathbb{P}\left[(g_{i,\ell})_{i:1\ldots n},(g_{i,k})_{i:1\ldots n}|\rho_\ell+\cdots+\rho_{k-1}\right] \\
&= \prod_{\ell,k:|\ell-k|<w} \sum_{h \text{ consistent with } g} \mathbb{P}\left[(h_{i,\ell})_{i:1\ldots n},(h_{i,k})_{i:1\ldots n}|\rho_\ell+\cdots+\rho_{k-1}\right].
\end{aligned}$$

In either case we now only need to phase two loci at a time, and having the product on the outside allows us to take the log and obtain a linear expression

$$\sum_{\ell,k:|\ell-k|<w}\left[\log\left(\sum_{h \text{ consistent with } g} \mathbb{P}\left[(h_{i,\ell})_{i:1\ldots n},(h_{i,k})_{i:1\ldots n}|\rho_\ell+\cdots+\rho_{k-1}\right]\right)\right].$$

Furthermore, we may precompute the values inside of the log by using our lookup table of haploid likelihoods at a grid of recombination rates. Summing over the consistent haplotypes may be performed efficiently using equations 10-12 in [124]. We may then use these new precomputed lookup tables as a drop-in replacement when running our optimization scheme.

Benchmarking

Timing

To obtain timings for our method and `LDhat` for a realistic use-case, we computed the time it took to infer a recombination map for chromosome-scale data. Both methods make use of the same precomputed lookup table of two-locus likelihoods, and so we did not benchmark the creation of those tables, which has been done previously [131]. Thus, we compare only the amount of time to infer a recombination map. Using `msprime` [85], we simulated ten replicates of data matching the length of chromosome 1 with the HapMap recombination map [9] and under the demography inferred for CEU, for a sample size of $n = 196$ haploids. Because `LDhat` does not allow for parallelization, we wrote a python script to separate the data into the same overlapping windows used in our method (windows of 4001 SNPs overlapping by 100 SNPs). We ran our method `pyrho` using 32 cores and also used 32 cores to parallelize `LDhat` runs. For the `LDhat` runs, we then used a python script to combine the output of the runs. Because our scripts for splitting and combining the data for `LDhat` are not optimized, we only timed the total runtime of `LDhat` and compared that to the total time `pyrho` required, which is slightly advantageous for `LDhat`. We used the "optimal" hyperparameters for `pyrho` as discussed below and used the default parameters for `LDhat`, which were tuned to a human-like setting. The timings are presented in Figure 4.5, showing that in our simulations `pyrho` was on average at least 10 times faster than `LDhat`. Yet, when generating the 1KG maps, `LDhat` was run on windows of 2000 SNPs, and the MCMC was run for 22.5 million iterations per window, whereas we used only one million iterations per window in our timing benchmark.

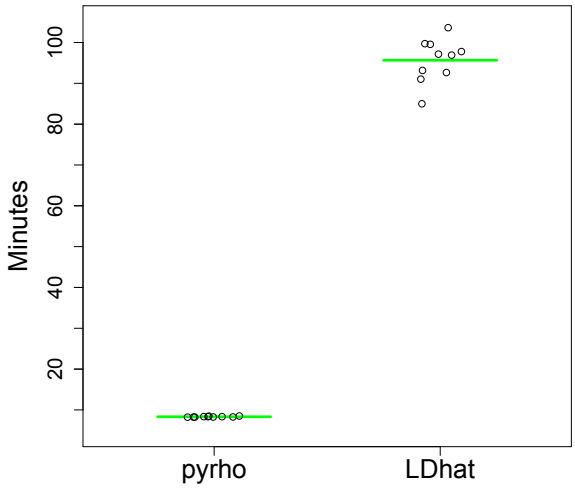


Figure 4.5: Time in minutes to infer a recombination map for a simulated chromosome 1 with 196 haploids, using 32 cores for 10 replicate simulations.

Computing the recombination maps for chromosome 1 for 1KG thus likely took between 22.5 and 45 times longer than the results reported here, suggesting that our method is closer to between 225 and 450 times faster.

Accuracy on simulated data

To assess the accuracy of our method, we used `msprime` [85] to simulate 100 sequences of 1 Mb with recombination maps randomly drawn from the HapMap recombination map [9] under the demography inferred for CEU. We then used the lookup table generated for CEU, which takes demography into account, for `pyrho`, while using a constant-demography lookup table for `LDhat`, as is the default for that program. For each simulation, we took the middle 500 kb and computed the correlation between the true recombination map and the inferred recombination map. We computed the Pearson correlation in both natural and log-scale, and also the Spearman correlation. To avoid issues with auto-correlation we look at windows centered at every $10,000^{\text{th}}$ position. To assess the correlation at different spatial scales, we considered windows of different sizes (1 bp, 1 kb, and 10 kb).

We also investigated whether the differences between `pyrho` and `LDhat` are due to the optimization scheme (i.e., fused-LASSO vs. MCMC) or due to the effect of taking demographic history into account. Using the same simulations as described above, we reran `LDhat` using the demography-aware lookup table used by `pyrho` and computed the same measures of correlation between these inferred maps and the true maps. We found that at fine-scales `pyrho` outperforms `LDhat` by any measure regardless of whether `LDhat` used a constant-demography lookup table or the demography-aware lookup table. At broader scales, `pyrho` outperformed

A B

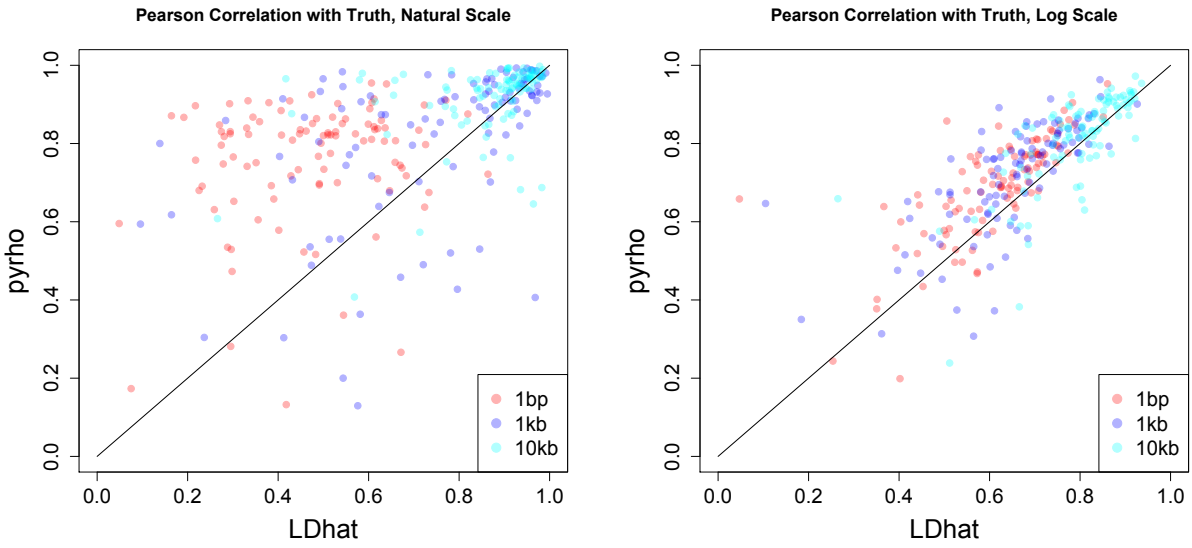


Figure 4.6: Additional measures of accuracy on simulated data. (**A**) Pearson correlation with the truth in natural scale and (**B**) Pearson correlation with the truth in log scale.

`LDhat` if `LDhat` used a constant-demography lookup table, but performed comparably to `LDhat` using the demography-aware lookup table. Meanwhile, the demography-aware version of `LDhat` outperformed the version of `LDhat` that assumed a constant-demography at all scales. The results are summarized in Table 4.2.

Comparison of r^2 on the 1000 Genomes Project dataset

To get a sense of accuracy on real data, we computed a measure of linkage disequilibrium, r^2, between pairs of nearby SNPs. We used `vcftools` [163] with the `-hap-r2`, `-ld-window 15`, `-thin 2000`, `-maf 0.1`, and `-max-missing 1` flags. Briefly, this removes all missing data, removes SNPs until they are all separated by at least 2 kb, and removes SNPs with a minor allele frequency (MAF) less than 10%, and then computes the r^2 for all SNPs within 15 SNPs of each other. For each pair of SNPs we then computed the recombination rate between them as determined by a given fine-scale recombination map. We sorted the pairs of SNPs by the recombination rate between them, and grouped them into bins of 1000 pairs of SNPs, and reported the empirical deciles from that bin. We compared these against the theoretical deciles of the distribution of r^2 for sample sizes matched to the observed sample sizes for SNPs with MAF greater than 10%, which we computed from the lookup tables we generated as discussed below. The results for YRI are presented in Figure 4.1B, and for CEU and CHB in Figure 4.7. See Table 4.1 for a list of three letter population codes.

To determine whether the differences between maps are statistically significant, we

Corr.	pyrho	$\texttt{LDhat}_{\text{demo}}$	$\texttt{LDhat}_{\text{const}}$
Pear., 1bp	$\mathbf{0.756 \pm 0.032}$	0.536 ± 0.033	0.471 ± 0.034
Pear., 1kb	$\mathbf{0.816 \pm 0.039}$	0.778 ± 0.039	0.746 ± 0.043
Pear., 10kb	0.916 ± 0.020	$\mathbf{0.921 \pm 0.018}$	0.864 ± 0.026
Pear., log-scale, 1bp	$\mathbf{0.687 \pm 0.013}$	0.627 ± 0.012	0.610 ± 0.013
Pear., log-scale, 1kb	$\mathbf{0.713 \pm 0.014}$	0.669 ± 0.124	0.648 ± 0.013
Pear., log-scale, 10kb	$\mathbf{0.811 \pm 0.012}$	0.808 ± 0.010	0.791 ± 0.011
Spear., 1bp	$\mathbf{0.659 \pm 0.014}$	0.605 ± 0.014	0.597 ± 0.014
Spear, 1kb	$\mathbf{0.689 \pm 0.014}$	0.653 ± 0.014	0.640 ± 0.014
Spear, 10kb	0.794 ± 0.012	$\mathbf{0.807 \pm 0.012}$	0.799 ± 0.012

Table 4.2: Correlation (Pearson in natural or log-scale and Spearman) at different spatial resolutions. The mean correlation across 100 simulations ($\pm$ 2 standard errors) is reported, and the best performing method for each measure of accuracy is presented in boldface. We present results for our method, `pyrho`, as well as for `LDhat` using a demography-aware lookup table ($\texttt{LDhat}_{\text{demo}}$) or assuming a constant demography ($\texttt{LDhat}_{\text{const}}$). Overall, the methods that take demography into account outperform $\texttt{LDhat}_{\text{const}}$, and `pyrho` substantially outperforms $\texttt{LDhat}_{\text{demo}}$ at fine-scales, and performs comparably at broader scales.

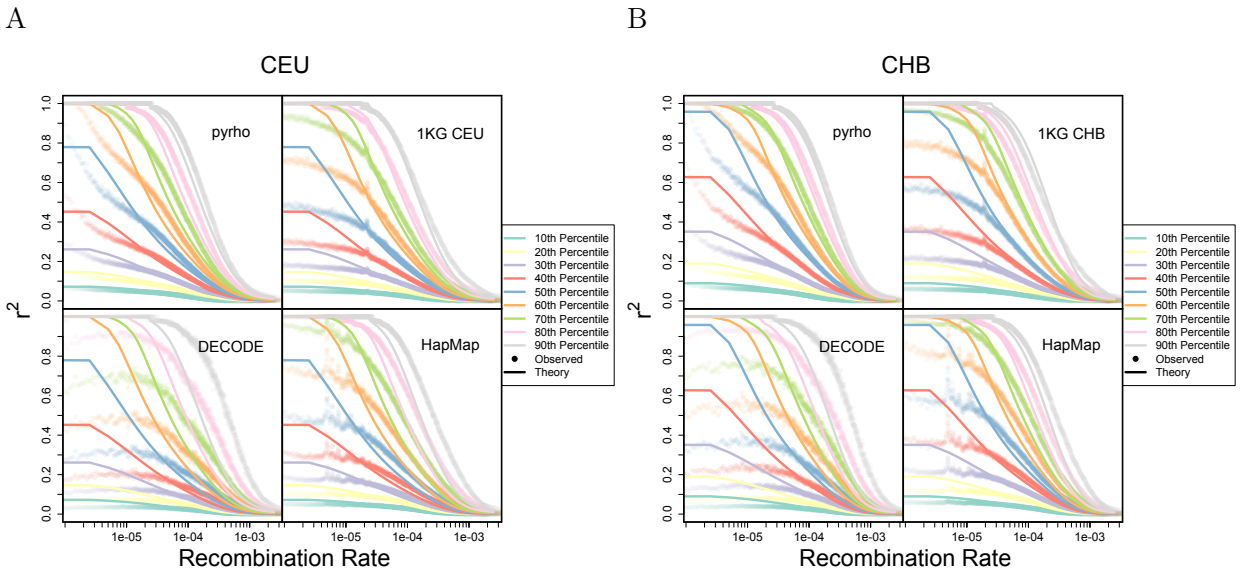


Figure 4.7: Fit of empirical (points) to theoretical (lines) deciles of r^2 between pairs of SNPs with MAF greater than 10% as a function of genetic distance between them.

computed the mean square error between the empirical and theoretical deciles, averaged across all of the bins of pairs of SNPs for different maps. To compare two maps, we used the difference in their mean square error as a test statistic and obtained a null distribution by performing 1,000,000 permutations of the bins (i.e., randomly assigning each bin to one map or the other, making sure each map has the correct number of bins). We compared the maps that we inferred to the LD-based maps, HapMap [9] and 1KG [79]; a trio-based map, DECODE [87]; and an admixture-based map [134]. We performed this comparison for CEU, CHB, and YRI, using the appropriate population for our population-specific maps and the population-specific maps of 1KG. That is, overall we performed 12 comparisons (comparing our maps against 4 others in 3 different populations). For each comparison, we found that our recombination maps had a lower mean square error between the empirical and theoretical deciles with no permutations providing an equal or greater improvement, conservatively implying $p < 1 \times 10^{-5}$ for each comparison.

Inference of population size histories

We applied `smc++` (v1.11.1) [140] to infer population size histories using a previous build of the genome (hg19). All individuals for a given population were included in the analysis with the first 5 individuals (alphabetically by sample name) being used as "distinguished" individuals in the composite likelihood. We assumed a mutation rate of 1.25×10^{-8} per-base per-generation, and masked out sites according to Stephan Schiffels' mappability mask available at https://oc.gnz.mpg.de/owncloud/index.php/s/RNQAkHcNiXZz2fd. Otherwise all default parameter settings of `smc++` were used, and a generation time of 29 years [164] was used to convert generations to years.

Lookup table generation

When using the population sizes inferred in the previous section to build lookup tables for `pyrho`, we made some approximations to reduce the computational cost. The population size functions returned by `smc++` using the `plot` command are piecewise constant, with many pieces. To reduce the number of pieces, we started at present and combined adjacent pieces by taking the harmonic mean of the population sizes for those pieces (weighted by their lengths) if all of the pieces that were combined had population sizes within 10% of the resulting harmonic mean. Furthermore, computing the initial stationary distribution of two-locus configurations, which depends on the most ancient population size, is computationally expensive, and so after reducing the number of pieces, the most ancient size was set to 19,067 for all populations. Computing the exact two-locus likelihoods requires $O(n^6)$ time, where n is the sample size, and is too computationally prohibitive for sample sizes in the hundreds for 26 populations. In previous work [131], we showed that downsampling approximate two-locus likelihoods for a larger sample size, N, results in little loss in accuracy, and these approximate likelihoods may be computed in $O(N^3)$ time and downsampled in $O(N^3 \times (N - n))$ time as well. As

such, we used this approximation, with $N = 256$ for each population, downsampling to the observed sample size, which ranged from $n = 122$ to $n = 226$ haploids.

Hyperparameter optimization

Our method has two important hyperparameters, namely the window size w, which determines how far apart pairs of SNPs must be before we ignore them, and the ℓ_1 regularization penalty, λ, that determines the smoothness of resulting map. Because our method is extremely fast, we were able to optimize these parameters for each population to account for differences in sample size and demography. For each population, we used `msprime` [85] to simulate 100 regions, each of 1 Mb in length, with a recombination map randomly drawn from the HapMap recombination map [9] and with sample size matching the observed sample size. On this dataset, we then ran our method with all possible combinations of $(w, \lambda) \in \{30, 40, 50, 60, 70, 80, 90\} \times \{15, 20, 25, 30, 35, 40, 45, 50\}$. Our method does not estimate a recombination rate before the first SNP or after the last SNP, so we took the estimated recombination rate in the region between the first and last SNP for each simulation and concatenated them together into a single vector, and did the same with the true recombination maps under which we had simulated. We then computed the Pearson correlation of these vectors in both natural and log scale, and also the Spearman correlation; we also computed these correlations at broader scales by taking our estimates and dividing them into non-overlapping windows of length 10kb or 100kb and concatenating the average recombination rate within each window and doing the same to the true recombination maps. We also computed the squared ℓ_2 norm between the inferred recombination maps and the true recombination maps in both natural and log scale. We computed all of these quantities for each setting of the hyperparameters. To choose the "best" hyperparameters, we looked at each measure of quality and ranked the hyperparameter settings for that measure (e.g., the hyperparameter setting that produced the smallest square ℓ_2 norm in natural scale between the estimates and the truth would be ranked 1 for that measure). We then chose the hyperparameter setting that minimized the sum of these ranks over all of the measures we considered. Interestingly, non-African populations tended to have higher values of λ and lower values of w than African populations, likely due to the lower SNP density in non-African populations resulting from the out-of-Africa bottleneck.

Prediction of PRDM9-A binding sites and SFS construction

To predict PRDM9-A binding sites, we obtained empirical position weight matrices (PWMs) from [113]. In [113], a number of different motifs are presented, but following that paper we only used their motifs Human1, ..., Human7 as "true" PRDM9-A binding motifs. These PWM matrices describe the probability $p_X(\ell)$ of observing a nucleotide $X \in \{A, C, G, T\}$ for each position ℓ in the motif. To determine a cutoff for whether to call a particular sequence as matching a particular binding motif or not, we generated 10,000,000 random nucleotide sequences by sampling each position independently, and drawing A or T with probability 0.3

and C or G with probability 0.2, which approximately matches the marginal distribution of nucleotides in the human genome. We then computed the log-likelihood, $\log \mathcal{L}$, of each sequence by

$$\log \mathcal{L}^{(i)} := \sum_{\ell=1}^{M} \log \left[p_{X_\ell^{(i)}}(\ell) \right],$$

where $X_\ell^{(i)}$ is the nucleotide at position ℓ in simulation i and M is the length of the motif. We chose the the 9,999,990$^{\text{th}}$ largest log-likelihood as the cutoff for calling a motif. This is equivalent to an approximate p-value of 1×10^{-6}.

We then called PRDM9-A alleles in each haploid sequence in the 1KG dataset on the hg38 genome build as follows. We considered only diallelic SNPs where all individuals have reported genotypes. Sites with more than two alleles or structural variants were treated as missing. Individuals were treated as having the reference allele at all other positions. Then, starting at the first base in the genome, we computed the log-likelihood, as above, for each motif (or its reverse complement) starting at that position, reporting log-likelihoods that are greater than the empirical cutoff for that motif, and then moving to the next base and repeating. We skipped any starting points where any motif overlapped a missing position. Instead of performing this for each haploid individually, we instead constructed all of the unique haplotypes in the dataset that spanned the region from the starting position to the end of the longest motif, and only computed the log likelihood of each motif on these unique haplotypes.

To construct the PRDM9-A binding site SFS, we took these calls and looked for starting positions where some individuals were called as matching one of the PRDM9-A binding motifs, and other individuals were not predicted to bind any PRDM9-A motif. We then treated binding and non-binding as the two alleles and constructed a standard diallelic SFS. We also constructed SFSs for each population by restricting to only sites with a recombination rate inferred in that population within some range.

Inference of selection coefficients

While a number of software packages exist to fit a selection coefficient to an SFS (e.g., [17, 75]), there were a number of peculiarities about the PRDM9 binding SFS that prevented us from using these previous methods; we expect selection to act against PRDM9 binding alleles regardless of whether they are ancestral or derived, and hence we want to "polarize" our SFS by considering the frequency of PRDM9 binding alleles, instead of the frequency of the derived allele or the frequency of the minor allele as is usual. Yet, mutations may act to introduce new PRDM9 binding sites or to disrupt PRDM9 binding, meaning that new mutants may arise at either end of the SFS. To account for this issue, we derived and implemented a method to fit selection coefficients for this particular setting.

Let $\hat{\tau}_n = (\hat{\tau}_{n,1}, \ldots, \hat{\tau}_{n,n-1})$ be the observed PRDM9 binding SFS. That is $\hat{\tau}_{n,k}$ is the number of segregating sites where k individuals have a haplotype that binds PRDM9 and $n-k$ individuals have a haplotype that does not bind PRDM9. As in previous methods [17, 75],

we fit a selection coefficient by maximizing a multinomial log-likelihood:

$$\log \mathcal{L}_{\text{mult}} \propto \sum_{k=1}^{n-1} \hat{\tau}_{n,k} \log \xi_{n,k}(s, \theta_{\text{bind}}, \theta_{\text{nonbind}}), \tag{4.1}$$

where $\xi_{n,k}(s, \theta_{\text{bind}}, \theta_{\text{nonbind}})$ is the probability that a segregating site has k binding alleles given a selection coefficient of s, a rate θ_{bind} of new PRDM9 binding sites appearing via mutation, and a rate θ_{nonbind} of all non-segregating PRDM9 binding sites generating a new non-binding PRDM9 allele. As has been shown previously [14], we have

$$\begin{aligned}
\xi_{n,k}(s, \theta_{\text{bind}}, \theta_{\text{nonbind}}) &= \mathbb{E}_{s,\theta_{\text{bind}},\theta_{\text{nonbind}}}\left[\frac{\hat{\tau}_{n,k}}{\sum_{\ell=1}^{n-1} \hat{\tau}_{n,\ell}}\right] \\
&\approx \frac{\mathbb{E}_{s,\theta_{\text{bind}},\theta_{\text{nonbind}}}[\hat{\tau}_{n,k}]}{\sum_{\ell=1}^{n-1} \mathbb{E}_{s,\theta_{\text{bind}},\theta_{\text{nonbind}}}[\hat{\tau}_{n,\ell}]} \\
&= \frac{\mathbb{E}_{s,1,\theta_{\text{nonbind}}/\theta_{\text{bind}}}[\hat{\tau}_{n,k}]}{\sum_{\ell=1}^{n-1} \mathbb{E}_{s,1,\theta_{\text{nonbind}}/\theta_{\text{bind}}}[\hat{\tau}_{n,\ell}]},
\end{aligned}$$

where the approximation is exact in the limit of small mutation rates and the final equality follows from the fact that absolute scaling of the mutation rates only determines the total number of segregating sites and not their relative proportions, causing a multiplicative factor to cancel in the numerator and denominator. Therefore, we only need to be able to compute $\mathbb{E}_{s,1,\phi}[\hat{\tau}_{n,k}]$, where $\phi = \theta_{\text{nonbind}}/\theta_{\text{bind}}$. Assuming a panmictic population, this expectation depends on both the unscaled effective population size history, $\eta(t)$, as well as s and ϕ.

We have thus far suppressed the dependence of this expectation on η for notational convenience, but now define $m_{n,k}^{s,\phi}(t)$ to be $\mathbb{E}_{s,1,\phi}[\hat{\tau}_{n,k}]$ for the population size history $\tilde{\eta}(t') = \eta(t + t')$. That is, we truncate the population size history at some point t and treat the resulting function as a new population size history to compute the expectation. Furthermore, define $\mathbf{m}_n^{s,\phi}(t) := (m_{n,1}^{s,\phi}(t), \ldots, m_{n,n-1}^{s,\phi}(t))$. The idea behind our method is to set up and solve a system of differential equations of the form

$$\frac{d}{dt}\mathbf{m}_n^{s,\phi}(t) = g(\mathbf{m}_n^{s,\phi}(t), t)$$

to obtain $\mathbf{m}_n^{s,\phi}(0)$, which is our desired expectation. In the case where $s = 0$, this system of equations turns out to be equivalent to the Moran model [165] with a continuous injection of new mutants into classes at the boundary, a result that follows from [166] and is further explored in [24, 167]. That is

$$\frac{d}{dt}\mathbf{m}_n^{0,\phi}(t) = -\big(\mathbf{M}_n(t,0)\big)^T \cdot \mathbf{m}_n^{0,\phi}(t) - \mathbf{e}_1 - \phi\mathbf{e}_{n-1},$$

where the minus signs arise from our convention of having time run backward, $\mathbf{e}_i$ is the i^{th} basis vector, and $\mathbf{M}_n(t,s) \in \mathbb{R}^{n-1\times n-1}$ is the well-known generator of the Moran process

scaled by the population size:

$$\big(\mathbf{M}_n(t,0)\big)_{ij} = \begin{cases} -\frac{i(n-i)}{\eta(t)}, & \text{if } j = i, \\ \frac{i(n-i)}{2\eta(t)}, & \text{if } j = i-i, \\ \frac{i(n-i)}{2\eta(t)}, & \text{if } j = i+1, \\ 0, & \text{otherwise.} \end{cases}$$

In the case where there is selection ($s \neq 0$), there is no closed system of differential equations exactly describing the evolution of this vector [166, 75]. Yet, it is known that the Moran model with selection converges to the Wright-Fisher diffusion with selection in the limit of large n [168]. We therefore approximate the dynamics with selection by the Moran process with selection, and we compute these expectations for a larger sample size and then downsample to our observed sample size. With this approximation, we obtain the following system of equations,

$$\frac{d}{dt}\mathbf{m}_n^{s,\phi}(t) \approx -\big(\mathbf{M}_n(t,s)\big)^T \cdot \mathbf{m}_n^{s,\phi}(t) - \mathbf{e}_1 - \phi\mathbf{e}_{n-1},$$

where

$$\big(\mathbf{M}_n(t,s)\big)_{ij} = \begin{cases} -\frac{i(n-i)}{\eta(t)} - s \times \frac{i(n-i)}{n}, & \text{if } j = i, \\ \frac{i(n-i)}{2\eta(t)} + s \times \frac{i(n-i)}{n}, & \text{if } j = i-i, \\ \frac{i(n-i)}{2\eta(t)}, & \text{if } j = i+1, \\ 0, & \text{otherwise.} \end{cases}$$

Now that we have set up the system of differential equations, we show how to efficiently solve it. We assume that η is piecewise constant, with sizes $\eta_1, \ldots, \eta_{T+1}$ and breakpoints $t_1, \ldots, t_T$, setting $t_0 := 0$ for ease of notation. For convenience denote the lengths of the pieces as $\Delta_1, \Delta_2, \ldots, \Delta_T$ where $\Delta_k = t_k - t_{k-1}$ for $k > 1$, and also let $\widetilde{\mathbf{M}}_n(k,s) := \mathbf{M}_n(t,s)$ for any t in the k^{th} epoch. We begin at the most ancient interval, which runs from t_T to ∞ and has size η_{T+1}. Since this epoch is infinitely long, we can compute $\mathbf{m}_n^{s,\phi}(t_T)$ by finding the stationary distribution of this process. That is, we solve

$$\mathbf{0} = -\big(\widetilde{\mathbf{M}}_n(T+1,s)\big)^T \cdot \mathbf{m}_n^{s,\phi}(t_T) - \mathbf{e}_1 - \phi\mathbf{e}_{n-1}$$

for $\mathbf{m}_n^{s,\phi}(t_T)$, using a sparse linear solver implemented in scipy [169].

Now, assume that we have computed $\mathbf{m}_n^{s,\phi}(t_k)$. We may compute $\mathbf{m}_n^{s,\phi}(t_{k-1})$ by separately considering what happens to mass already in the system and what happens to mass that is injected during this epoch. Mass already in the system simply evolves according to $\widetilde{\mathbf{M}}(k,s)$, so the contribution of existing mass is $\exp\left\{\Delta_k\left(\widetilde{\mathbf{M}}_n(k,s)\right)^T\right\} \cdot \mathbf{m}_n^{s,\phi}(t_k)$ which can be efficiently computed using `expm_multiply` implemented in scipy [170]. For newly arising mass, we

further condition on when the mass arose, resulting in

$$\int_0^{\Delta_k} \exp\left\{t\Big(\widetilde{\mathbf{M}}_n(k,s)\Big)^T\right\}\cdot(\mathbf{e}_1+\phi\mathbf{e}_{n-1})dt = \Big(\widetilde{\mathbf{M}}_n(k,s)\Big)^{-T}\exp\left\{\Delta_k\Big(\widetilde{\mathbf{M}}_n(k,s)\Big)^T\right\}\cdot(\mathbf{e}_1+\phi\mathbf{e}_{n-1}),$$

which can be computed efficiently, again with `expm_multiply` and a sparse linear solver to avoid needing to invert a matrix. Combining, this results in

$$\begin{aligned}\mathbf{m}_n^{s,\phi}(t_{k-1}) = &\exp\left\{\Delta_k\Big(\widetilde{\mathbf{M}}_n(k,s)\Big)^T\right\}\cdot\mathbf{m}_n^{s,\phi}(t_k)\\ &+\Big(\widetilde{\mathbf{M}}_n(k,s)\Big)^{-T}\exp\left\{\Delta_k\Big(\widetilde{\mathbf{M}}_n(k,s)\Big)^T\right\}\cdot(\mathbf{e}_1+\mathbf{e}_{n-1}),\end{aligned}$$

and iterating this computation we arrive at $\mathbf{m}_n^{s,\phi}(0)$ as desired.

Finally, to find selection coefficients, we can numerically maximize Equation (4.1) using Powell's direction set method [171] as implemented in scipy.

To minimize the effect of using the Moran process with selection to approximate the Wright-Fisher process, we used a larger sample size of 256, and then downsampled to the desired sample size. We performed this for each population for each window of recombination rates, using the decimated `smc++` inferred population sizes.

We were concerned about potential biases arising from either misspecification of the population sizes or differences in background selection due to differences in the recombination rate. To alleviate this bias, we also computed selection coefficients using only SNPs as a putatively neutral control. We tabulated SFSs for SNPs in the same recombination bins, say $\hat{\tau}_n^{\text{SNP}}$, and then computed

$$\tilde{\tau}_n^{\text{SNP}} = \alpha\hat{\tau}_n^{\text{SNP}} + (1-\alpha)\text{reverse}\Big[\hat{\tau}_n^{\text{SNP}}\Big]$$

where the reverse$[\cdot]$ operator reverses the indexing of the vector, with α chosen to match the ratio of the 1^{st} to the $(n-1)^{\text{th}}$ entries of the SFS between the SNP and PRDM9 binding cases. We then inferred selection coefficients for $\tilde{\tau}_n^{\text{SNP}}$ using the same method as described above. Our reported de-biased estimate for a given recombination bin and population are then the selection coefficients inferred for the PRDM9 binding SFS minus the selection coefficient inferred for the matched SNP SFS.

Differences in SFS for private mutations

To compute the expected difference between a normalized SNP SFS and a normalized PRDM9 SFS for mutations that are private to a continental group, we computed the expected SFS as above, but instead of starting with the stationary distribution at the most ancient epoch, we

set $\mathbf{m}_n^{s,\phi}(t_{\text{div}})$ to be zero – ignoring all mutations that occurred prior to the divergence time, t_{div}. We then computed the normalized SFS under some selection coefficient s and under neutrality ($s = 0$), and reported the difference. We assumed that t_{div} was 50 ka ago, but the results are qualitatively similar for all the divergence times we tried between 40 ka and 70 ka ago. To compute the observed difference for Europeans, we looked at only PRDM9 binding alleles (or SNPs) where the binding allele (in the SNP case either SNP allele) was at frequency 1 in all African populations and East Asian populations, pooled across all recombination rate bins. We matched the SNP SFS to the PRDM9 binding SFS in terms of recombination rate as follows. First we partitioned SNPs and PRDM9 binding alleles into different bins based on the recombination rate at the SNP or PRDM9 binding site, and computed an unnormalized SFS for the SNPs within each bin. We then computed an overall SNP SFS by weighting each bin proportional to the number of PRDM9 binding alleles within that bin. The binning scheme we chose was

$$\left(-\infty, \frac{e^{-20}}{40000}\right), \left[\frac{e^{-20}}{40000}, \frac{e^{-19}}{40000}\right), \left[\frac{e^{-19}}{40000}, \frac{e^{-18}}{40000}\right), \ldots \left[\frac{e^{0}}{40000}, \infty\right).$$

We then normalized the PRDM9 binding SFS by this weighted SNP SFS and took the difference. We repeated this procedure swapping the roles of Europeans and East Asians.

Data processing and analysis for determinants of recombination rate variation

For the analyses presented in Figure 4.4 and Figure 4.8, we preprocessed the data as follows. We first restricted our analyses to only sites satisfying the previously mentioned mappability mask for which we inferred recombination rates. Then, to partially alleviate issues of spatial dependency, we subsetted these data by taking every 1000^{th} element. Throughout we converted the population-scaled recombination rates inferred by `pyrho` to per-generation rates by multiplying by μ/θ where μ is the per-generation mutation rate (assumed to be 1.25×10^{-8}) and θ is the population-scaled mutation rate (chosen to be 5×10^{-4}). We calculated the expected number of recombinations per chromosome by averaging the subsetted data within each chromosome and then multiplying by the chromosome length. For analyzing the subtelomeres we averaged all entries within the first 10 Mb of each chromosome to obtain an average for the "left subtelomere" and the last 10 Mb for the "right subtelomere". We ignored the missing subtelomeric regions in the acrocentric chromosomes 13, 14, 15, 21, and 22 and only presented results for the right subtelomere for these chromsomes.

For PRDM9 binding we would ideally use actual, measured PRDM9-A binding sites (e.g., determined by ChIP-seq), but no such dataset exists. Binding locations of the PRDM9-B allele were determined by ChIP-seq in [113] and binding locations of a PRDM9 variant were inferred in the mouse genome using affinity-seq [172]. Pratto and colleagues determined putative PRDM9-A binding sites by performing ChIP-seq on DMC1, a protein recruited to double strand breaks, in individuals with different PRDM9 alleles [173]. This approach

is problematic for our purposes because inferring PRDM9-A binding positions by their induced double strand breaks effectively conditions on those binding sites having elevated recombination rates. We were interested in finding genomic features that modulate the effect of PRDM9 binding on recombination rate, which would be impossible if we only included PRDM9-binding sites with high recombination rates. Ultimately, we labeled each position as affected by PRDM9 binding if it is within 100bp of a computationally predicted PRDM9-A allele binding motif. Note that we focused on PRDM9-A because that is the predominant allele in humans, and is primarily responsible for the historical recombinations we implicitly used in our inference of the recombination maps.

When analyzing the effect of putative PRDM9 binding or chromatin status, we performed all our analyses in log-space. In our benchmarking, we found that `pyrho` produces errors that are approximately normally distributed in log-space, making the use of t-tests, ANOVA, and linear models more appropriate in log-space. All statistical tests were performed in R [174].

Comparison with previous recombination maps

We used LiftOver [175] to re-map previously inferred recombination maps to the current genome build (hg38). We compared our maps with maps released with the 1KG project for CEU, CHB, and YRI [79]; the sex-averaged DECODE recombination map [87]; the HapMap recombination map [9]; and the admixture-based maps reported in Hinch *et al.* [134] and Wegmann *et al.* [135]. We then computed correlation (Pearson in natural scale and log-scale, and Spearman, at various spatial resolutions, as described above) between all pairs of maps. The results are presented in Figure 4.9.

Effect of genome build

We inferred recombination maps on both the current genome build (hg38) and the previous genome build (hg19) to explore the effect of using LiftOver [175] to move recombination maps from one coordinate system to another. This is common in practice, with, for example, the DECODE map being originally called on hg18 [87] but commonly used on hg19 following LiftOver. There appears to be only a modest overall effect: even at the single base-pair resolution the Spearman correlation between maps inferred on hg38 and those inferred on hg19 and lifted to hg38 ranged from $\rho = 0.986$ to $\rho = 0.998$ across all populations. Similarly the Pearson correlation in log-space varied from $r = 0.984$ to $r = 0.998$. The Pearson correlation in natural scale was somewhat less reliable, however, ranging from $r = 0.474$ to $r = 0.987$, likely due to the extreme leverage of hotspots in natural scale. The results are summarized in Table 4.3.

Effect of background selection on inferred recombination rates

To investigate the effect of background selection on our inferred recombination rates, we downloaded a genome-wide measure of background selection (B-statistics) from [176]. B-

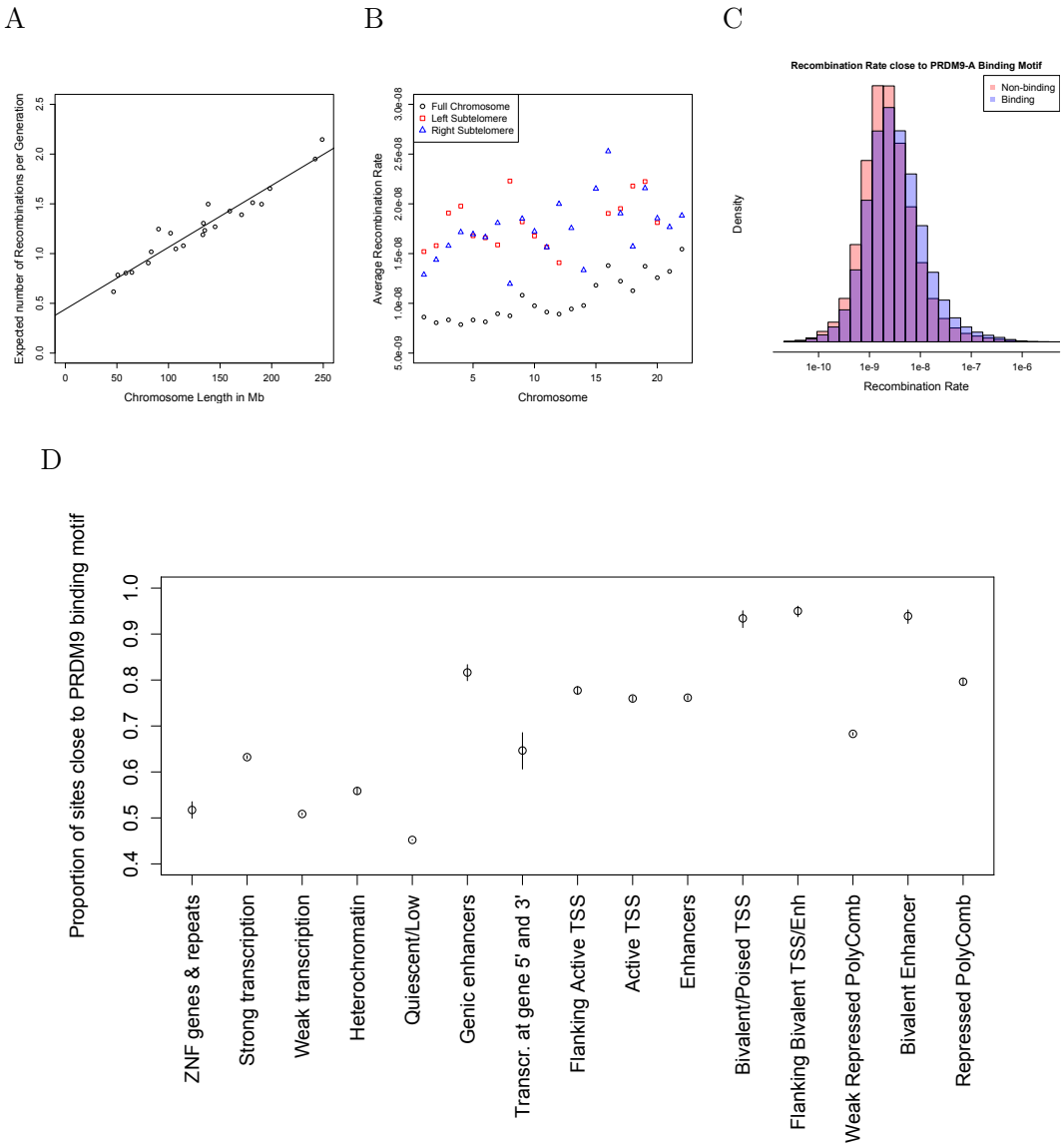


Figure 4.8: (**A**) Expected number of recombinations by chromosome length. The non-zero intercept suggests that there is a minimum number of crossovers required per meiosis, but the positive slope suggests that longer chromosomes can have more than this minimal number. (**B**) Subtelomeres show elevated rates of recombination on all chromosomes. (**C**) Regions within 100bp of a PRDM9-A binding motif have higher recombination rates on average, but the effect explains only a small amount of the variation in recombination rate. (**D**) Sites in some chromatin states are far more likely to be within 100bp of a PRDM9-A binding motif.

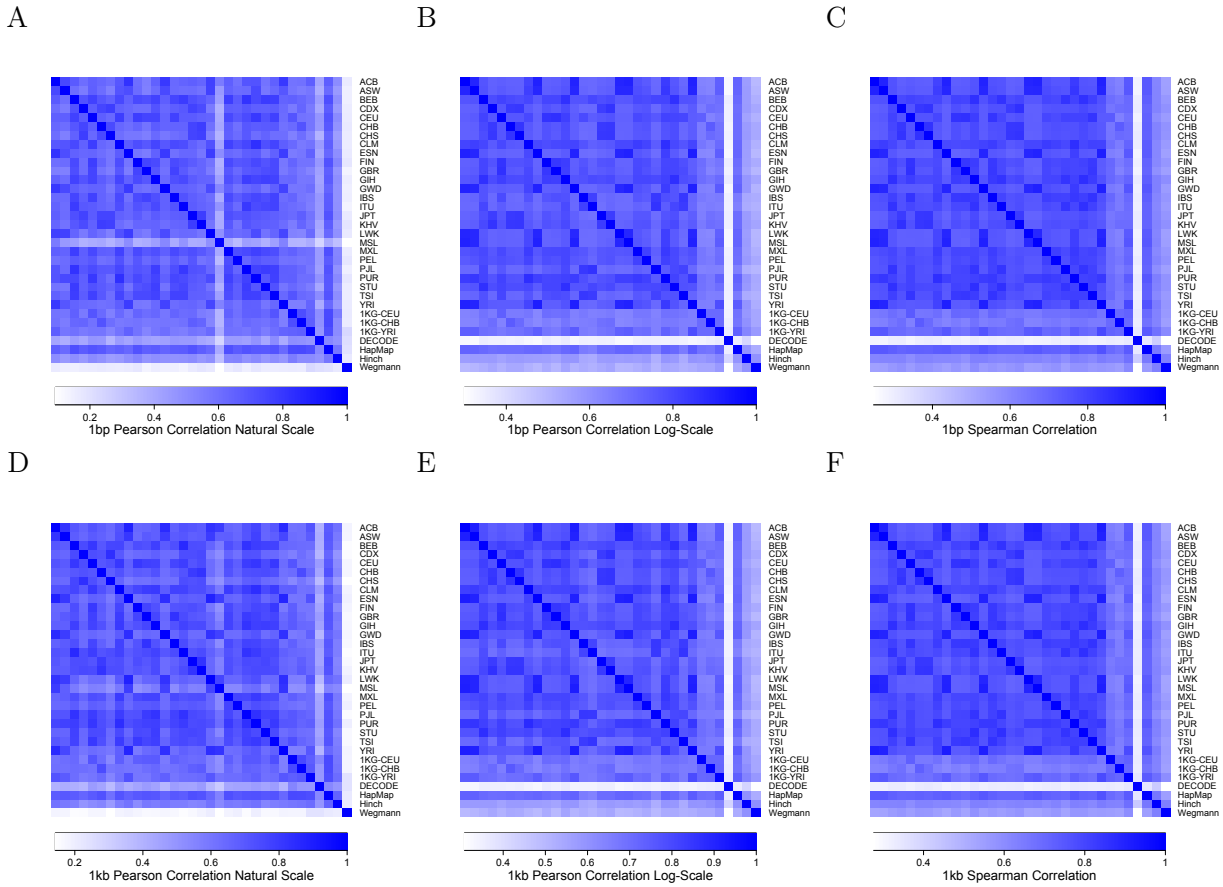


Figure 4.9: Correlation between maps inferred by `pyrho` and maps inferred by previous methods. Results inferred by `pyrho` for each population are indicated by their three letter population code (see Table 4.1). Population specific 1KG maps are described in [79]; DECODE is the sex-averaged DECODE recombination map [87]; HapMap is described in [9]; Hinch in [134]; Wegmann in [135]. Each panel is a different measure of correlation (Pearson in natural and log scales and Spearman) at a different scale (1 bp or 1 kb).

Population	$r^{1bp}_{\text{nat. scale}}$	$r^{1kb}_{\text{nat. scale}}$	$r^{1bp}_{\text{log scale}}$	$r^{1kb}_{\text{log scale}}$	ρ^{1bp}	ρ^{1kb}
ACB	0.814	0.818	0.997	0.997	0.997	0.997
ASW	0.877	0.859	0.997	0.997	0.997	0.997
BEB	0.855	0.867	0.986	0.987	0.987	0.987
CDX	0.958	0.967	0.988	0.988	0.989	0.989
CEU	0.828	0.826	0.991	0.991	0.992	0.992
CHB	0.863	0.843	0.984	0.985	0.986	0.986
CHS	0.916	0.925	0.985	0.985	0.986	0.986
CLM	0.840	0.825	0.991	0.990	0.992	0.991
ESN	0.742	0.748	0.998	0.998	0.998	0.998
FIN	0.986	0.987	0.990	0.990	0.990	0.990
GBR	0.980	0.982	0.990	0.990	0.990	0.990
GIH	0.916	0.913	0.989	0.990	0.990	0.990
GWD	0.492	0.499	0.994	0.994	0.994	0.995
IBS	0.920	0.904	0.988	0.988	0.990	0.990
ITU	0.787	0.799	0.985	0.986	0.986	0.986
JPT	0.904	0.900	0.987	0.987	0.988	0.988
KHV	0.800	0.807	0.989	0.989	0.989	0.989
LWK	0.945	0.944	0.996	0.996	0.996	0.997
MSL	0.693	0.585	0.996	0.997	0.997	0.997
MXL	0.984	0.942	0.991	0.990	0.991	0.991
PEL	0.961	0.885	0.986	0.985	0.987	0.986
PJL	0.928	0.914	0.992	0.984	0.990	0.991
PUR	0.836	0.837	0.994	0.994	0.995	0.995
STU	0.474	0.475	0.984	0.984	0.986	0.986
TSI	0.925	0.895	0.987	0.986	0.987	0.987
YRI	0.920	0.927	0.994	0.994	0.995	0.995

Table 4.3: Correlation between maps inferred on hg38 and those inferred on hg19 and lifted over to hg38. Pearson correlation is denoted by r with the subscript denoting whether it is in log-scale or natural scale. Spearman correlation is denoted by ρ. The amount of smoothing performed is denoted by the superscript. Population codes are listed in Table 4.1

statistics range in value from 0 to 1,000 and reflect the relative loss in genetic diversity as a result of background selection, with 0 being a total loss in diversity and 1,000 representing the truly neutral level of genetic diversity. The available B-statistics are reported in terms of coordinates on hg18, so we used LiftOver [175] to re-map the coordinates to hg38. We took the data as processed in Section 4.6 and further restricted to sites with reported B-statistics. We repeated the analyses of the effect of putative PRDM9 binding and chromatin state while controlling for background selection by including the B-statistics as linear covariates. The results are presented in Figure 4.10. While we observe a high correlation between the inferred recombination rate and the B-statistics (Spearman's $\rho = 0.375$, $p < 2.2 \times 10^{-16}$), the overall impact of chromatin state and PRDM9 binding remains comparable whether or not we control for B-statistics.

Note that B-statistics were originally computed by fitting distributions of selection coefficients for exonic and non-exonic regions to observed patterns of diversity [176]. Importantly, the impact of these distributions on the diversity at linked neutral sites depends on the genetic distance between the selected site and the neutral site, and hence requires knowledge of the fine-scale recombination map. Due to this circularity, it is difficult to determine whether the observed correlation between B-statistics and our inferred recombination rates is due to lower recombination rates directly causing higher levels of background selection (and hence lower inferred recombination rates being associated with lower B-statistics), or if higher levels of background selection result in a lower apparent effective population size, resulting in underestimated recombination rates. It may be possible to disentangle background selection from changes in local recombination rate by jointly inferring B-statistics and fine-scale recombination rates, but we leave such an undertaking for future work.

Results on unphased data

To test the performance of our method on unphased data, we performed the hyperparameter optimization described above for each population with genotype data from diploid individuals. We then tested our method on the same benchmarking data used in Section 4.6 using the optimal hyperparameters for CEU. The results are presented in Figure 4.11, which also shows a scatterplot of the inferred recombination rates compared to the true recombination rates for both phased an unphased data. Both settings are fairly unbiased for all but the smallest recombination rates. For the most part inference using unphased individuals results in performance indistinguishable from that on perfectly phased data. As such we recommend using genotype calls when phasing may be inaccurate.

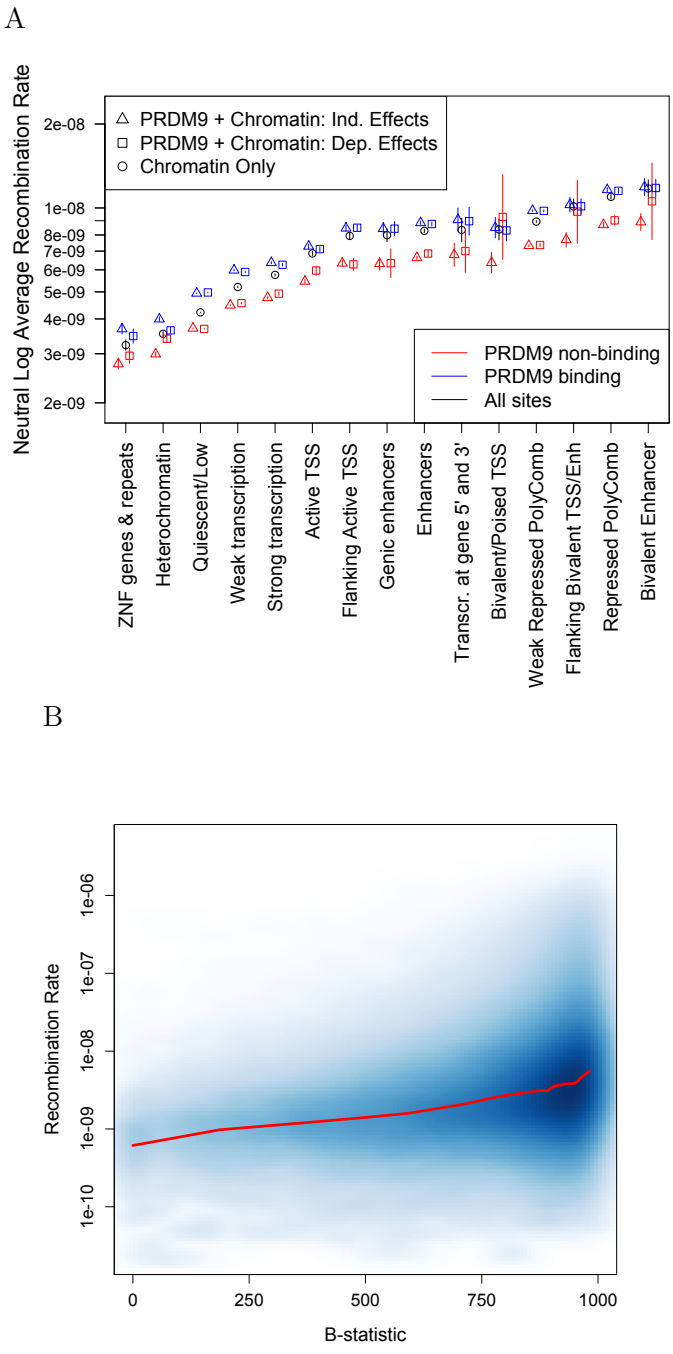


Figure 4.10: (**A**) Inferred regression coefficients after controlling for B-statistics. The results are comparable to Figure 4.4 in terms of relative ordering and relative effect size. ZNF: Zinc Finger Genes, TSS: Transcription Start Site. (**B**) Inferred recombination rates as a function of background selection, presented as a smoothed scatter plot with a local average.

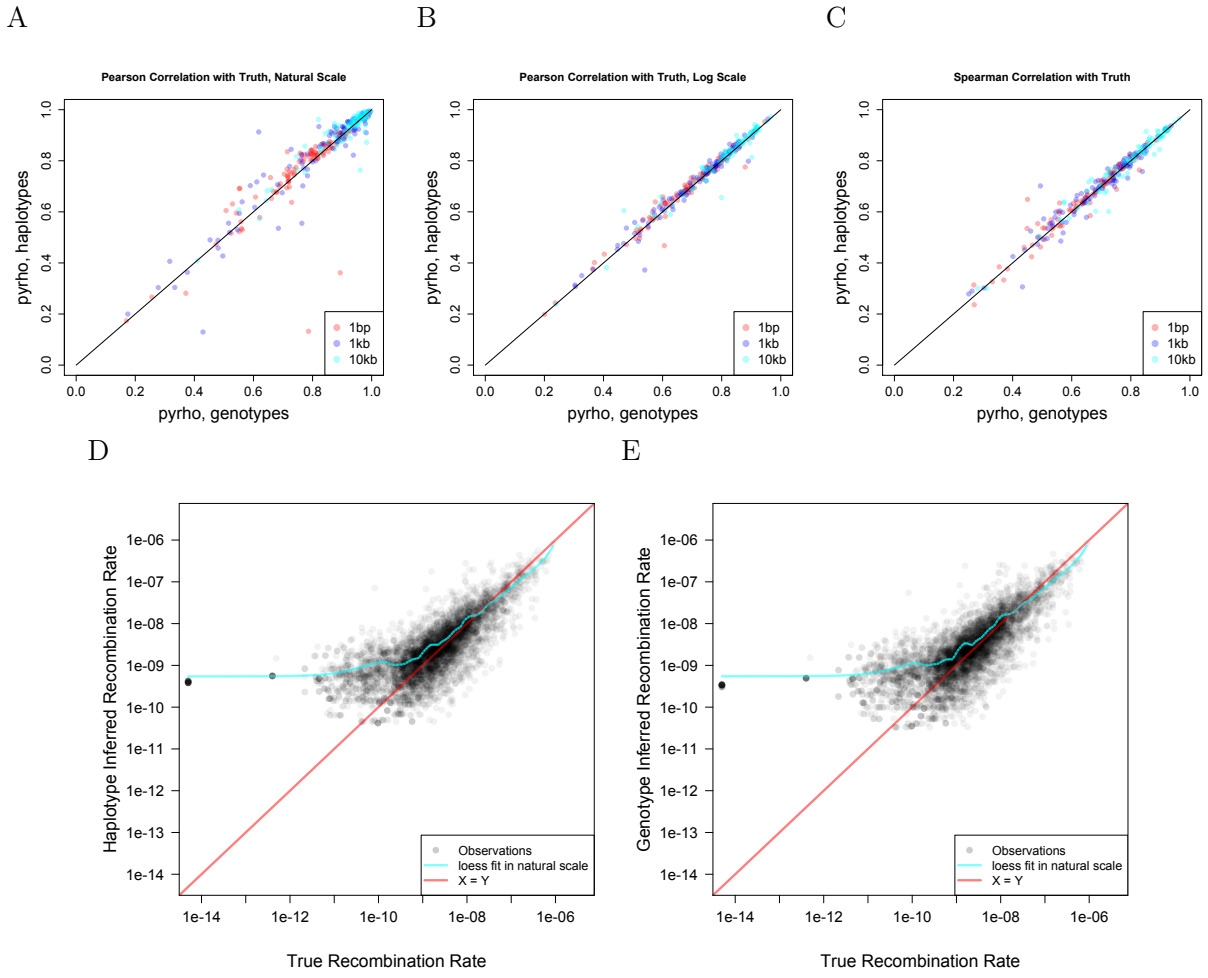


Figure 4.11: Accuracy on simulated data for `pyrho` using phased and unphased data. Accuracy is measured in terms of (**A**) Pearson correlation with the truth in natural scale, (**B**) Pearson correlation with the truth in log scale, and (**C**) Spearman correlation with the truth. (**D**) Scatter plot of the inferred recombination rates using phased data at a 1bp resolution. (**E**) Scatter plot of the inferred recombination rates using unphased data at a 1bp resolution. There is little bias except at the smallest recombination rates when using either phased or unphased data.

Bibliography

[1] John H. Gillespie. *Population genetics: a concise guide*. JHU Press, 2004.

[2] Sara Sheehan and Yun S. Song. "Deep learning for population genetic inference". In: *PLoS computational biology* 12.3 (2016), e1004845.

[3] Lex Flagel, Yaniv Brandvain, and Daniel R. Schrider. "The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference". In: *Molecular Biology and Evolution* 36.2 (2018), pp. 220–238.

[4] Daniel R. Schrider and Andrew D. Kern. "Supervised machine learning for population genetics: a new paradigm". In: *Trends in Genetics* 34.4 (2018), pp. 301–312.

[5] Jeffrey Chan et al. "A likelihood-free inference framework for population genetic data using exchangeable neural networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8594–8605.

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

[7] Kelley Harris and Rasmus Nielsen. "The genetic cost of Neanderthal introgression". In: *Genetics* 203.2 (2016), p. 881.

[8] Ivan Juric, Simon Aeschbacher, and Graham Coop. "The strength of selection against Neanderthal introgression". In: *PLoS Genetics* 12.11 (2016).

[9] Simon Myers et al. "A fine-scale map of recombination rates and hotspots across the human genome". In: *Science* 310.5746 (2005), pp. 321–324.

[10] Simon Myers et al. "Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination". In: *Science* 327.5967 (2010), pp. 876–879.

[11] Jeffrey P. Spence, John A. Kamm, and Yun S. Song. "The site frequency spectrum for general coalescents". In: *Genetics* 202.4 (2016), pp. 1549–1561.

[12] John F. C. Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13 (1982), pp. 235–248.

[13] Yun-Xin Fu. "Statistical properties of segregating sites". In: *Theoretical Population Biology* 48 (1995), pp. 172–197.

[14] Robert C. Griffiths and Simon Tavaré. "The age of a mutation in a general coalescent tree". In: *Communications in Statistics. Stochastic Models* 14.1-2 (1998), pp. 273–295.

[15] Andrzej Polanski, Adam Bobrowski, and Marek Kimmel. "A note on distributions of times to coalescence, under time-dependent population size". In: *Theoretical Population Biology* 63.1 (2003), pp. 33–40.

[16] Andrzej Polanski and Marek Kimmel. "New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth". In: *Genetics* 165.1 (2003), pp. 427–436.

[17] 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.

[18] Anand Bhaskar, Y. X. Rachel Wang, and Yun S. Song. "Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data". In: *Genome Research* 25.2 (2015), pp. 268–279.

[19] Alex Coventry et al. "Deep resequencing reveals excess rare recent variants consistent with explosive population growth". In: *Nature Communications* 1 (2010), p. 131.

[20] Laurent Excoffier et al. "Robust demographic inference from genomic and SNP data". In: *PLoS Genetics* 9.10 (2013), e1003905.

[21] Simon Gravel et al. "Demographic history and rare allele sharing among human populations". In: *Proceedings of the National Academy of Sciences* 108.29 (2011), pp. 11983–11988.

[22] Rasmus Nielsen. "Estimation of population parameters and recombination rates From single nucleotide polymorphisms". In: *Genetics* 154.2 (2000), pp. 931–942.

[23] Feng Gao and Alon Keinan. "Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models". In: *Genetics* 202.1 (2016), pp. 235–245.

[24] John A. Kamm, Jonathan Terhorst, and Yun S. Song. "Efficient computation of the joint sample frequency spectra for multiple populations". In: *Journal of Computational and Graphical Statistics* 26.1 (2017), pp. 182–194.

[25] Jim Pitman. "Coalescents with multiple collisions". In: *Annals of Probability* 27 (1999), pp. 1870–1902.

[26] Serik Sagitov. "The general coalescent with asynchronous mergers of ancestral lines". In: *Journal of Applied Probability* 36.4 (1999), pp. 1116–1125.

[27] Martin Möhle and Serik Sagitov. "A classification of coalescent processes for haploid exchangeable population models". In: *The Annals of Probability* 29.4 (2001), pp. 1547–1562.

[28] Bjarki Eldon and John Wakeley. "Coalescent processes when the distribution of offspring number among individuals is highly skewed". In: *Genetics* 172 (2006), pp. 2621–2633.

[29] Richard Durrett and Jason Schweinsberg. "Approximating selective sweeps". In: *Theoretical Population Biology* 66 (2004), pp. 129–138.

[30] Richard Durrett and Jason Schweinsberg. "A coalescent model for the effect of advantageous mutations on the genealogy of a population". In: *Stochastic Processes and their Applications* 115 (2005), pp. 1628–1657.

[31] Richard A. Neher and Oskar Hallatschek. "Genealogies of rapidly adapting populations". In: *Proceedings of the National Academy of Sciences* 110.2 (2013), pp. 437–442.

[32] Jason Schweinsberg. "Rigorous results for a population model with selection II: genealogy of the population". In: *Electronic Journal of Probability* 22 (2017).

[33] Jason Schweinsberg. "Coalescents with Simultaneous Multiple Collisions". In: *Electronic Journal of Probability* 5 (2000), pp. 1–50.

[34] Thierry E. Huillet. "Pareto genealogies arising from a Poisson branching evolution model with selection". In: *Journal of Mathematical Biology* 68.3 (2014), pp. 727–761.

[35] Matthias Birkner et al. "A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks". In: *ALEA* 6 (2009), pp. 25–61.

[36] Martin Möhle and Serik Sagitov. "Coalescent patterns in diploid exchangeable population models". In: *Journal of Mathematical Biology* 47.4 (2003), pp. 337–352.

[37] Peter Donnelly and Thomas G. Kurtz. "Particle representations for measure-valued population models". In: *The Annals of Probability* 27.1 (1999), pp. 166–205.

[38] Matthias Birkner, Jochen Blath, and Bjarki Eldon. "Statistical properties of the site-frequency spectrum associated with Λ-coalescents". In: *Genetics* 195.3 (2013), pp. 1037–1053.

[39] Jochen Blath et al. "The site-frequency spectrum associated with Ξ-coalescents". In: *Theoretical Population Biology* 110 (2016), pp. 36–50. ISSN: 0040-5809.

[40] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. "Beta-coalescents and continuous stable random trees". In: *The Annals of Probability* (2007), pp. 1835–1887.

[41] Julien Berestycki, Nathanaël Berestycki, and Vlada Limic. "Asymptotic sampling formulae for Λ-coalescents". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 50. 3. Institut Henri Poincaré. 2014, pp. 715–731.

[42] Nathanaël Berestycki. "Recent progress in coalescent theory". In: *Ensaios Matematicos* 16.1 (2009), pp. 1–193.

[43] Martin Möhle and Helmut Pitters. "A spectral decomposition for the block counting process of the Bolthausen-Sznitman coalescent". In: *Electronic Communications in Probability* 19.47 (2014), pp. 1–11.

[44] Erwin Bolthausen and Alain-Sol Sznitman. "On Ruelle's probability cascades and an abstract cavity method". In: *Communications in Mathematical Physics* 197 (1998), pp. 247–276.

[45] Jere Koskela, Paul A. Jenkins, and Dario Spanò. "Bayesian non-parametric inference for Λ-coalescents: Posterior consistency and a parametric method". In: *Bernoulli* 24.3 (2018), pp. 2122–2153.

[46] Matthias Birkner, Jochen Blath, and Bjarki Eldon. "An ancestral recombination graph for diploid populations with skewed offspring distribution". In: *Genetics* 193.1 (2013), pp. 255–290.

[47] Simon Myers, Charles Fefferman, and Nick Patterson. "Can one learn history from the allelic spectrum?" In: *Theoretical Population Biology* 73.3 (2008), pp. 342–348.

[48] Anand Bhaskar and Yun S. Song. "Descartes' rule of signs and the identifiability of population demographic models from genomic variation data". In: *Annals of Statistics* 42.6 (2014), pp. 2469–2493.

[49] Bjarki Eldon et al. "Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents?" In: *Genetics* 199.3 (2015), pp. 841–856.

[50] Einar Árnason. "Mitochondrial cytochrome *b* DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy." In: *Genetics* 166 (2004), pp. 1871–1885.

[51] Dennis Hedgecock and Alexander I. Pudovkin. "Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary". In: *Bulletin of Marine Science* 87.4 (2011), pp. 971–1002.

[52] Anand Bhaskar, Andrew G. Clark, and Yun S. Song. "Distortion of genealogical properties when the sample is very large". In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pp. 2385–2390.

[53] Yun-Xin Fu. "Exact coalescent for the Wright-Fisher model". In: *Theoretical Population Biology* 69.4 (2006), pp. 385–394.

[54] Jonathan Terhorst and Yun S. Song. "Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum". In: *Proceedings of the National Academy of Sciences* 112.25 (2015), pp. 7677–7682.

[55] Matthias Steinrücken et al. "Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans". In: *Molecular Ecology* 27.19 (2018), pp. 3873–3888.

[56] Iain Mathieson et al. "Genome-wide patterns of selection in 230 ancient Eurasians". In: *Nature* 528 (2015), pp. 499–503.

[57] Matthias Meyer et al. "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104 (2012), p. 222.

[58] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505 (2013), pp. 43–49.

[59] Kay Prüfer et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia". In: *Science* 358.6363 (2017), p. 655.

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

[61] Theodosius Dobzhansky. "Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids". In: *Genetics* 21.2 (Mar. 1936), pp. 113–135.

[62] H. Allen Orr. "The population genetics of speciation: the evolution of hybrid incompatibilities." In: *Genetics* 139.4 (1995), p. 1805.

[63] Nicholas J. Brideau et al. "Two Dobzhansky-Muller Genes Interact to Cause Hybrid Lethality in *Drosophila*". In: *Science* 314.5803 (Nov. 2006), p. 1292.

[64] Llla Fishman and John H. Willis. "Evidence for Dobzhansky-Muller incompatibilities contributing to the sterility of hybrids between *Mimulus guttatus* and *M. nasutus*". In: *Evolution* 55.10 (2001), pp. 1932–1942.

[65] Benjamin M. Fitzpatrick. "Dobzhansky–Muller model of hybrid dysfunction supported by poor burst-speed performance in hybrid tiger salamanders". In: *Journal of Evolutionary Biology* 21.1 (2008), pp. 342–351.

[66] Daven C. Presgraves. "Sex chromosomes and speciation in *Drosophila*". In: *Trends in Genetics* 24.7 (2008), pp. 336–343.

[67] Michael Dannemann, Kay Prüfer, and Janet Kelso. "Functional implications of Neandertal introgression in modern humans". In: *Genome Biology* 18.1 (2017), p. 61.

[68] Rachel M. Gittelman et al. "Archaic hominin admixture facilitated adaptation to out-of-Africa environments". In: *Current Biology* 26.24 (2016), pp. 3375–3382.

[69] Rebekah L. Rogers. "Chromosomal rearrangements as barriers to genetic homogenization between archaic and modern humans". In: *Molecular Biology and Evolution* 32.12 (2015), pp. 3064–3078.

[70] Corinne N. Simonti et al. "The phenotypic legacy of admixture between modern humans and Neandertals". In: *Science* 351.6274 (2016), p. 737.

[71] Molly Schumer et al. "Natural selection interacts with recombination to shape the evolution of hybrid genomes". In: *Science* 360 (6389 2018), pp. 656–660.

[72] Sriram Sankararaman et al. "The combined landscape of Denisovan and Neanderthal ancestry in present-day humans". In: *Current Biology* 26.9 (2016), pp. 1241–1247.

[73] Benjamin Vernot et al. "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals". In: *Science* 352.6282 (2016), p. 235.

[74] Matthias Steinrücken, John A. Kamm, and Yun S. Song. "Inference of complex population histories using whole-genome sequences from multiple populations". In: *bioRxiv* (2015), p. 026591. eprint: https://www.biorxiv.org/content/early/2015/09/16/026591.full.pdf.

[75] Julien Jouganous et al. "Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation". In: *Genetics* 206.3 (2017), pp. 1549–1567.

[76] Aylwyn Scally and Richard Durbin. "Revising the human mutation rate: implications for understanding human evolution". In: *Nature Reviews Genetics* 13 (2012), pp. 745–753.

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

[78] Richard E. Green et al. "A draft sequence of the Neandertal genome". In: *Science* 328.5979 (2010), p. 710.

[79] The 1000 Genomes Project Consortium. "A global reference for human genetic variation". In: *Nature* 526 (2015), pp. 68–74.

[80] Vincent Plagnol and Jeffrey D. Wall. "Possible Ancestral Structure in Human Populations". In: *PLoS Genetics* 2.7 (July 2006).

[81] Joshua S. Paul and Yun S. Song. "A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination". In: *Genetics* 186 (2010), pp. 321–338.

[82] Joshua S. Paul, Matthias Steinrücken, and Yun S. Song. "An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination". In: *Genetics* 187 (2011), pp. 1115–1128.

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

[84] Na Li and Matthew Stephens. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". In: *Genetics* 165 (2003), pp. 2213–2233.

[85] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. "Efficient coalescent simulation and genealogical analysis for large sample sizes". In: *PLoS Computational Biology* 12.5 (2016), pp. 1–22.

[86] Benedict Paten et al. "Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs". In: *Genome Research* 18.11 (2008), pp. 1814–1828.

[87] Augustine Kong et al. "Fine-scale recombination rate differences between sexes, populations and individuals". In: *Nature* 467 (2010), pp. 1099–1103.

[88] Eran Eden et al. "Discovering motifs in ranked lists of DNA sequences". In: *PLoS Computational Biology* 3.3 (2007).

[89] Eran Eden et al. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC Bioinformatics* 10 (2009), pp. 48–48.

[90] Bettina Malnic, Paul A. Godfrey, and Linda B. Buck. "The human olfactory receptor gene family". In: *Proceedings of the National Academy of Sciences* 101.8 (2004), p. 2584.

[91] Cathie Sudlow et al. "UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age". In: *PLoS Medicine* 12.3 (2015).

[92] Gregory McInnes et al. "Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics". In: *Bioinformatics* (2018).

[93] Karla L. Miller et al. "Multimodal population brain imaging in the UK Biobank prospective epidemiological study". In: *Nature Neuroscience* 19 (2016), pp. 1523–1536.

[94] 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.

[95] Montgomery Slatkin and Fernando Racimo. "Ancient DNA and human history". In: *Proceedings of the National Academy of Sciences* 113.23 (2016), p. 6380.

[96] Qiaomei Fu et al. "An early modern human from Romania with a recent Neanderthal ancestor". In: *Nature* 524 (2015), pp. 216–219.

[97] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44.D1 (2015), pp. D733–D745.

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

[99] Maria I. Klapa et al. "Reconstruction of the experimentally supported human protein interactome: what can we learn?" In: *BMC Systems Biology* 7.1 (2013), p. 96.

[100] Aris Gioutlakis, Maria I. Klapa, and Nicholas K. Moschonas. "PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology". In: *PLoS ONE* 12.10 (2017).

[101] Fernando Racimo et al. "Evidence for archaic adaptive introgression in humans". In: *Nature Reviews Genetics* 16 (2015), pp. 359–371.

[102] Aaron J. Sams et al. "Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans". In: *Genome Biology* 17.1 (2016), p. 246.

[103] Fernando Racimo, Davide Marnetto, and Emilia Huerta-Sánchez. "Signatures of archaic adaptive introgression in present-day human populations". In: *Molecular Biology and Evolution* 34.2 (2016), pp. 296–317.

[104] Yi-Ju Li, Yoko Satta, and Naoyuki Takahata. "Paleo-demography of the *Drosophila melanogaster* subgroup: application of the maximum likelihood method". In: *Genes & Genetic Systems* 74.4 (1999), pp. 117–127.

[105] Yaniv Brandvain et al. "Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*". In: *PLOS Genetics* 10.6 (2014).

[106] H. Bradley Shaffer and Mark L. McKnight. "The polytypic species revisited: Genetic differentiation and molecular phylogenetics of the tiger salamander *Ambystoma tigrinum* (Amphbia: Caudata) complex". In: *Evolution* 50.1 (1996), pp. 417–433.

[107] Benjamin M. Fitzpatrick et al. "Rapid spread of invasive genes into a threatened native species". In: *Proceedings of the National Academy of Sciences* 107.8 (2010), pp. 3606–3610.

[108] Krishna R. Veeramah et al. "An early divergence of KhoeSan ancestors from those of other modern humans Is supported by an ABC-based analysis of autosomal resequencing data". In: *Molecular Biology and Evolution* 29.2 (2011), pp. 617–630.

[109] Ilan Gronau et al. "Bayesian inference of ancient human demography from individual genome sequences". In: *Nature Genetics* 43 (2011), pp. 1031–1034.

[110] Swapan Mallick et al. "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". In: *Nature* 538 (2016), pp. 201–206.

[111] Jeffrey P. Spence and Yun S. Song. "Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations". In: *bioRxiv* (2019). eprint: https://www.biorxiv.org/content/early/2019/01/28/532168.full.pdf.

[112] Frédéric Baudat et al. "PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice". In: *Science* 327 (2010), pp. 836–840.

[113] Nicolas Altemose et al. "A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis". In: *eLife* 6 (2017), e28383.

[114] Natalie R. Powers et al. "The meiotic recombination activator PRDM9 trimethylates both H3K36 and H3K4 at recombination hotspots *in vivo*". In: *PLoS Genetics* 12.6 (2016), e1006146.

[115] Sonal Singhal et al. "Stable recombination hotspots in birds". In: *Science* 350.6263 (2015), pp. 928–932.

[116] Julian Lange et al. "The landscape of mouse meiotic double-strand break formation, processing, and repair". In: *Cell* 167.3 (2016), 695–708.e16.

[117] Benjamin Davies et al. "Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice". In: *Nature* 530 (2016), pp. 171–176.

[118] Sona Gregorova et al. "Modulation of *Prdm9*-controlled meiotic chromosome asynapsis overrides hybrid sterility in mice". In: *eLife* 7 (2018), e34282.

[119] Jiri Forejt and Pavol Iványi. "Genetic studies on male sterility of hybrids between laboratory and wild mice (*Mus musculus* L.)" In: *Genetical Research* 24.2 (1974), pp. 189–206.

[120] Graham Coop and Simon R. Myers. "Live hot, die young: Transmission distortion in recombination hotspots". In: *PLOS Genetics* 3.3 (2007), pp. 1–10.

[121] Francisco Úbeda and Jon F. Wilkins. "The Red Queen theory of recombination hotspots". In: *Journal of Evolutionary Biology* 24.3 (2010), pp. 541–553.

[122] Zachary Baker et al. "Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates". In: *eLife* 6 (2017), e24133.

[123] Thibault Latrille, Laurent Duret, and Nicolas Lartillot. "The Red Queen model of recombination hot-spot evolution: a theoretical investigation". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 372.1736 (2017).

[124] Richard R. Hudson. "Two-locus sampling distributions and their application". In: *Genetics* 159.4 (2001), pp. 1805–1817.

[125] Gil McVean, Philip Awadalla, and Paul Fearnhead. "A coalescent-based method for detecting and estimating recombination from gene sequences". In: *Genetics* 160.3 (2002), pp. 1231–1241.

[126] Andrew H. Chan, Paul A. Jenkins, and Yun S. Song. "Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*". In: *PLoS Genetics* 8.12 (2012), e1003090.

[127] Paul Fearnhead. "Consistency of estimators of the population-scaled recombination rate". In: *Theoretical Population Biology* 64 (2003), pp. 67–79.

[128] Robert Tibshirani et al. "Sparsity and Smoothness via the Fused Lasso". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67.1 (2005), pp. 91–108.

[129] Jun Liu, Lei Yuan, and Jieping Ye. "An efficient algorithm for a Class of fused Lasso problems". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. Washington, DC, USA: ACM, 2010, pp. 323–332.

[130] David H. Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Research* 19.9 (2009), pp. 1655–1664.

[131] John A. Kamm et al. "Two-locus likelihoods under variable population size and fine-scale recombination rate estimation". In: *Genetics* 203.3 (2016), pp. 1381–1399.

[132] Henry R Johnston and David J Cutler. "Population demographic history can cause the appearance of recombination hotspots". In: *American Journal of Human Genetics* 90.5 (2012), pp. 774–783.

[133] Amy L. Dapper and Bret A. Payseur. "Effects of Demographic History on the Detection of Recombination Hotspots from Linkage Disequilibrium". In: *Molecular Biology and Evolution* 35.2 (2018), pp. 335–353.

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

[135] Daniel Wegmann et al. "Recombination rates in admixed individuals identified by ancestry-based inference". In: *Nature Genetics* 43 (2011), pp. 847–853.

[136] Takeshi Kawakami et al. "Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds". In: *Molecular Ecology* 26.16 (2017), pp. 4158–4172.

[137] Joshua R. Puzey, John H. Willis, and John K. Kelly. "Population structure and local selection yield high genomic variation in *Mimulus guttatus*". In: *Molecular Ecology* 26.2 (2017), pp. 519–535.

[138] Tom R. Booker, Rob W. Ness, and Peter D. Keightley. "The recombination landscape in wild house mice inferred using population genomic data". In: *Genetics* 207.1 (2017), p. 297.

[139] Alice F. Shanfelter, Sophie L. Archambeault, and Michael A. White. "Divergent fine-scale recombination landscapes between a freshwater and marine population of threespine stickleback fish". In: *Genome Biology and Evolution* (2019).

[140] Jonathan Terhorst, John A. Kamm, and Yun S. Song. "Robust and scalable inference of population history from hundreds of unphased whole genomes". In: *Nature Genetics* 49 (2017), pp. 303–309.

[141] Ingrid L. Berg et al. "PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans". In: *Nature Genetics* 42.10 (2010), pp. 859–863.

[142] Laure Ségurel. "The complex binding of PRDM9". In: *Genome Biology* 14.4 (2013), p. 112.

[143] Yann Lesecque et al. "The Red Queen model of recombination hotspots evolution in the light of archaic and modern human genomes". In: *PLoS genetics* 10.11 (2014).

[144] Jerrod J. Schwartz et al. "Primate evolution of the recombination regulator PRDM9". In: *Nature Communications* 5.4370 (2014).

[145] Anjali G. Hinch et al. "Recombination in the human pseudoautosomal region PAR1". In: *PLOS Genetics* 10.7 (2014).

[146] Amandine Batté et al. "Recombination at subtelomeres is regulated by physical distance, double-strand break resection and chromatin status". In: *The EMBO Journal* 36.17 (2017), p. 2609.

[147] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature Methods* 9 (2012), pp. 215–216.

[148] The Roadmap Epigenomics Consortium. "Integrative analysis of 111 reference human epigenomes". In: *Nature* 518 (2015), pp. 317–330.

[149] Yao Wang et al. "Reprogramming of meiotic chromatin architecture during spermatogenesis". In: *Molecular Cell* 73.3 (2019), 547–561.e6.

[150] Adam Auton et al. "Genetic recombination is targeted towards gene promoter regions in dogs". In: *PLoS Genetics* 9.12 (2013), e1003984.

[151] Danny M. Chou et al. "A chromatin localization screen reveals poly (ADP ribose)-regulated recruitment of the repressive polycomb and NuRD complexes to sites of DNA damage". In: *Proceedings of the National Academy of Sciences* 107.43 (2010), p. 18475.

[152] Weipeng Mu et al. "Repression of the soma-specific transcriptome by Polycomb-repressive complex 2 promotes male germ cell development". In: *Genes & Development* 28.18 (2014), pp. 2056–2069.

[153] Claude Bhérer, Christopher L. Campbell, and Adam Auton. "Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales". In: *Nature Communications* 8.14994 (Apr. 2017).

[154] Bjarni V. Halldorsson et al. "Characterizing mutagenic effects of recombination through a sequence-level genetic map". In: *Science* 363.6425 (2019), eaau1043.

[155] Kao Lin, Andreas Futschik, and Haipeng Li. "A fast estimate for the population recombination rate based on regression". In: *Genetics* 194.2 (2013), pp. 473–484.

[156] Feng Gao et al. "New software for the fast estimation of population recombination rates (FastEPRR) in the genomic era". In: *G3: Genes|Genomes|Genetics* 6.6 (2016), pp. 1563–1571.

[157] Philipp Hermann et al. "LDJump: Estimating variable recombination rates from population genetic data". In: *Molecular Ecology Resources* 19.3 (2019), pp. 623–638.

[158] Gustavo Valadares Barroso, Natasa Puzovic, and Julien Dutheil. "Inference of recombination maps from a single pair of genomes and its application to archaic samples". In: *bioRxiv* (2018), p. 452268. eprint: https://www.biorxiv.org/content/early/2018/10/25/452268.full.pdf.

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

[160] Adam Auton and Gil McVean. "Recombination rate estimation in the presence of hotspots". In: *Genome Research* 17.8 (2007), pp. 1219–1227.

[161] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. "Numba: A LLVM-based Python JIT compiler". In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. LLVM '15. Austin, Texas: ACM, 2015, 7:1–7:6.

[162] Brent S. Pedersen and Aaron R. Quinlan. "cyvcf2: fast, flexible variant analysis with Python". In: *Bioinformatics* 33.12 (2017), pp. 1867–1869.

[163] Petr Danecek et al. "The variant call format and VCFtools". In: *Bioinformatics* 27.15 (2011), pp. 2156–2158.

[164] Jack N. Fenner. "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies". In: *American Journal of Physical Anthropology* 128.2 (2005), pp. 415–423.

[165] Patrick A. P. Moran. "Random processes in genetics". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54 (01 1958), pp. 60–71.

[166] Steven N. Evans, Yelena Shvets, and Montgomery Slatkin. "Non-equilibrium theory of the allele frequency spectrum". In: *Theoretical Population Biology* 71.1 (2007), pp. 109–119.

[167] John A. Kamm et al. "Efficiently inferring the demographic history of many populations with allele count data". In: *bioRxiv* (2018). eprint: https://www.biorxiv.org/content/early/2018/03/23/287268.full.pdf.

[168] Stephen M. Krone and Claudia Neuhauser. "Ancestral processes with selection". In: *Theoretical Population Biology* 51.3 (1997), pp. 210–237.

[169] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001.

[170] Awad H. Al-Mohy and Nicholas J. Higham. "Computing the action of the matrix exponential, with an application to exponential integrators". In: *SIAM Journal on Scientific Computing* 33 (2) (2011), pp. 488–511.

[171] Michael J. D. Powell. "An efficient method for finding the minimum of a function of several variables without calculating derivatives". In: *The Computer Journal* 7.2 (1964), pp. 155–162.

[172] Michael Walker et al. "Affinity-seq detects genome-wide PRDM9 binding sites and reveals the impact of prior chromatin modifications on mammalian recombination hotspot usage". In: *Epigenetics & Chromatin* 8.1 (2015), p. 31.

[173] Florencia Pratto et al. "Recombination initiation maps of individual human genomes". In: *Science* 346.6211 (2014), p. 1256442.

[174] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.

[175] W. James Kent et al. "The Human Genome Browser at UCSC". In: *Genome Research* 12.6 (2002), pp. 996–1006.

[176] Graham McVicker et al. "Widespread genomic signatures of natural selection in hominid evolution". In: *PLoS Genetics* 5.5 (2009), e1000471.

Appendix A

Additional results for Chapter 3

A.1 Gene ontology analysis results

Here we present the Gene Ontology enrichment results for Chapter 3. All Gene Ontology categories with an FDR q value, as reported by GOrilla [88, 89], less than 0.05 are reported in Tables A.1, A.2, A.3, and A.4. We note that GOrilla does not properly account for spatial correlations or uncertainty in our mean posterior introgression probabilities and so the reported q values may be anticonservative.

GO Catgory	GO ID	Description	FDR q-value
Process:	GO:0050907	detection of chemical stimulus involved in sensory perception	2.72×10^{-28}
	GO:0050911	detection of chemical stimulus involved in sensory perception of smell	6.64×10^{-28}
	GO:0009593	detection of chemical stimulus	2.6×10^{-27}
	GO:0050906	detection of stimulus involved in sensory perception	1.21×10^{-24}
	GO:0051606	detection of stimulus	4.01×10^{-17}
	GO:0007186	G-protein coupled receptor signaling pathway	4.71×10^{-13}
	GO:0007608	sensory perception of smell	3.31×10^{-6}
	GO:0007606	sensory perception of chemical stimulus	8.09×10^{-5}
Function:	GO:0004984	olfactory receptor activity	3.98×10^{-28}
	GO:0004930	G-protein coupled receptor activity	5.59×10^{-17}
	GO:0005549	odorant binding	1.3×10^{-13}
	GO:0004888	transmembrane signaling receptor activity	5.34×10^{-10}
	GO:0099600	transmembrane receptor activity	4.51×10^{-10}
	GO:0038023	signaling receptor activity	4.71×10^{-9}
	GO:0004872	receptor activity	7.19×10^{-7}
	GO:0060089	molecular transducer activity	9.93×10^{-7}
	GO:0004871	signal transducer activity	1.76×10^{-6}
	GO:0045236	CXCR chemokine receptor binding	0.0273

Table A.1: Gene ontology terms associated with lack of introgression in CHB+CHS

GO Catgory	GO ID	Description	FDR q-value
Process:	GO:0050907	detection of chemical stimulus involved in sensory perception	1.11×10^{-9}
	GO:0009593	detection of chemical stimulus	7.21×10^{-9}
	GO:0050906	detection of stimulus involved in sensory perception	1.41×10^{-8}
	GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1.11×10^{-7}
	GO:0051606	detection of stimulus	2.7×10^{-5}
	GO:0007608	sensory perception of smell	1.34×10^{-3}
	GO:0007606	sensory perception of chemical stimulus	3.04×10^{-3}
	GO:0018149	peptide cross-linking	0.0178
Function:	GO:0004984	olfactory receptor activity	1.33×10^{-7}
	GO:0004930	G-protein coupled receptor activity	6.29×10^{-4}
	GO:0045236	CXCR chemokine receptor binding	1.44×10^{-3}
	GO:0033038	bitter taste receptor activity	0.0143

Table A.2: Gene ontology terms associated with lack of introgression in CEU

GO Catgory	GO ID	Description	FDR q-value
Process:	GO:0071493	cellular response to UV-B	1.5×10^{-3}
	GO:0030214	hyaluronan catabolic process	7.99 × 10−3
	GO:0010224	response to UV-B	0.0103
	GO:0045926	negative regulation of growth	0.0355
	GO:0060337	type I interferon signaling pathway	0.0314
	GO:0030212	hyaluronan metabolic process	0.0288
	GO:0033141	positive regulation of peptidyl-serine phosphorylation of STAT protein	0.038
	GO:0033139	regulation of peptidyl-serine phosphorylation of STAT protein	0.0424
	GO:0071482	cellular response to light stimulus	0.0452
	GO:0061099	negative regulation of protein tyrosine kinase activity	0.0409
Function:	GO:0004415	hyaluronon glucosaminidase activity	2.37×10^{-4}
	GO:0005132	type I interferon receptor binding	4.4×10^{-4}
	GO:0015929	hexosaminidase activity	1.2×10^{-3}
	GO:0033906	hyaluronoglucuronidase activity	1.14×10^{-3}
	GO:0031433	telethonin binding	0.0177

Table A.3: Gene ontology terms associated with enrichment of introgression in CHB+CHS

GO Catgory	GO ID	Description	FDR q-value
Process:	GO:0050911	detection of chemical stimulus involved in sensory perception of smell	3.32×10^{-9}
	GO:0050907	detection of chemical stimulus involved in sensory perception	1.13×10^{-7}
	GO:0009593	detection of chemical stimulus	1.14×10^{-7}
	GO:0050906	detection of stimulus involved in sensory perception	1.14×10^{-6}
	GO:0051606	detection of stimulus	2.7×10^{-5}
	GO:0007186	G-protein coupled receptor signaling pathway	2.65×10^{-3}
	GO:0006342	chromatin silencing	0.0399
Function:	GO:0005549	odorant binding	9.43×10^{-13}
	GO:0004984	olfactory receptor activity	4.97×10^{-10}
	GO:0001730	2'-5'-oligoadenylate synthetase activity	5.41×10^{-6}
	GO:0004930	G-protein coupled receptor activity	9.2×10^{-5}
	GO:0004950	chemokine receptor activity	0.0122
	GO:0001637	G-protein coupled chemoattractant receptor activity	0.0102
	GO:0015125	bile acid transmembrane transporter activity	0.0104
	GO:0070566	adenylyltransferase activity	0.033
	GO:0004715	non-membrane spanning protein tyrosine kinase activity	0.0421
Component:	GO:0000786	nucleosome	2.49×10^{-7}
	GO:0044815	DNA packaging complex	7.36×10^{-7}
	GO:0032993	protein-DNA complex	2.49×10^{-4}
	GO:0017101	aminoacyl-tRNA synthetase multienzyme complex	5.27×10^{-3}
	GO:0045095	keratin filament	0.0287

Table A.4: Gene ontology terms associated with enrichment of introgression in CEU

A.2 Simulation study

Here we present some results of the simulation study we performed to assess the accuracy of the method presented in Chapter 3.

ROC and Precision-Recall Curves from Simulated Data

Precision-recall curves for data simulated under different models, and analyzed using the same model as simulated and the "true" model are presented in Figures A.1, A.2, A.3, A.4, A.5, A.6.

Power as a function of tract length

We used the marginal posterior obtained from analyzing the data simulated under the true "model", using the "true" model in the analysis, categorizing introgressed fragments by their length. In Figure A.7, we plot the percentage of correctly called bases for tracts of a certain length:

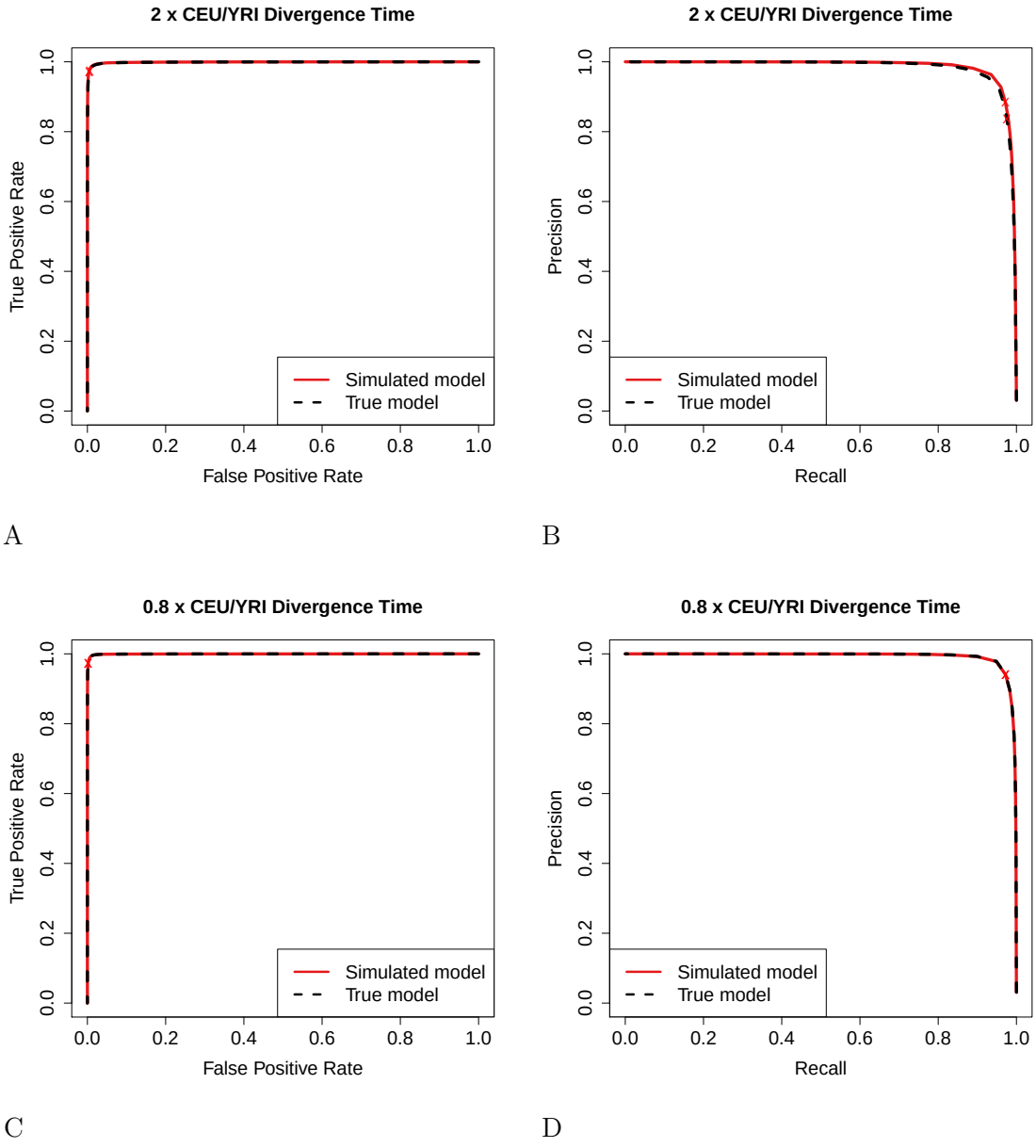


Figure A.1: Sensitivity of ROC and PR curves to CEU/YRI divergence time.

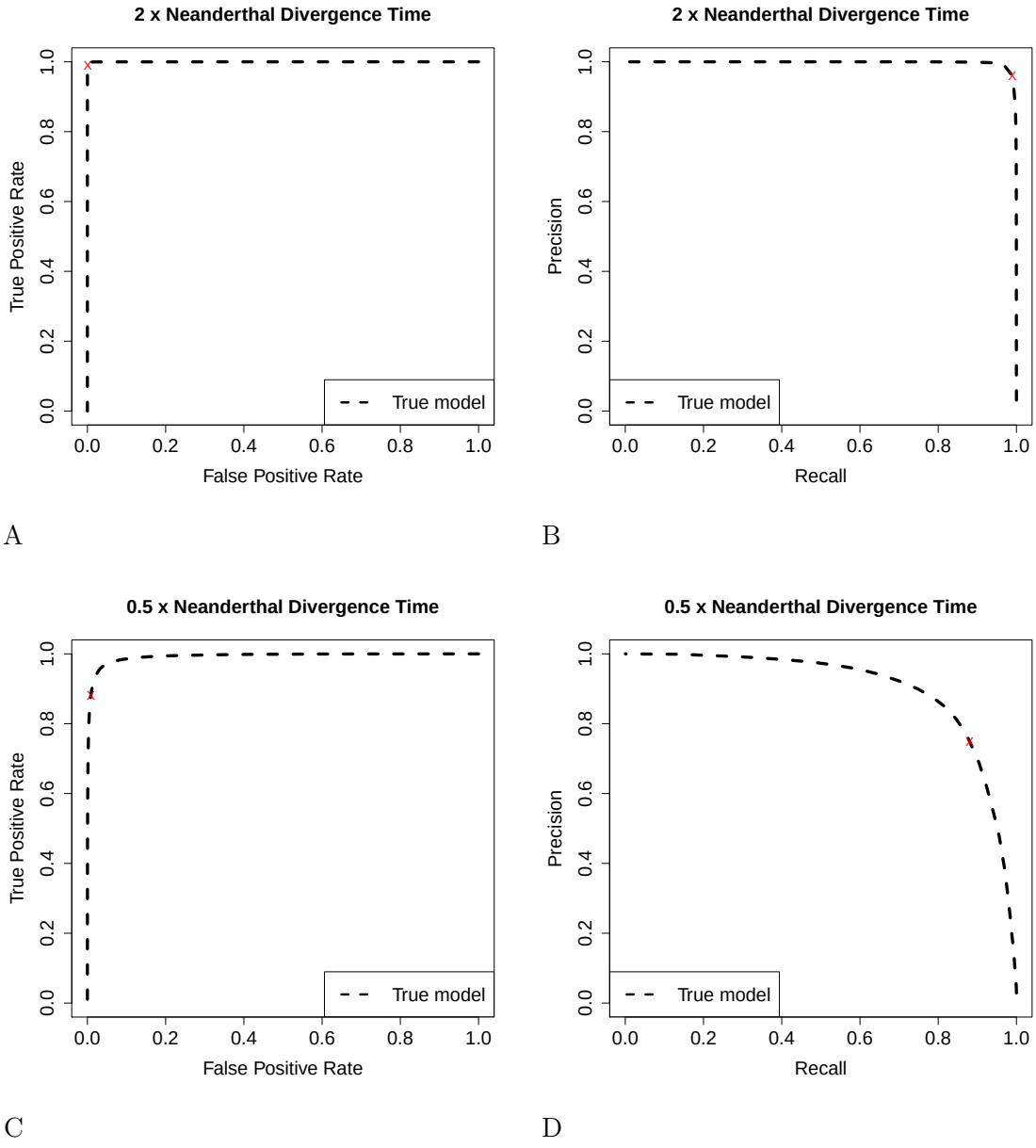


Figure A.2: Sensitivity of ROC and PR curves to Neanderthal divergence time.

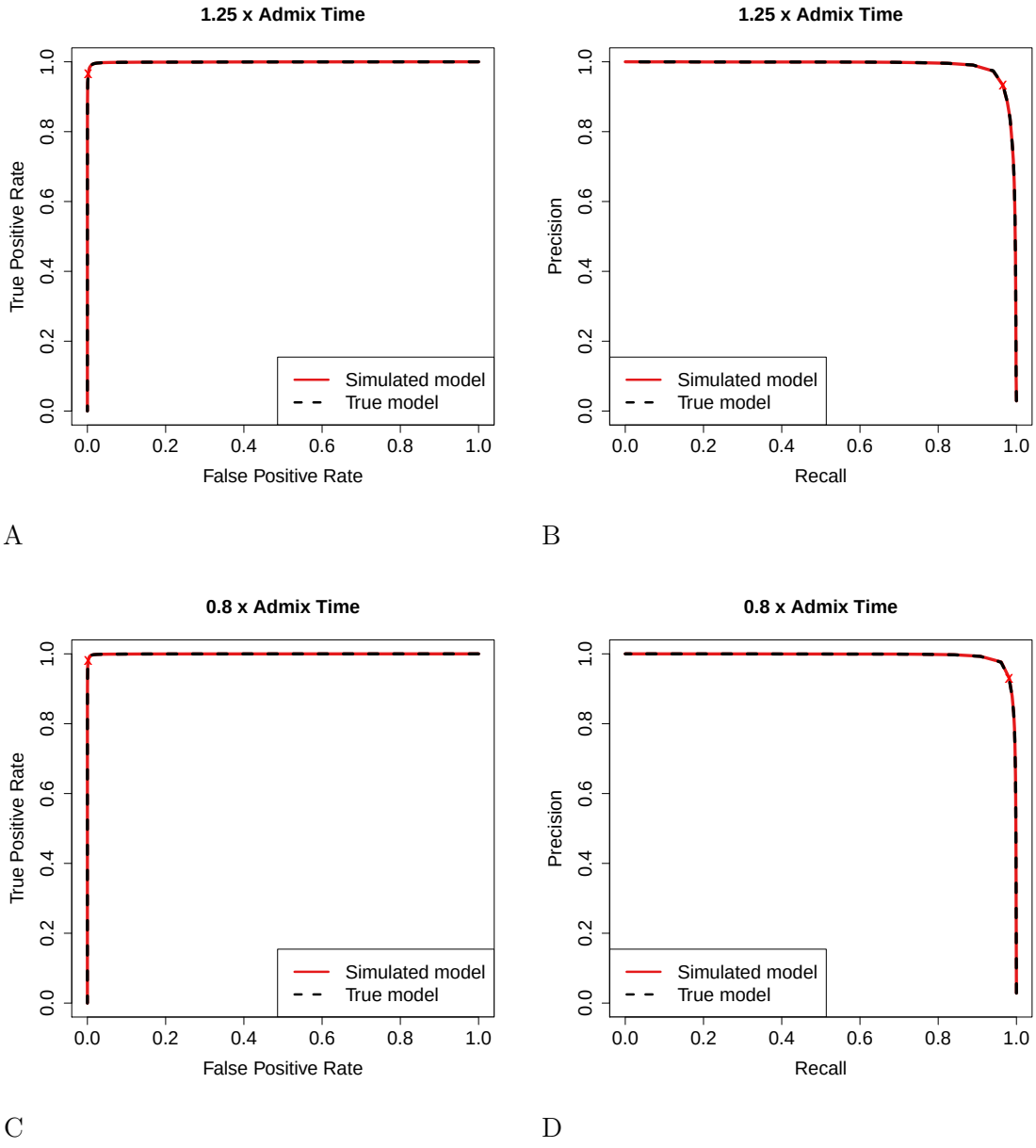
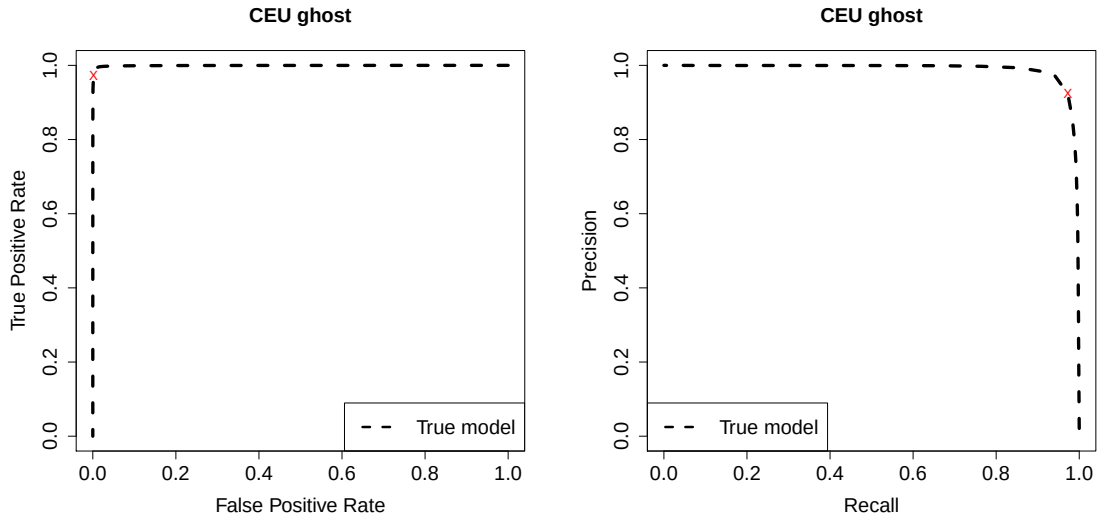
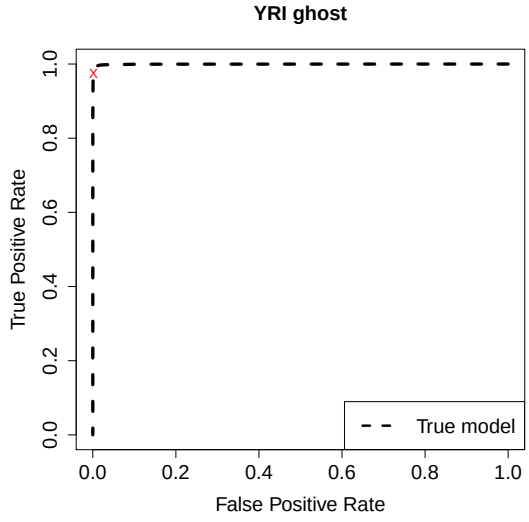


Figure A.3: Sensitivity of ROC and PR curves to the timing of admixture.

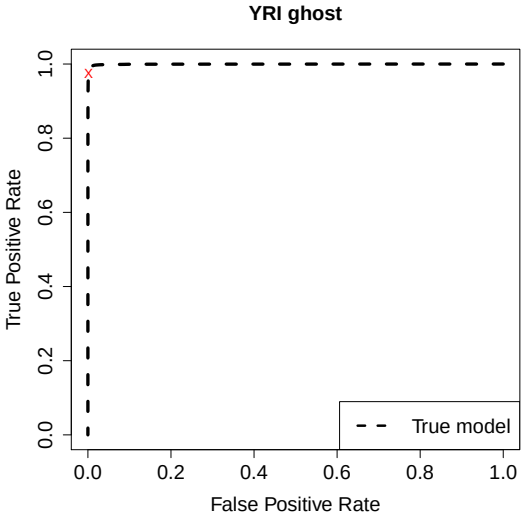


A

B

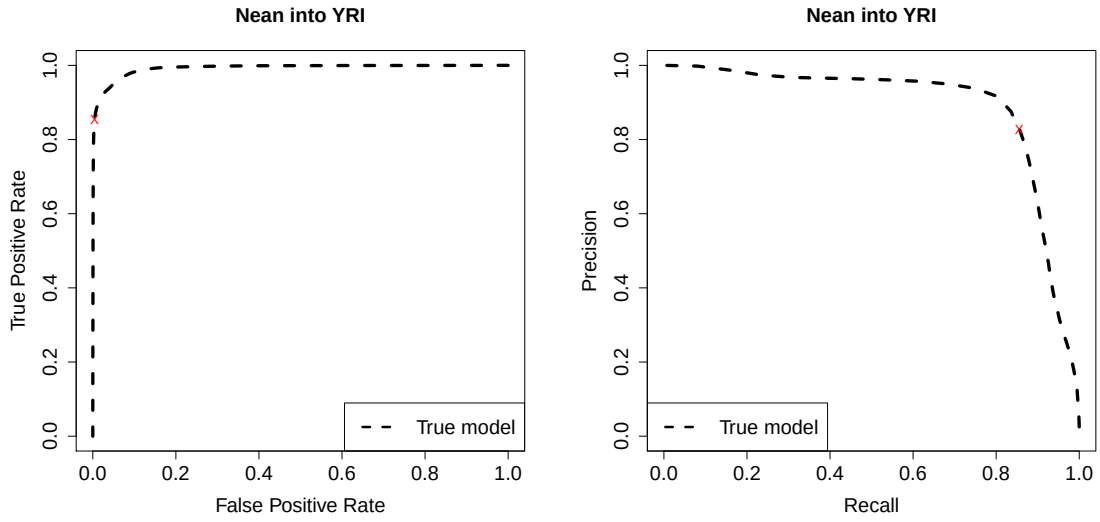


C



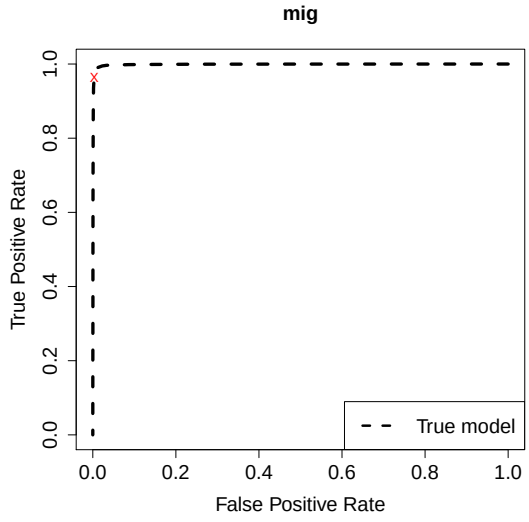
D

Figure A.4: Sensitivity of ROC and PR curves to migration from a ghost population into either CEU (a), (b) or YRI (c), (d).

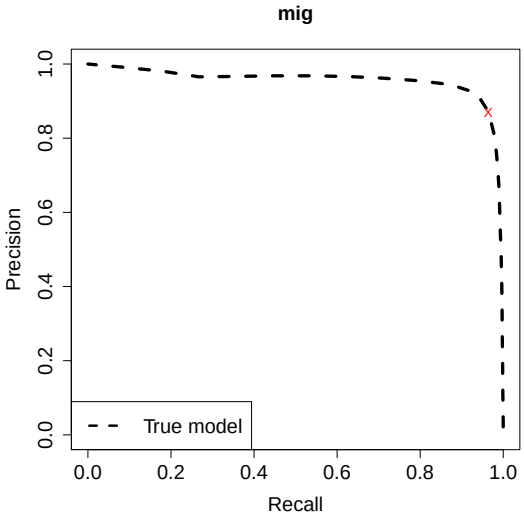


A

B

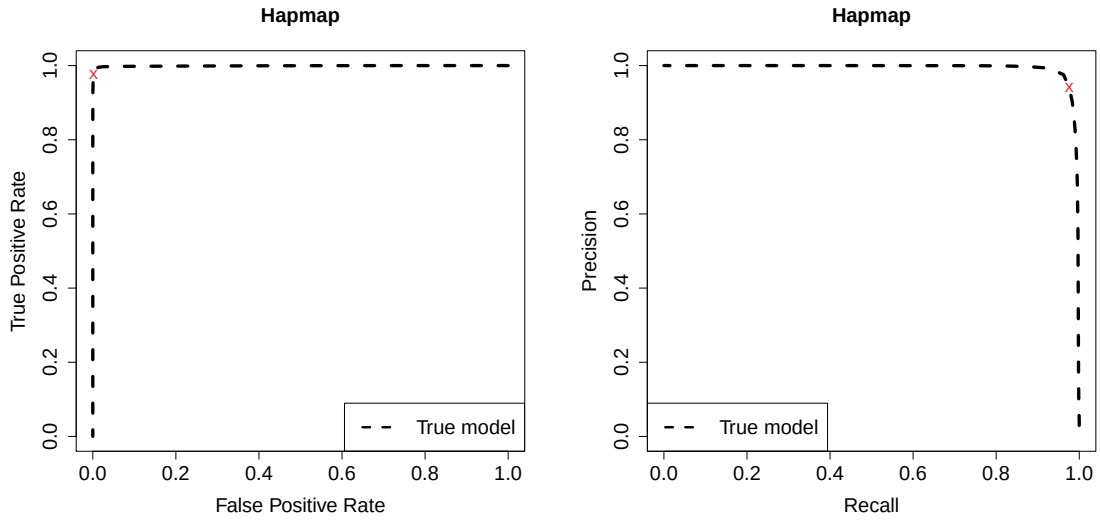


C



D

Figure A.5: Sensitivity of ROC and PR curves to introgression from Neanderthals into YRI (a), (b) or migration after the admixture between CEU and YRI (c), (d).



A B

Figure A.6: Sensitivity of ROC and PR curves to differences in the assumed constant recombination map to a realistic recombination taken from HapMap [9].

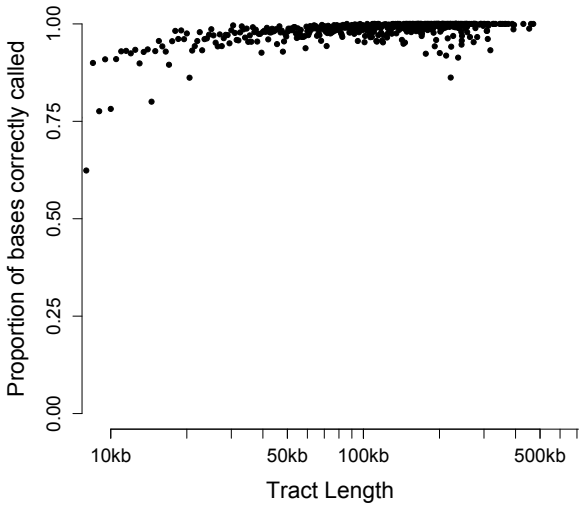


Figure A.7: Power as a function of tract length.

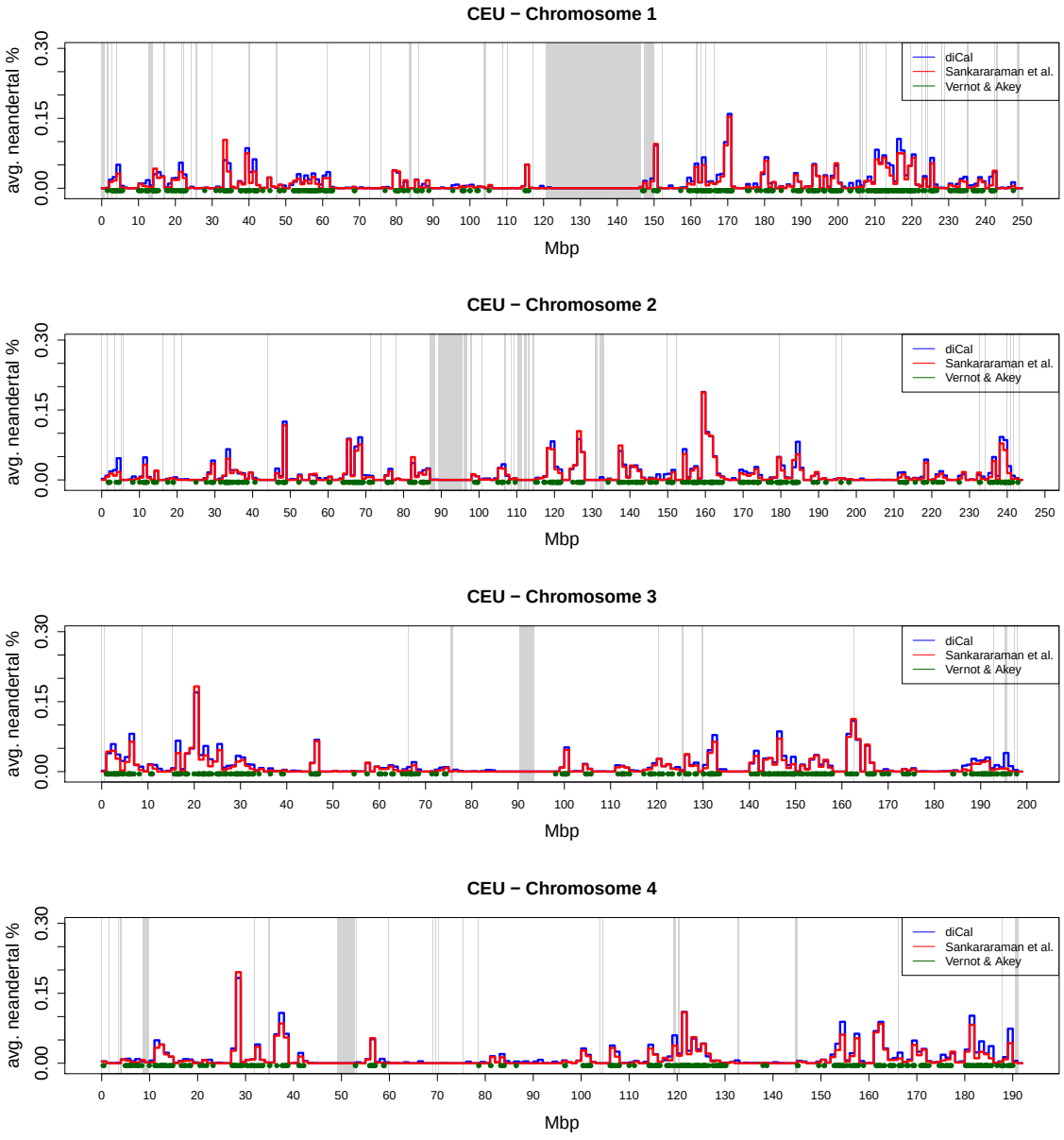
A.3 Fine-scale population average introgression

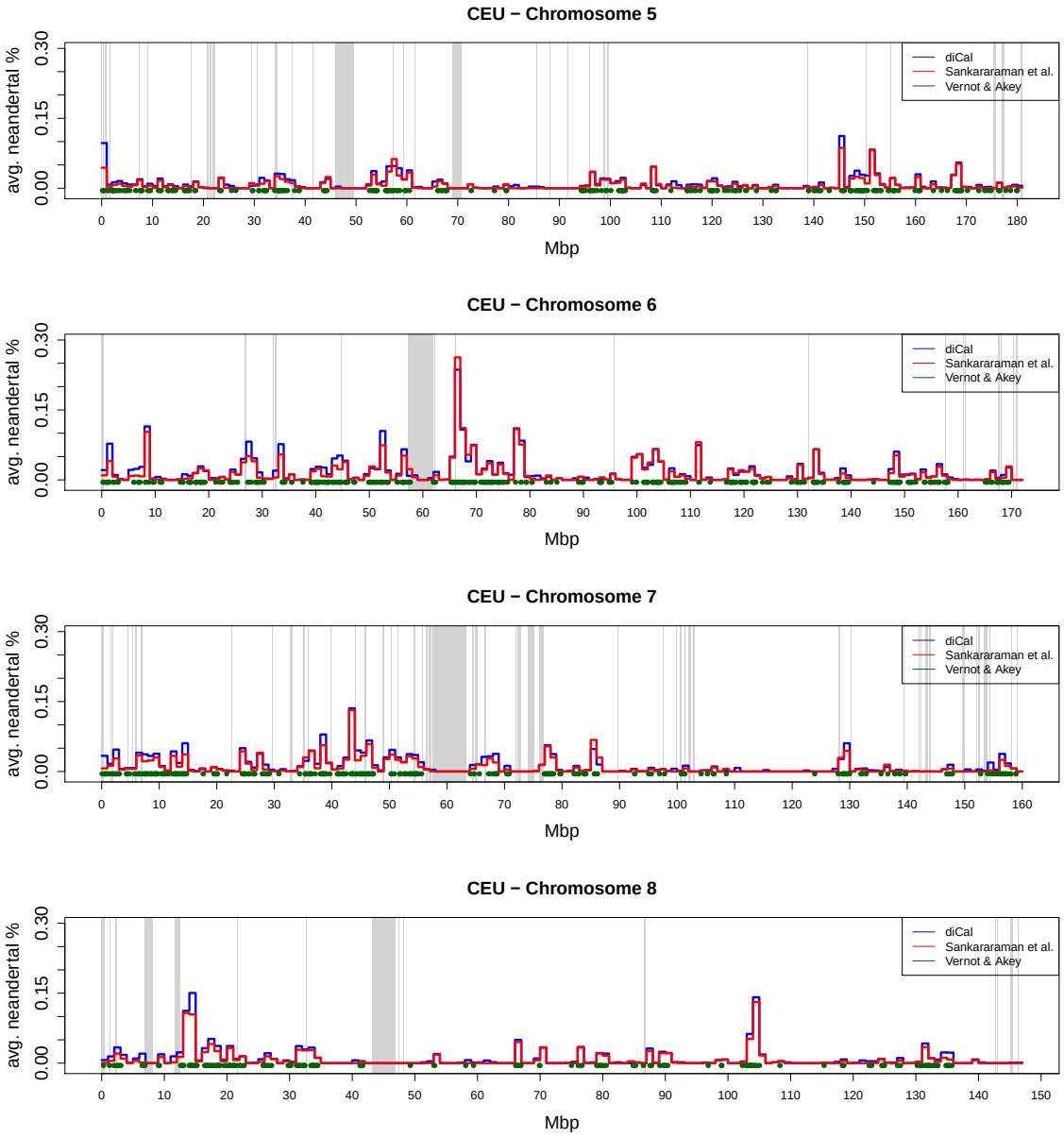
CEU

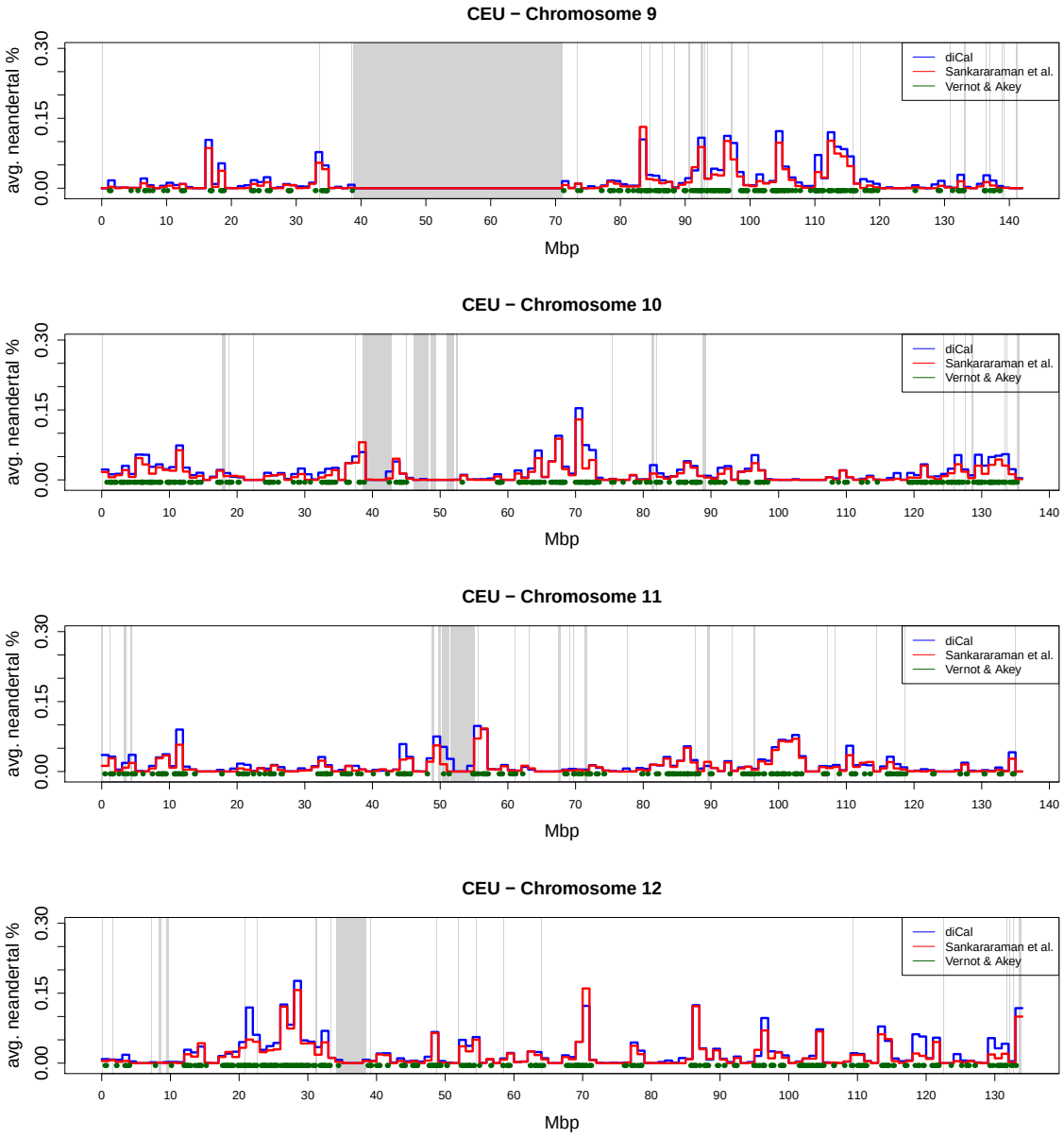
Here we present skyline plots of the amount of Neanderthal introgression in modern individuals inferred using the method presented in Chapter 3. Figure A.8 presents that amount of introgression in the CEU samples on the different chromosomes, averaged over all individuals in 1 Mbp windows. The results from *diCal* are indicated in blue, and the results from [6] indicated in red. The regions reported as introgressed by [60] are indicated in green. The gray bars denote the regions were no calls were made in the 1000 genomes dataset, which include the centromeres.

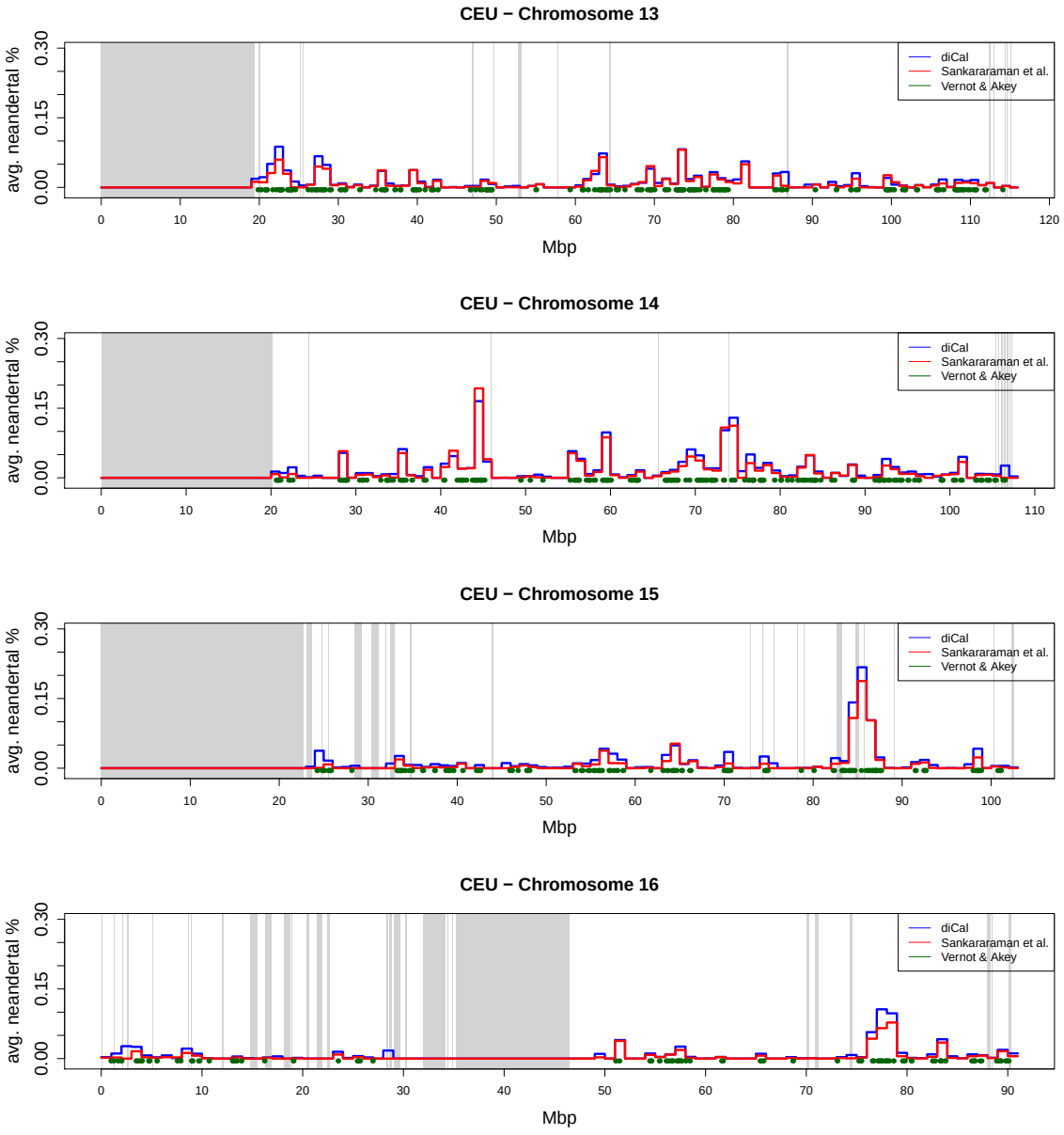
CHB+CHS

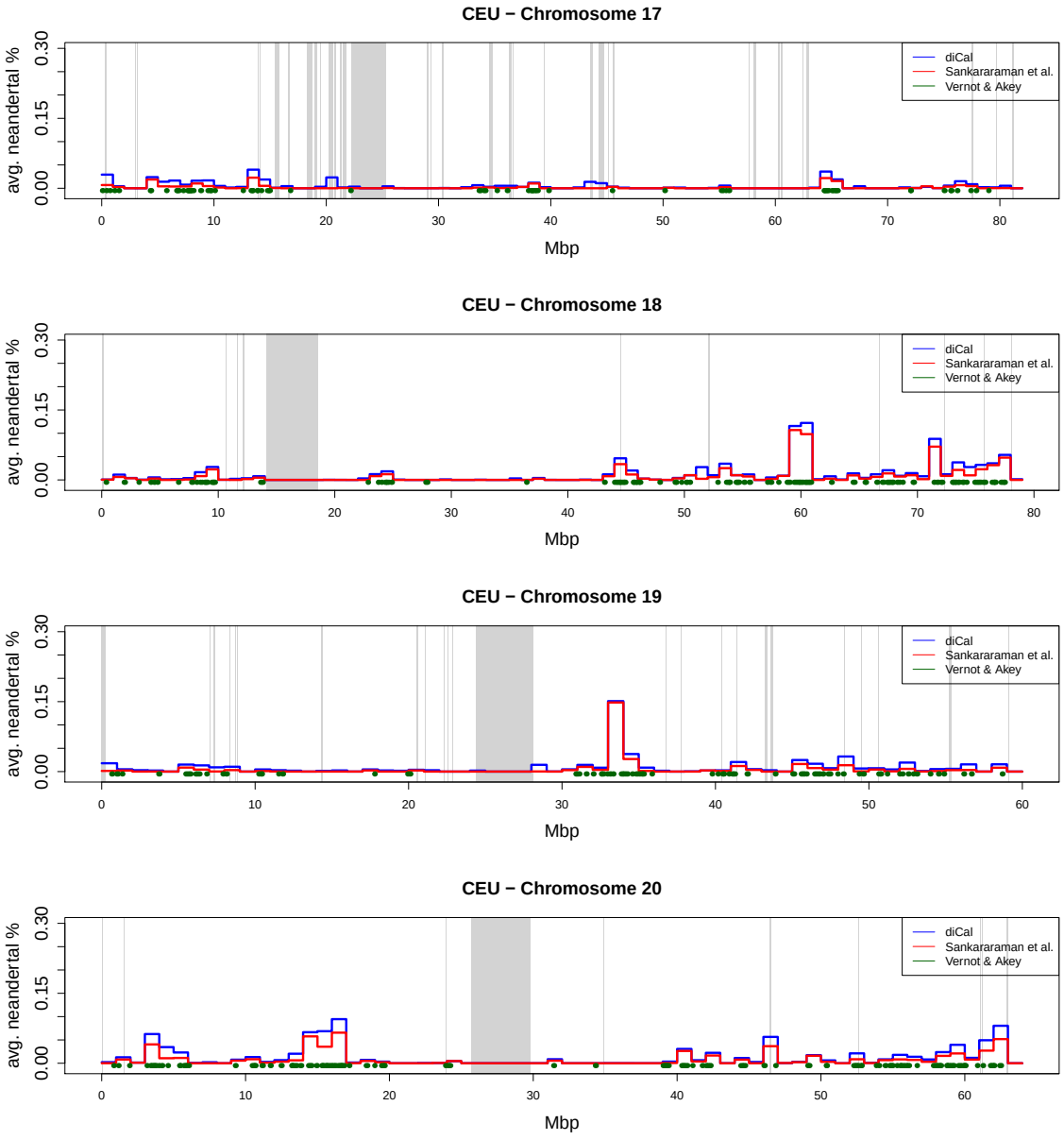
A skyline plot of the amount of Neanderthal introgression in the individuals of the CHB and CHS population on the different chromosomes, averaged over all individuals in 1 Mbp windows is presented in Figure A.9. The results from *diCal* are indicated in blue, and the results from [6] indicated in red. The regions reported as introgressed by [60] are indicated in green. The gray bars denote the regions were no calls were made in the 1000 genomes dataset, which include the centromeres.


CEU − Chromosome 1
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
Mbp
CEU − Chromosome 2
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
Mbp
CEU − Chromosome 3
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200
Mbp
CEU − Chromosome 4
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
Mbp


CEU − Chromosome 5
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 6
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 7
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 8
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Mbp
diCal
Sankararaman et al.
Vernot & Akey


CEU − Chromosome 9
CEU − Chromosome 10
CEU − Chromosome 11
CEU − Chromosome 12
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Mbp
diCal
Sankararaman et al.
Vernot & Akey


CEU − Chromosome 13
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
100
110
120
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 14
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
100
110
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 15
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
100
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 16
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
Mbp
diCal
Sankararaman et al.
Vernot & Akey


CEU − Chromosome 17
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 18
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 19
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CEU − Chromosome 20
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
Mbp
diCal
Sankararaman et al.
Vernot & Akey

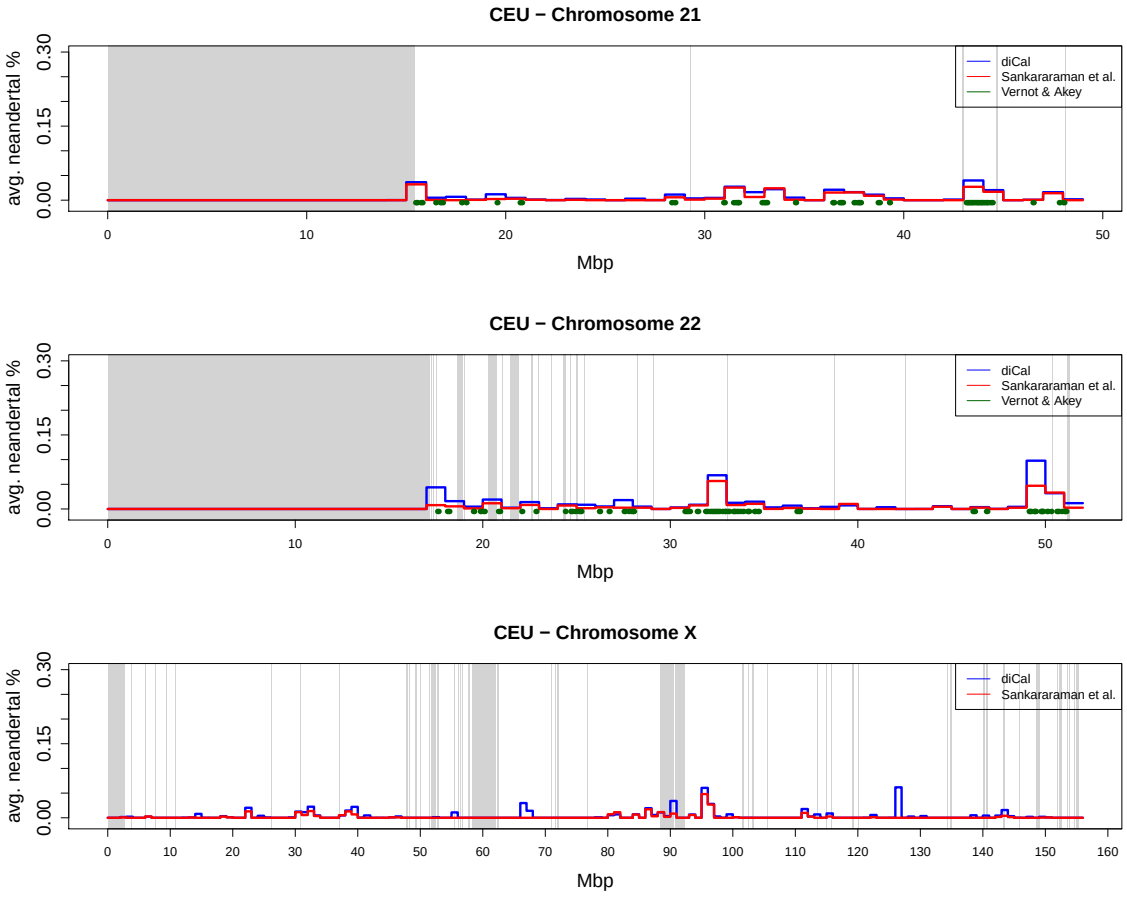
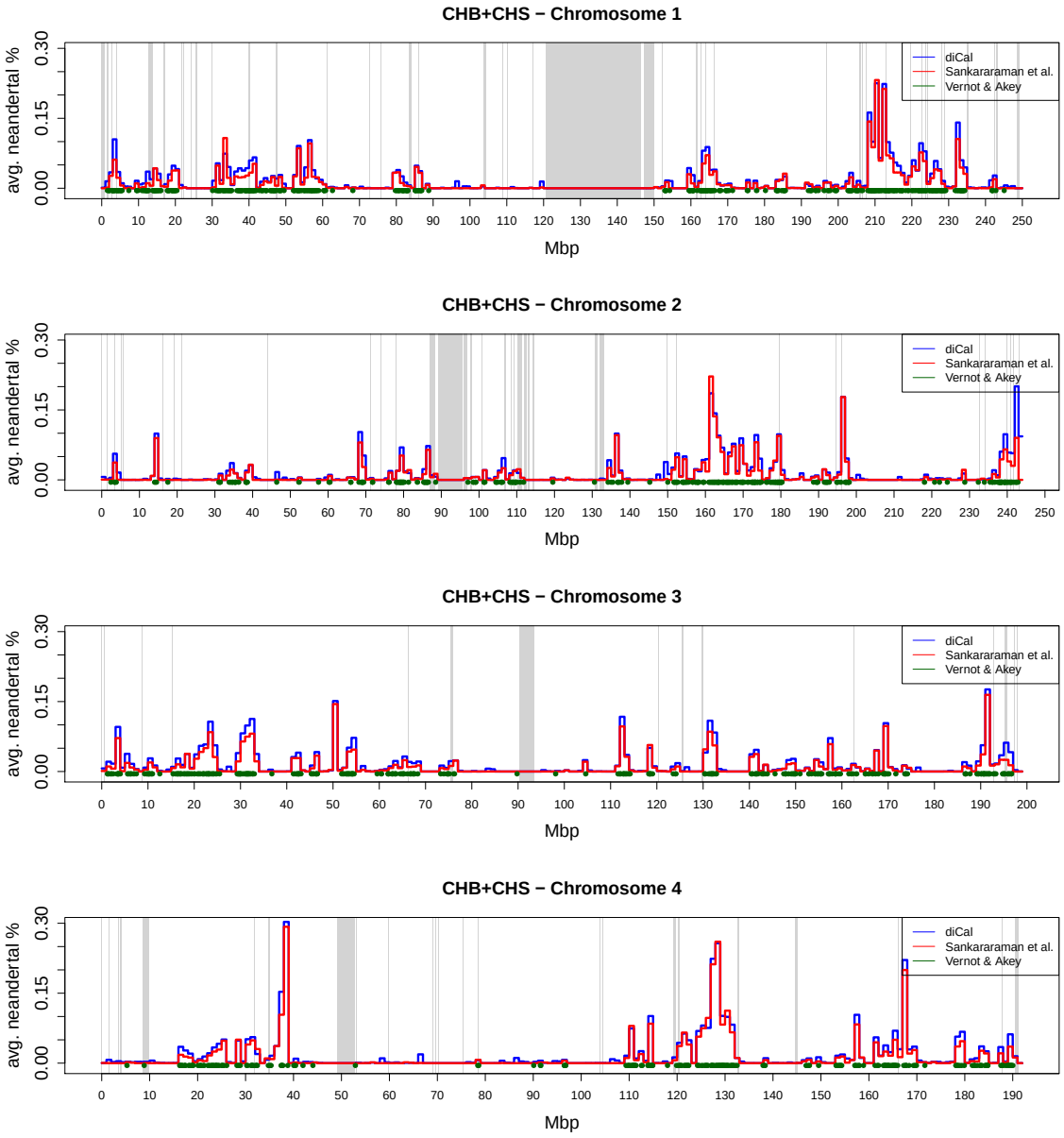
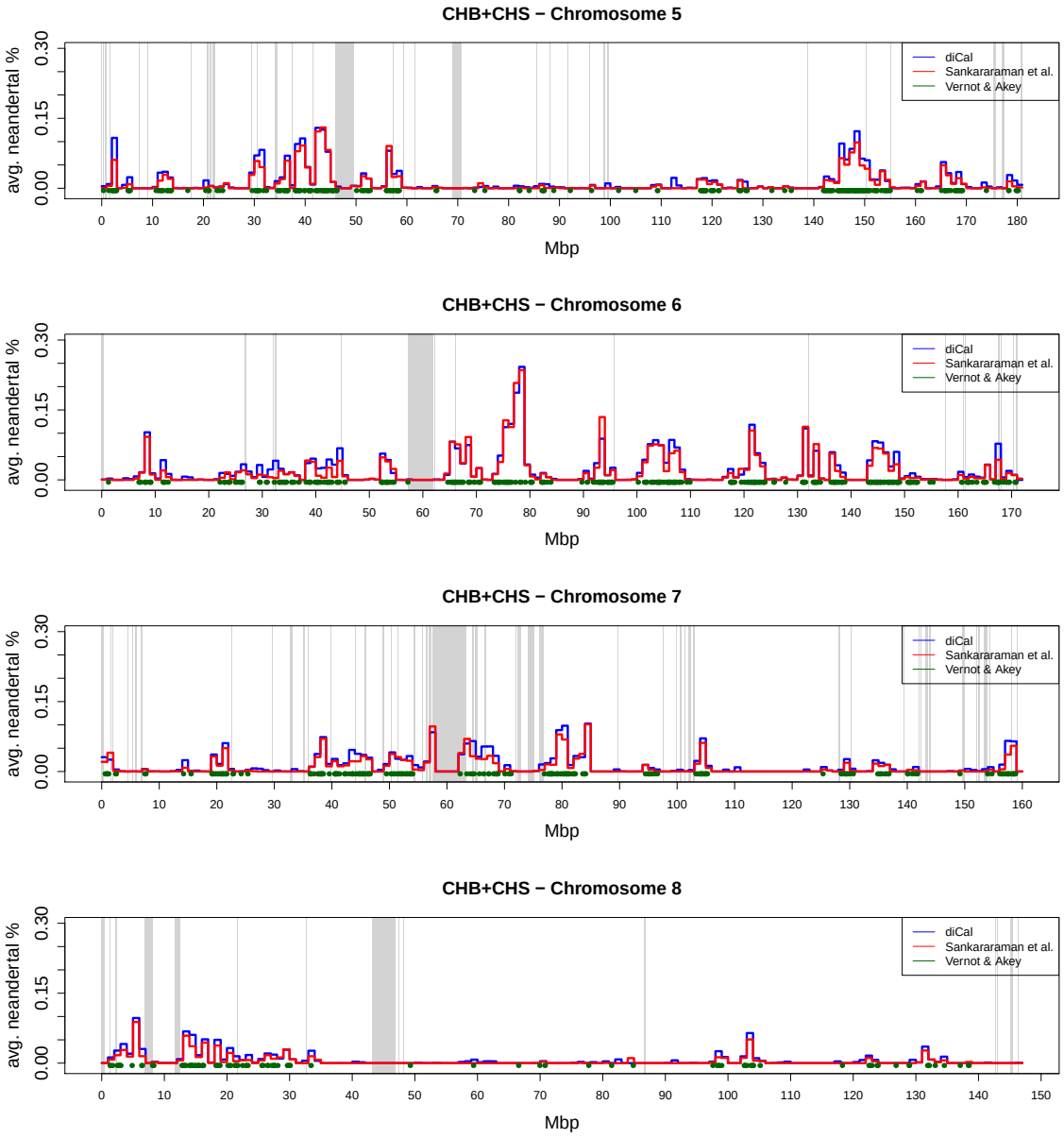
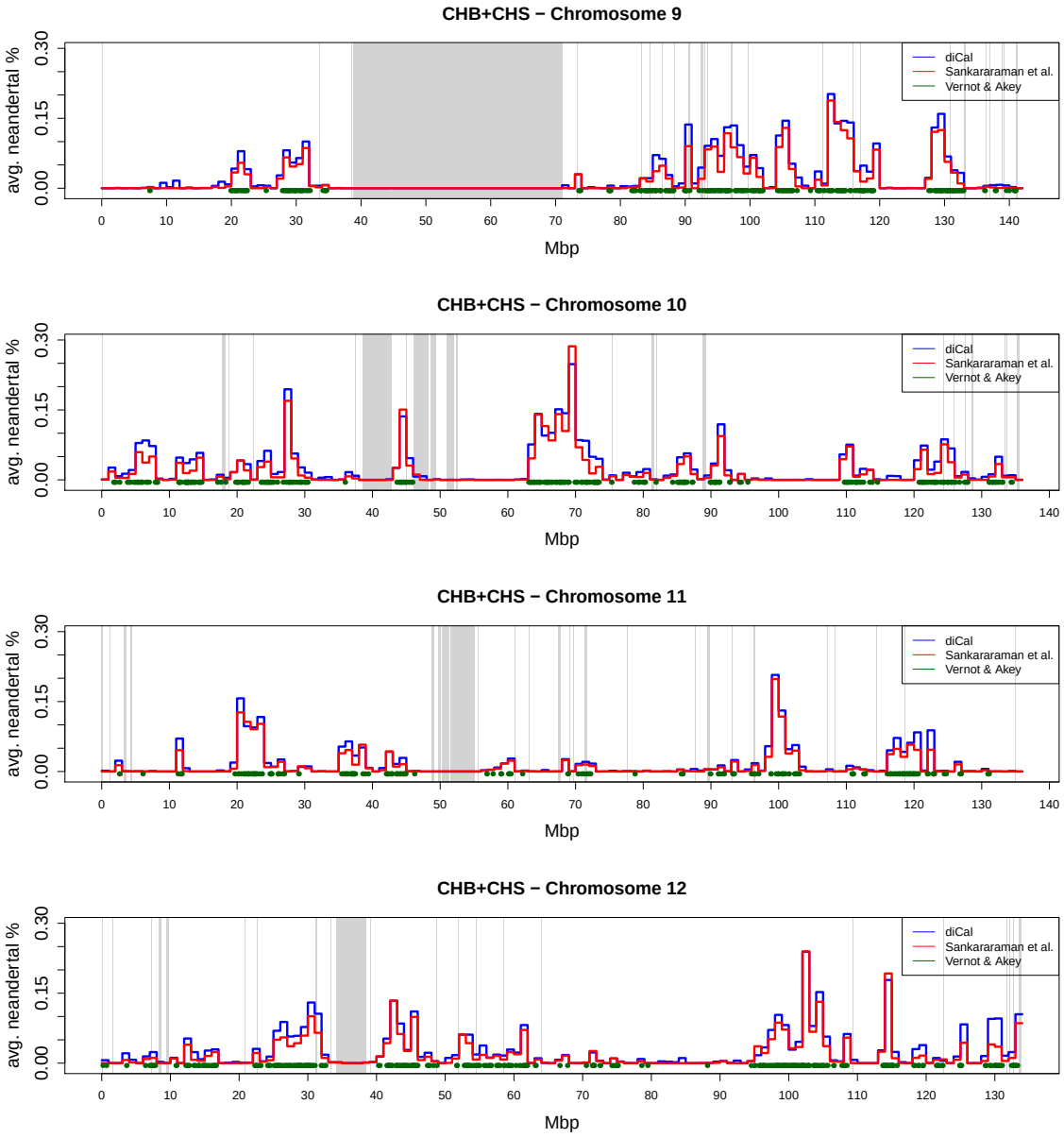
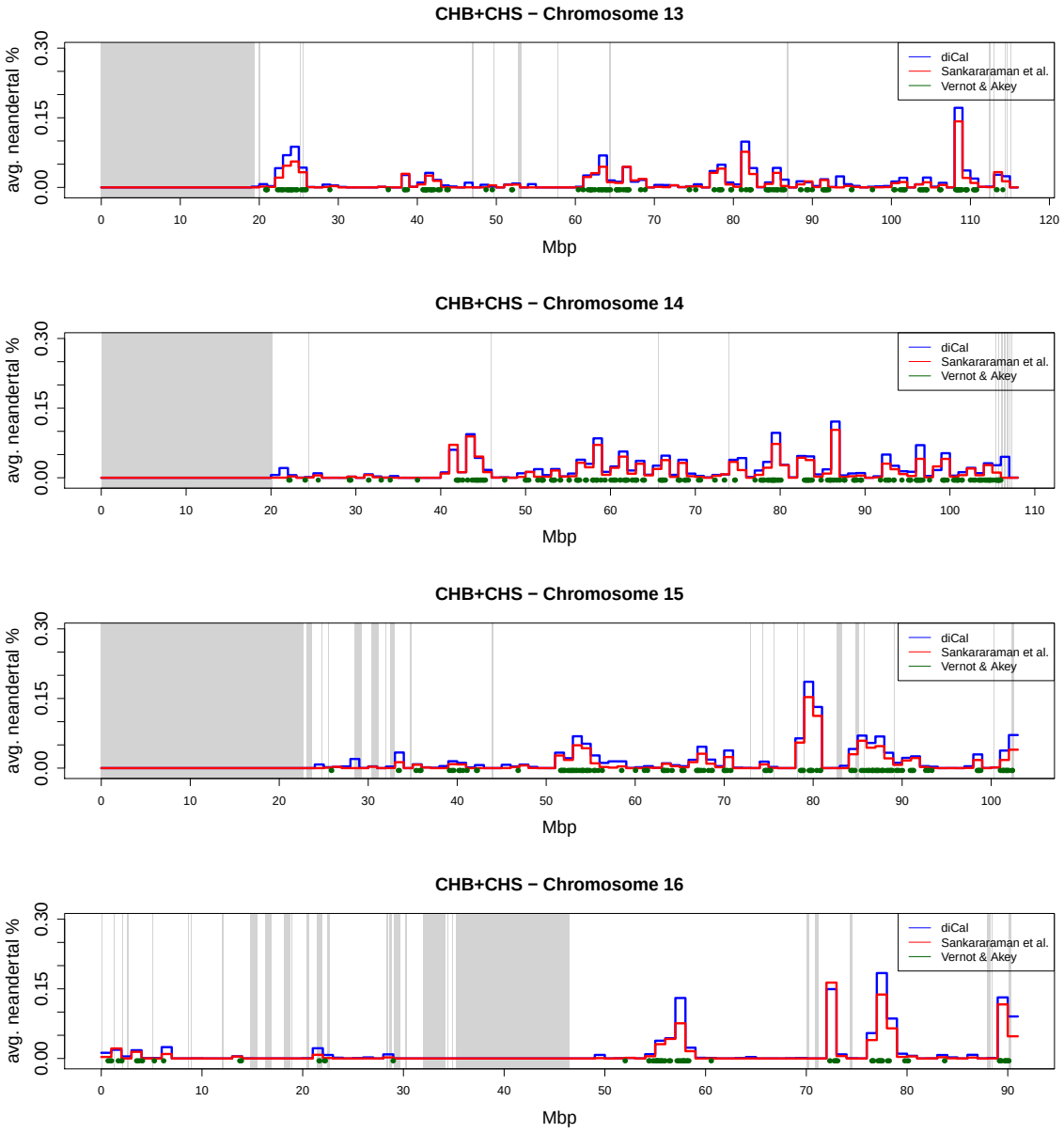


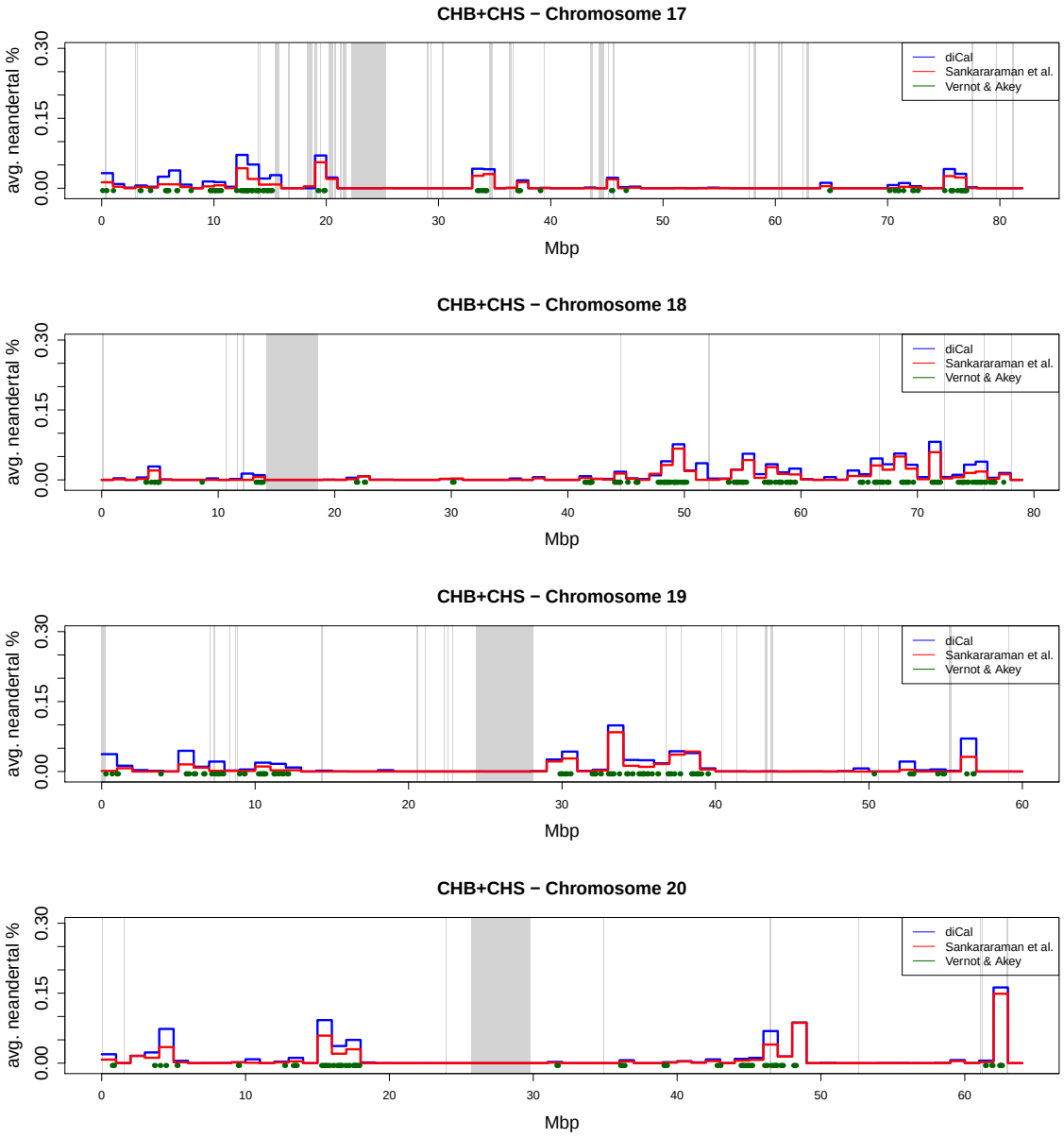
Figure A.8: Skyline plots of Neanderthal introgression in CEU.


CHB+CHS − Chromosome 1
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
Mbp
CHB+CHS − Chromosome 2
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250
Mbp
CHB+CHS − Chromosome 3
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200
Mbp
CHB+CHS − Chromosome 4
avg. neandertal %
0.00
0.15
0.30
diCal
Sankararaman et al.
Vernot & Akey
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190
Mbp


CHB+CHS − Chromosome 5
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 6
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 7
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 8
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Mbp
diCal
Sankararaman et al.
Vernot & Akey


CHB+CHS − Chromosome 9
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 10
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 11
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 12
avg. neandertal %
0.00
0.15
0.30
0 10 20 30 40 50 60 70 80 90 100 110 120 130
Mbp
diCal
Sankararaman et al.
Vernot & Akey


CHB+CHS − Chromosome 13
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
100
110
120
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 14
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
100
110
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 15
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
100
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 16
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
90
Mbp
diCal
Sankararaman et al.
Vernot & Akey


CHB+CHS − Chromosome 17
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 18
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
70
80
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 19
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
Mbp
diCal
Sankararaman et al.
Vernot & Akey
CHB+CHS − Chromosome 20
avg. neandertal %
0.00
0.15
0.30
0
10
20
30
40
50
60
Mbp
diCal
Sankararaman et al.
Vernot & Akey

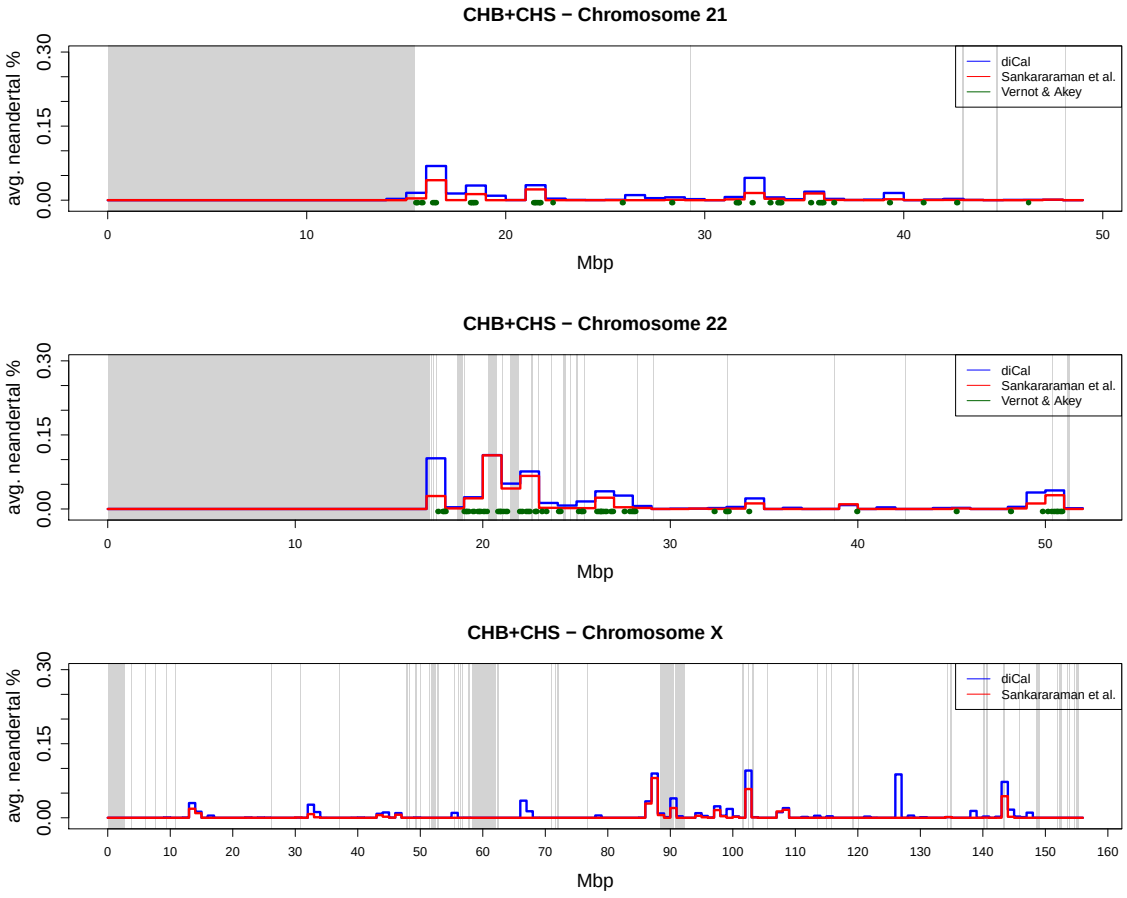


Figure A.9: Skyline plots of Neanderthal introgression in CHB + CHS.