Population Genetic Effects of Recent Admixture

By

Weiyi Mason Liang

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Integrative Biology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Rasmus Nielsen, Chair
Professor Montgomery Slatkin
Professor James Pitman

Fall 2014

Population Genetic Effects of Recent Admixture

Abstract

Population Genetic Effects of Recent Admixture

by

Weiyi Mason Liang

Doctor of Philosophy in Integrative Biology

University of California, Berkeley

Professor Rasmus Nielsen, Chair

Admixture has played an important role in shaping genetic diversity in many human populations. Quantifying these effects is important not only for answering historical questions, but also for detecting selection, mapping disease genes, and estimating recombination rates. Many existing methods for estimating admixture times use spatial information from the genomes of admixed individuals, such as the distribution of admixture tract lengths or the two-point covariance function of their local ancestries. I first discuss some theoretical results about the length distribution of admixture tracts. I use simulations to show that, for recent admixture events, no existing population genetic model approximates this length distribution well. I introduce a new model, based on dyadic intervals, which is accurate in this regime more mathematically tractable. I then show how the distribution of admixture proportions within a population, as estimated by programs such as STRUCTURE, gives information about the population's admixture history and relate the moments of this distribution to the theory of multi-locus linkage disequilibria. Finally, I show how measures of three-locus linkage disequilbria can be used to improve on the resolution of existing admixture history inference methods.

?

To my parents.

Contents

List of Figures

List of Tables

Acknowledgments

In addition to my parents, many people have helped me on this long journey. I'd like to thank the members of the Nielsen and Slatkin labs, including Mel Yang, Kelley Harris, Amy Ko, Fernando Racimo, Ben Peter, Emilia Huerta-Sanchez, and Anna Ferrer-Admetlla, for putting up with me and giving me advice over the past five and a half years. I'd especially like to thank Wayne Lee, for his encouragement, Joshua Schraiber, for his advice about the research and LaTeX template, and Alan Malek, for letting me use his apartment in Berkeley.

This dissertation was typeset using the ucastrothesis LaTeX template.

Chapter 1

Introduction

Over the course of human history, trade, conquest, slavery, and migration have all led to gene flow between previously isolated source populations and the creation of admixed populations, such as African Americans (Parra et al. 1998), Indians (Moorjani et al. 2013), or Rapanui (Moreno-Mayar et al. 2014). Understanding these admixture histories is important, not only for answering historical or anthropological questions, but also from a biological perspective, because of the population genetic effects of admixture. Gene flow from source populations into an admixed population is expected to cause genome-wide correlations which would otherwise not be present. Over the course of generations, this correlation is then broken down through recombination and drift in the admixture population. Accounting for these correlations, and their decay as a function of time, is a crucial step in answering many biological questions, e.g. mapping disease gene mapping, estimating recombination rates, or inferring local ancestries.

The population genetic effects of admixture are closely related to the theory of junctions, which were first studied by (Fisher 1949). Junctions for an individual can be defined with respect to a collection of ancestors of that individual, and are positions in that individual's chromosome which mark transitions in inheritance. For example, a junction may mark the base pair where an individual's chromosome transitions from being inherited by one grandparent to being inherited from another. Although junctions are passed down in a population in the same manner as genetic markers, junctions are not physical, and their existence can only be inferred. In analyzing admixture, we are interested in transitions in the local ancestry i.e. the junctions with respect to source populations instead of collections of ancestors. The junctions are positions at which the chromosome transitions from being inherited from one source population to being inherited from another. For example, a junction in an African American individual may demarcate a section of a chromosome that is inherited from an African ancestor from a section that is inherited from a European ancestor.

A frequently used model of admixture is a one-pulse model (Gravel 2012) and (Moorjani et al. 2011), in which, after the founding generation, there is no additional gene flow from any of the source populations into the admixed population. In the second chapter, I analyze the distribution of admixture tract lengths that arises from this model. Admixture tracts are the

contiguous sections of genome descended from a single source population, i.e. the segments between consecutive junctions. This length distribution is commonly approximated by an exponential distribution. I show that the accuracy of this approximation depends on several factors, including the age of the admixture event and the effective admixed population size. For recent admixture events, no existing model is accurate, so I introduce a new model, based on dyadic intervals, which has the correct admixture tract length distribution for recent admixture events.

A commonly used technique in admixture analyses is estimating the admixture proportions of samples via programs such as STRUCTURE or ADMIXTURE. Admixture proportions are the proportions of admixed individuals' chromosomes which trace their ancestry back to each source population. This can be thought of as an integral of the local ancestry over each individual's entire genome. In the third chapter, I show that the distribution of these admixture proportions gives information about the population's admixture history. The moments of this random distribution are related to the n-point correlation functions of the local ancestry. I then show how to compute the expectations of these correlation functions in terms of the population's admixture history and additional population genetic parameters.

Existing inference methods for admixture histories are generally limited to a one-pulse model, but the complexities of many populations' admixture histories cannot be adequately captured by such a coarse model. In the final chapter, I show how existing methods for estimating admixture histories can be improved by using a statistic based three-locus linkage disequalibrium. These existing methods, based on two-locus linkage disequalibrium, are limited to estimating the time for the most recent pulse of migration. I relate the linkage disequalibrium created by admixture to the two and three point covariance functions of the local ancestry, which were computed in the preceding chaper. With this, we can fit more complex admixture histories to the observed statistics. I show that the addition of a third locus improves the resolution of the method, allowing it to estimate the timing of multiple pulses of migration.

Chapter 2

Admixture Tracts Lengths

2.1 Introduction

There has been interest in analyzing population genomic data by using methods that partition an admixed individual's genome into blocks originating from different ancestral populations. An early version of the popular program Structure (Falush et al. 2003) accomplished this with a hidden Markov model (HMM), indexed along the genome, with hidden states corresponding to the ancestral population each position was inherited from. The contiguous blocks of the genome inherited from a population are called "admixture/migrant tracts/segments", depending on the context. For consistency, we will use the term "admixture tract". Admixture tracts are unobservable, and their existence can only be inferred from genomic data. The process of doing so is called "admixture deconvolution" or "ancestry painting", and has been used in a number of different contexts, such as in admixture mapping for identifying human disease associated genes (Hoggart et al. 2003; Reich et al. 2005), population genetic inferences aimed at understanding human ancestry (Bryc et al. 2010; Henn et al. 2012), or identifying regions affected by natural selection (Tang et al. 2007).

The technique of using HMMs to partition an individual's genome into admixture tracts has been used in subsequent methods. Hoggart et al. (2003) and Smith et al. (2004) used HMMs for inferring admixture tracts with the purpose of admixture mapping and controlling for population stratification, similar to the method of Falush et al. (2003). More recent publications have focused on admixture deconvolution for more general population genetic purposes, such as Tang et al. (2006) and Sundquist et al. (2008).

In HapMix (Price et al. 2009), the HMM model of Li & Stephens (2003) for modeling linkage disequilibrium is extended to include admixture between two populations. HapMix uses a genotype-based state space and so does not require phased data.

LAMP (Sankararaman et al. 2008; Paşaniuc et al. 2009; Baran et al. 2012) is similar to HapMix, in that it also can be considered an extension of the Li and Stephens model. However, the size of its state space does not depend on the number of reference haplotypes, which allows it to run faster than HapMix.

PCAdmix (Bryc et al. 2010; Brisbin et al. 2012; Henn et al. 2012) also uses an HMM to

identify admixture tracts, but replaces observed data with admixture scores inferred from principle component analyses (PCA). As in the case of LAMP, it is applicable to multiple populations. Brisbin et al. (2012) argue that the method performs better than LAMP in simulations and has performance comparable to that of HapMix, which is limited to two populations.

There are also methods for estimating population genetic parameters of admixture events from genomic data without first inferring admixture tracts, such as ROLLOFF (Moorjani et al. 2011). Other more general methods for estimating population genetic parameters, such as $\partial a \partial i$ (Gutenkunst et al. 2009), can also be used to estimate time and the strength of admixture events. Finally, there are a many pre-genomic methods for analyzing divergence and gene-flow exemplified by the IM methods developed in (Hey & Nielsen 2004; Hey 2010). However, these methods do not directly use the information contained in the distribution of admixture tract lengths.

As a result of these efforts, there has been considerable interest in the relationship between admixture tract lengths and the time of admixture (T) and admixture fraction (m), to be defined mathematically later. Pool & Nielsen (2009) derived the admixture tract length distribution under the assumptions that inbreeding is not significant and that tracts are so rare that they are unlikely to recombine with each other. Gravel (2012) relaxed this second assumption to model tracts descended from multiple migrant ancestors, but under simplified model of reproduction called the Markovian Wright-Fisher (MWF).

The methods for ancestry deconvolution discussed above use an HMM, assuming that the spacing between recombination events is independent and exponentially distributed, and that ancestries of these recombination segments are independent. This is equivalent to assuming that admixture tracts have lengths which are independent and exponentially distributed. Population genetic models which are designed to be Markov along the genome, such as the MWF, sequentially Markov coalescent (SMC) (McVean & Cardin 2005), or SMC' (Marjoram & Wall 2006) models generate admixture tracts with these properties. Under the Wright-Fisher (WF) model with recombination, which is not Markov along the genome, we show that admixture tracts lengths do not have an exponential distribution, and furthermore, that these lengths can be highly correlated. When T is small, these properties are a result of inheritance from a small, fixed sample pedigree, and when T is large, they are a result of inbreeding (in the sense of identity by descent due to genetic drift, as opposed to non-random mating). This former cause was first discussed by Wakeley et al. (2012) in examining the convergence of the ancestral recombination graph (Hudson 1983; Griffiths & Marjoram 1996) to the WF genealogical process. Because of this integration over pedigrees, the ancestral recombination graph diverges from the WF model when T is small, and, like the Markov population genetic models, generates independent, exponential tract lengths.

Parallel to the literature on inference methods for admixture deconvolution, there is a well-developed literature on the segregation of tracts in pedigrees. This starts with Fisher's theory of junctions (Fisher 1949). A junction is defined with respect to an ancestral population, and is a point in the chromosome where, due to a crossover, the segments to the left and right trace their descent back to different members of the ancestral population. The

distribution of the distances between junctions is of prime interest in this body of theory and is closely related to the distribution of admixture tract lengths. Fisher (1949) was interested in determining the expected number of junctions under different models of inbreeding. Stam (1980) extended Fisher's original results results by considering a randomly breeding population of constant size, and derived a number of different results under the assumption of independent and exponentially distributed tract lengths. Many studies have subsequently focused on the amount of genetic material passed from an individual to its descendants, given a known pedigree. Donnelly (1983) showed that the probability that an individual contributes no genes to a descendant T generations in the future is approximately $\exp(-TR/2^T)$, where R is the recombination map length. Barton & Bengtsson (1986) looked at the inheritance of blocks of loci under selection in hybridizing populations. Other studies have subsequently studied properties of the distribution of junctions and the distances between between junctions, for fixed pedigrees including (Guo 1994; Bickeböller & Thompson 1996a,b; Stefanov 2000; Ball & Stefanov 2005; Cannings 2003; Dimitropoulou & Cannings 2003; Walters & Cannings 2005; Rodolphe et al. 2008).

Baird et al. (2003) also consider the distribution of surviving tracts among the descendants of an individual. They model the number of descendants as a branching process and the lengths of inherited material carried by all descendants as a branching random walk. Assuming complete cross-over interference (i.e., at most one recombination event per chromosome), they derive the generating function for these lengths as a function of T and the map length. They also derive expressions for the mean number of tracts of a certain length under both the complete cross-over interference model and a Poisson process of recombination. Baird et al. (2003) notice that their results can be used to understand the process of genetic fragments between introgressed species, similar to the admixture problem considered here. In particular, they note that the standard deviations of both tract lengths and number of tracts are comparable to their means, indicating a high degree of variability. These results have been extended in other applications, for example to derive the distribution of reproductive values (Barton & Etheridge 2011).

Chapman & Thompson (2002) derive general expressions for the mean and variance of the number of junctions. Their results can be applied under different demographic models because they show that these two moments depend only on the recombination map length and the one and two-locus probabilities of identity-by-descent.

Beyond the fact that we focus on the effect on an admixed population, these approaches differ from our work in two ways. First, we consider the backwards-in-time process of the ancestry of a sample, instead of considering the forward-time process describing the descendants of an individual. We also by consider the merger of multiple fragments inherited from the a group of individuals (migrants), instead of the contributions from just one. The effect of such mergers is particularly important when the number of migrants is large.

As no models other than the full WF model are available for accurate analyses of tract lengths for recent admixture times, we present a new model of genealogical structure that can be used to analyze and approximate tract lengths distributions, and short-term pedigree based-processes more generally. This model assumes the sample has a full pedigree, and

represents the genealogical history of a sample in terms of dyadic intervals. It is accurate for time scales and population sizes in which pedigree structure is important but inbreeding is not.

2.2 Models

For simplicty, we consider a simple admixture scenario in which, T generations ago, two source populations contributed to form a third, admixed, population. Founders of this admixed population come from the "migrant" population with probability m and from the "non-migrant" population with probability $1 - m$. Note that the labels on the two source populations are arbitrary.

Each of the population-genetic models analyzed in this chapter model the reproduction and recombination in this monecious population of $2N$ chromosomes subsequent to the admixture event. We assume that recombination events follow a Poisson process with rate 1 crossover/Morgan. This assumption of no crossover interference is not biologically accurate, but it is mathematically tractable. We will later argue that this assumption is conservative with respect to the major conclusions of this chapter and show how our results can be extended to incorporate some models of interference.

Haploid Wright-Fisher with Recombination

This is the standard haploid version of the WF model with recombination considered by Gravel (2012), Wakeley et al. (2012), and others. Each chromosome is produced by recombining two parents from the previous generation, chosen independently and uniformly at random. We consider this to be the more appropriate model for understanding tract lengths distributions and compare the following models to it.

Markovian Wright-Fisher

Gravel (2012) introduced this mathematically tractable approximation of the diploid WF model. It assumes that chromosomes are formed from the recombination of *all* $2N$ chromosomes from the previous generation, instead of just two. At each recombination point the offspring copies from one of the $2N$ chromosomes from previous generation, uniformly at random. Additionally, it assumes that $2N$ is large, so that each crossing-over results in a new parent contributing genetic material. As its name implies, the MWF model is a Markov process along the genome.

Coalescent with Recombination

In the coalescent limit ($2N \to \infty$ with time measured in units $2N$ generations and recombination distance in units of crossovers/$4N$), Griffiths & Marjoram (1996) showed that the genealogical process of a sample from the haploid WF model converges in distribution

to the ancestral recombination graph (ARG), which can be constructed as a Markov process going backwards in time. Wiuf & Hein (1999) presented a sequential construction of the ARG along the genome. This sequential process is not Markov. Instead, the conditional distribution of a marginal trees depends on all the trees that have appeared to the left of it. The case of admixture tracts is slightly different than other uses of the coalescent, because here we start with one lineage and stop the process at the fixed time, $T/2N$, instead of the more common case, where we start with more than one lineage and stop the process when only one lineage is left.

Sequentially Markov Coalescent

McVean & Cardin (2005) developed an approximation of the coalescent in which the sequence of marginal trees form a Markov process along the sequence. In the sequentially Markov coalescent (SMC), the only allowed coalescence events are for lineages with overlapping ancestral material. The model is otherwise identical to the coalescent.

Majoram and Wall's SMC'

Marjoram & Wall (2006) presented a related model (SMC') which loosens the restrictions of the SMC while retaining its Markov property. In addition to the coalescence events allowed in the SMC, the SMC' further allows coalescence events for lineages with abutting ancestral material. This extra possibility allows for back-coalescences in the ancestral recombination graph, which produces a significant improvement for this model's predictive powers when these events are likely.

Perfect Binary Tree Model

As we will argue in the Results section, none of the four previous models approximate the tract length distribution well when T is small relative to $2N$. We therefore introduce the perfect binary tree model (PBT), so named because it assumes that sample has 2^T distinct great^{T-2} grandparents, i.e., that the pedigree of the sample, up to generation T, is a perfect binary tree with depth T. From simulations, we found that this approximation produces accurate results when $2^T < N$ which is the parameter space for which the coalescent approximation does not. For most biological populations, this restricts T to a rather limited set of parameter values, but often, this is a region of great interest. Some definitions and properties of this process are discussed in the following section, which can be skipped by the less mathematically interested reader.

Our goal is to characterize the stochastic process by which segments of ancestral genetic material are recombined to form the genome of a particular person of interest (the proband). We call this the ancestor copying process, which represents the line of descent of the proband's genome as a function of the genomic position. Label the parents of an individual as the 'left' and a 'right' parent, respectively. The ancestry of an individual in a particular position in

the genome is then determined by the choices of left and right parents back in time on the pedigree.

In investigating IBD probabilities, Donnelly (1983) considered this ancestry as a random walk on a hypercube, with each vertex corresponding to the set choices of left or right parents for *every* individual in the pedigree. For a perfect binary tree, the size of this state space is super-exponential in T, which Donnelly (1983) was able to considerably reduce by using symmetries in the transition matrix. For the ancestry copying process, we cannot use these symmetries in the same way, and instead directly integrate over hidden recombination events.

We instead represent this ancestry using dyadic intervals. At a position in the genome, x, the ancestor copying process N_x takes a value from the half-open interval $[0, 1)$. The dyadic intervals N_x is contained in correspond to the ancestors this position was inherited from. We define dyadic intervals to be half-open intervals of the real line of the form $I_{j,k} = [k2^{-j}, (k+1)2^{-j})$ for $j, k \in \mathbb{Z}$, $k < 2^j$. Dyadic intervals are isomorphic to the nodes of binary trees in that every dyadic interval is the union of two unique disjoint dyadic intervals. We use the following notation to denote the left and right halves of a dyadic interval $I_{j,k}$:

$$I^{\ell}_{j,k} = [k2^{-j}, (2k+1)2^{-j-1})$$
$$I^{r}_{j,k} = [(2k+1)2^{-j-1}, (k+1)2^{-j}).$$

We denote the length of a dyadic interval by $|I_{j,k}| = 2^{-j}$ and define the distance between two dyadic intervals, $d(I, J)$, to be the length of the shortest dyadic interval containing both. For a dyadic interval I, we define I' to be the dyadic interval with $2|I| = |I'|$ such that $I \subset I'$ and I^* to be the set difference of I' and I.

We associate an ancestor to each dyadic interval in $[0, 1)$: the proband to $I_{0,0}$, the left parent to $I_{1,0}$, the right parent to $I_{1,1}$, the left parent's left parent to $I_{2,0}$, etc. The value of the ancestor copying process at a particular position represents the ancestors the proband inherited that position from, e.g. if the ancestor copying process is less than $\frac{1}{2}$, then the proband inherited that position from the left parent, or if is greater than or equal to $\frac{3}{4}$, then the proband inherited that position from the right-most grandparent (and consequently the right parent). A realization of the ancestor copying process is given in Figure 2.1.

The defining property of the ancestor copying process is that its distribution does not change after a generation of recombination. The process of recombination between two parental genomes can be described by a two-state Markov process, R_x, which switches between 0 and 1 at rate 1. If N_x and N'_x are the independent ancestor copying processes of the two parent, which are jointly independent of R_x, then

$$N_x \overset{\mathrm{d}}{=} \tfrac{1}{2} R_x N_x + \tfrac{1}{2}(1 - R_x)(1 + N'_x). \tag{2.1}$$

This property makes it clear that conditional on R_x, the behavior of N_x in the range $[0, \frac{1}{2})$ is independent of its behavior in $[\frac{1}{2}, 1)$. In fact, this property can be extended to any mutually disjoint collection of dyadic intervals:

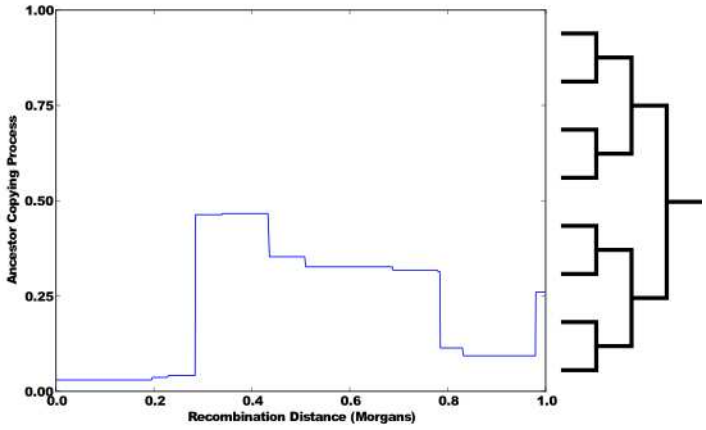

Figure 2.1: **A realization of the ancestor copying process.** In this case, the process stays in the interval $[0, \frac{1}{2})$, indicating that this length of chromosome was inherited entirely from the proband's left parent. The process jumps between $[0, \frac{1}{4})$ and $[\frac{1}{4}, \frac{1}{2})$ three times, indicating that each left grandparent contributed two blocks to the proband. The pedigree, up to the proband's 8 great-grandparents is shown on the right. Each ancestor has been placed in their corresponding dyadic interval.

Theorem 2.2.1 *For a dyadic interval A, the processes $N_x \mathbf{1}\{N_x \in A\}$ and $N_x \mathbf{1}\{N_x \notin A\}$ are conditionally independent given $\mathbf{1}\{N_x \in A\}$.*

An intuitive explanation for this theorem is that because there is no inbreeding, ancestors which are not lineal descendants will be unrelated, and hence independent. The mathematical proof, as with all others in the chapter, is presented in the appendix at the end of the chapter.

To characterize the ancestor copying process, we want to find the rate at which N_x leaves a dyadic interval I:

$$n_I = \lim_{x \downarrow 0} \frac{1 - \mathbb{P}_I(N_x \in I | \mathcal{N}_0)}{x}$$

and the transition rates between disjoint dyadic intervals I and J:

$$n_{I,J} = \lim_{x \downarrow 0} \frac{\mathbb{P}_I(N_x \in J | \mathcal{N}_0)}{x},$$

where $\mathbb{P}_I$ is the measure induced by conditioning on $N_0 \in I$ and $\mathcal{N}_0 = (\{N_x : x \leq 0\}$.

Theorem 2.2.2 *The length over which N_x remains in a dyadic interval is exponentially distributed, with rate given by*

$$n_{I_{j,k}} = j.$$

Theorem 2.2.3 *The transition rates between disjoint dyadic intervals is given by*

$$n_{I,J} = \prod_{i \in P(I,J)} \frac{1}{2} + \left(\mathbf{1}\{T_i > T_{i^*}\} - \frac{1}{2} \right) \exp(-2T_{i'})$$

with

$$T_I = \sup\{x < 0 : N_x \in I\}$$

and

$$P(I,J) = \{i \in \mathcal{I} : |i| < d(I,J), J \subset i\}.$$

The rate at which N_x leaves dyadic intervals depends only on the length of the dyadic interval, which is in accord with the results of Baird et al. (2003), Pool & Nielsen (2009), and Gravel (2012) regarding the exponential distribution of genetic distance between recombination events. However, the process is not Markov, because the transition rates depend on the the values of N_x for $x \leq 0$ and not just N_0.

The MWF and SMC models assume that segments are inherited from distinct ancestors, but for the PBT model, multiple segments can be inherited from the same ancestor. The probability of this event decreases as T increases, confirming the prediction given in (Baird et al. 2003).

2.3 Simulations

As we explain in the results, when there is a single pulse of admixture, the Markov models, (MWF, SMC, and SMC') produce admixture tracts whose lengths are independent and exponentially distributed. For the other models, we first wrote Monte-Carlo simulations which assigned an ancestor to each recombination segment. For the coalescent model, we used code which was essentially identical to the program ms (Hudson 2002), with two modifications: the backwards process stops at the time of admixture, instead of when only one lineage remains, and the simulation starts with just one lineage. The extant lineages at the time of admixture are then traced forward in time to find which recombination segments they contribute.

For the PBT model, we used the transition rates from theorem 2.2.3 to efficiently simulate N_x on the dyadic intervals with size at least 2^{-T} in the following manner: The stationary distribution of N_x is uniform on $[0,1)$, so we put N_0 in a dyadic interval, I, with length 2^T, chosen uniformly at random. The length for which N_x remains in this interval has an Exp(T) distribution. Note that $n_{I,I^*} = n_{I,(I')^*} = n_{I,(I'')^*} = \cdots = 1$, and that $I, I^*, (I')^*, \ldots$ form a partition of $I_{0,0}$ so we first determine which of these dyadic intervals N_x jumps to. Conditional on this, we then recursively determine which of the left and right dyadic intervals contains N_x, until we have narrowed N_x down to a dyadic interval of length 2^{-T}. As we do this, we also update the values of the T_I's. One of the advantages of the dyadic interval

representation is that it allows efficient simulations of pedigree structure by simulating a stochastic process on $[0,\ 1)$ instead of representing full pedigrees for each segment of the genome as a linked list in the computer memory.

The WF model is the same as the PBT model, with the exception that inbreeding is allowed. We still represent the pedigree as a perfect binary tree, with the caveat that some of the nodes are taken to represent the same ancestor. For the simulation, this means that some of the T_i's for different dyadic intervals which represent the same ancestor will in fact be equal. Generating the entire pedigree is computationally expensive for large T, so we only extend the pedigree as is needed i.e., as N_x jumps to previously unvisited dyadic intervals.

After assigning an ancestor to each recombination segment, we then independently label each ancestor as migrant or non-migrant, with probabilities m and $1-m$, respectively. allowing us to demarcate admixture tracts. For each set of admixture parameters, we used a simulated a segment of genome 30 times longer than the average tract length. To minimize edge effects, we only examine the tracts from the middle third of this segment.

2.3.1 Models of multiple admixture pulses

The Markov models (MWF, SMC, and SMC') predict that admixture tracts resulting from one pulse of admixture will have exponentially distributed lengths, while those resulting from two (or more) pulses of admixture will have length distributions which are the mixture of two (or more) exponentials. On the other hand, the Wright-Fisher model produces admixture tracts which are non-exponential, even in the one-pulse scenario. As a result, when analyzing the data using a Markov model, it is possible to mistakenly conclude that the observed tract length distribution cannot be explained by just one pulse of admixture, when in fact it can be, but only by using the more complex Wright-Fisher model.

We investigated the probability of this happening when using a likelihood ratio test to distinguish between an exponential distribution vs. a mixture of two exponentials. To draw from the null distribution, we simulated 10^4 admixture tracts with exponentially distributed lengths and found the maximum log-likelihood of these under a mixture model, with two exponentials, i.e.

$$\mathcal{L}(p, a, b|x) = \prod_{i=1}^{10^4} \left[pae^{-ax_i} + (1-p)be^{-bx_i}\right],$$

where each x_i is the length of a admixture tract. This maximization was done by a standard Expectation Maximization (EM) algorithm. The 100 initial random values p_0, a_0, and b_0 were repeatedly updated by first computing the posterior probabilities:

$$r_{i,t} = \frac{p_t a_t e^{-a_t x_i}}{p_t a_t e^{-a_t x_i} + (1-p_t) b_t e^{-b_t x_i}},$$

and then the likelihood-maximizing posterior means:

$$\hat{p}_{t+1} = \frac{\sum_{i=1}^{10^4} r_{i,t}}{10^4}$$
$$\hat{a}_{t+1} = \frac{\sum_{i=1}^{10^4} r_{i,t}}{\sum_{i=1}^{10^4} r_{i,t} x_i}$$
$$\hat{b}_{t+1} = \frac{\sum_{i=1}^{10^4} (1 - r_{i,t})}{\sum_{i=1}^{10^4} (1 - r_{i,t}) x_i}.$$

The values were updated until the log-likelihood improvement was less than 10^{-3}. We took the highest log-likelihood value resulting from these 100 optimizations to be the maximum log-likelihood under the mixture model for this sample.

2.3.2 Tests of a single admixture pulse

To test the null hypothesis of a single admixture event, we define a likelihood ratio test statistic, S, by subtracting the maximum log likelihood value under the full model with two admixture events from that obtained for a model allowing only a single admixture event. The asymptotic distribution for this test statistic is not known, because some parameters of the alternative hypothesis are not estimable under the null hypothesis. This implies that the general asymptotic likelihood theory is not applicable. To obtain critical values for this test statistic we instead used parametric simulations under the null hypothesis and assuming independent exponentially distributed tract lengths. We simulated 10^5 samples to approximate the critical values corresponding to significance levels of $p = 0.05$ and $p = 0.02$ a range of values for T and for $m = 0.1$, 0.3, and 0.5. We then compared this distribution of log-likelihood ratios to log-likelihood ratios obtained in the same way for simulated datasets of 10^4 tracts generated under the Wright-Fisher model with a single admixture event.

2.4 Simulation Results

The models predict that the sampled chromosome can be viewed as a mosaic of recombination segments from chromosomes in generation T. The models agree in predicting that the distance between recombination events, and hence the length of a recombination segment, is exponentially distributed, with scale T^{-1}, but differ in their predictions regarding how recombination segments are inherited from ancestors from the admixing generation. In the following, we use simulations to illuminate these differences.

2.4.1 Admixture tracts lengths, neither *iid* nor exponentially distributed

Recombination fragments are exponentially distributed in the WF model. Under the assumption that all ancestors are distinct, theorem 2.2.2 shows that the distribution of the length of fragments in which an individual has any particular ancestor T generations ago, is also exponentially distributed, with scale T^{-1}. If admixture tracts are assumed to be so rare that they are unlikely to recombine with each other, then admixture tract lengths will therefore also be exponentially distributed, and the process will be well-modeled using the independence assumption of Pool & Nielsen (2009). However, admixture tracts are different from recombination segments, as multiple recombination segments can recombine to form a single admixture tract. This was the situation considered by Gravel (2012). In general, if the lengths of recombination tracts are independent and identically distributed (*iid*) exponential random variables, and each segment is migrant independently and with probability m, then the length distribution of admixture tracts would be found as a geometric mixture of exponential random variables, and consequently be exponentially distributed with scale $[T(1-m)]^{-1}$. However, the second condition is not true. There are two reasons for this. First, as shown by theorem 2.2.3 the ancestry copying process is not Markov. An individual has a finite number of ancestors and recombination can bring together recombination fragments inherited from the same ancestor. As a result, the lengths of migrants tracts will be correlated when T is small. Another factor that contributes to this correlation is the variance in the number of migrant ancestors an individual has. For instance, an individual with one migrant grandparent will have admixture tracts which tend to be shorter than those for an individual with 3 migrant grandparents. The effect of this is illustrated in Figure 1 for $T = 5$. In addition, when T is large, the number of genetic ancestors will be significantly smaller than 2^T. It might be useful to think of this effect forward in time as an effect of inbreeding, in which admixture tracts introduced into the population are broken up by recombination but also joined again by inbreeding. As a result, many fragments in the population segregating after time T will likely be descendants of a relatively few number of larger fragments. The location of smaller fragments will therefore be correlated in the genome, corresponding to the location of the initial admixture fragments, and back recombination has a higher probability than under the *iid* assumption. This effect is illustrated in Figure 1 for $T = 2000$.

Baird et al. (2003) also simulated and commented on the clustering of tracts in the genome. A single tract spanning a larger region may survive the first generations, and then be broken up into smaller fragments in different individuals in the same region of the genome. Martin & Hospital (2011) also examined the problem of correlated tract lengths, but in the context of recombinant inbred lines, and similarly concluded that tract lengths are not independent.

As a consequence of the correlation in tracts lengths along the chromosome, admixture tracts are not accurately modeled as a geometric mixture of *iid* recombination fragments. This effect is illustrated in Figure 2.2. The strongest deviations occur when T is large, or when the admixture proportion is large. The length distribution of admixture fragments

when the admixture proportion is m, corresponds to the distribution of distances between fragments when the admixture proportion equals $1 - m$. In terms of HMM modeling, deviations from exponential distribution of either admixture fragments, or distances between admixture fragments, will violate the model assumptions.

Related results have previously been obtained relating to the theory of junctions. Chapman & Thompson (2002) examined an assumption of independent Poisson distributed junctions among individuals, and independence of junctions within individuals. They noticed that this assumption tends to underestimate the true variance when $T/N > 1$. Although the assumptions in their study is different from ours, in particular we consider descent from multiple migrant individuals and the possibility of recombination between tracts from these individuals, the conclusion reached by (Chapman & Thompson 2002) is essentially similar to the one reached here: tracts are not exponentially distributed when T is large relative to N. Martin & Hospital (2011) examined this problem further in the context of recombinant inbred lines and similarly concluded that tract lengths are not exponential.

The interplay of the non-independence and non-exponentiality of the admixture tract distribution can be illustrated by looking at the distribution of admixture proportions, the proportion of a window which is inherited from migrant ancestors. This is presented in Figure 4, using a window size of 1 cM, in an admixture scenario in which the pattern of admixture tracts is expected to have fixed in the population. The PBT, MWF, and SMC models do not account for the effect of inbreeding, so they predict that admixture tracts will become ever smaller as T becomes larger. As a result, they predict degenerate admixture proportions, i.e. an atom on m. Consequently, these models were not included in figure 2.3. The coalescent, SMC', and WF models do take inbreeding into account, and consequently predict non-degenerate limiting distributions for the admixture proportion.

For both values of m, the distribution predicted by the WF and coalescent models has a larger variance than that predicted by SMC', while having the same mean. For small values of m, this is because admixture tracts are likely to be clustered, and have either zero or a larger number of tracts than predicted by SMC'. For large values of m, this higher variance is better explained by the fat tails of the admixture tract length distribution.

2.4.2 Coalescent with Recombination

The coalescent provides an approximation to the WF model that is in general excellent, but may be less so when considering the dynamics shortly before the time of sampling (Wakeley et al. 2012). In the present context this means that the coalescent approximates the WF model well when T is large, but not necessarily so for small values of T. The correlation that arises due to inbreeding is well-modeled by the ARG, but the correlation due to a small number of ancestors in the pedigree in the very recent ancestry is not. This is shown in Figure 1. For small values of T, the coalescent does not accurately capture the correlation structure. As a consequence, the distribution of admixture tract lengths is not well-modeled when T is small (Figure 2), particularly for large migration fractions ($m = 0.9$). In an admixed population, the distribution of tracts originating from the population contributing

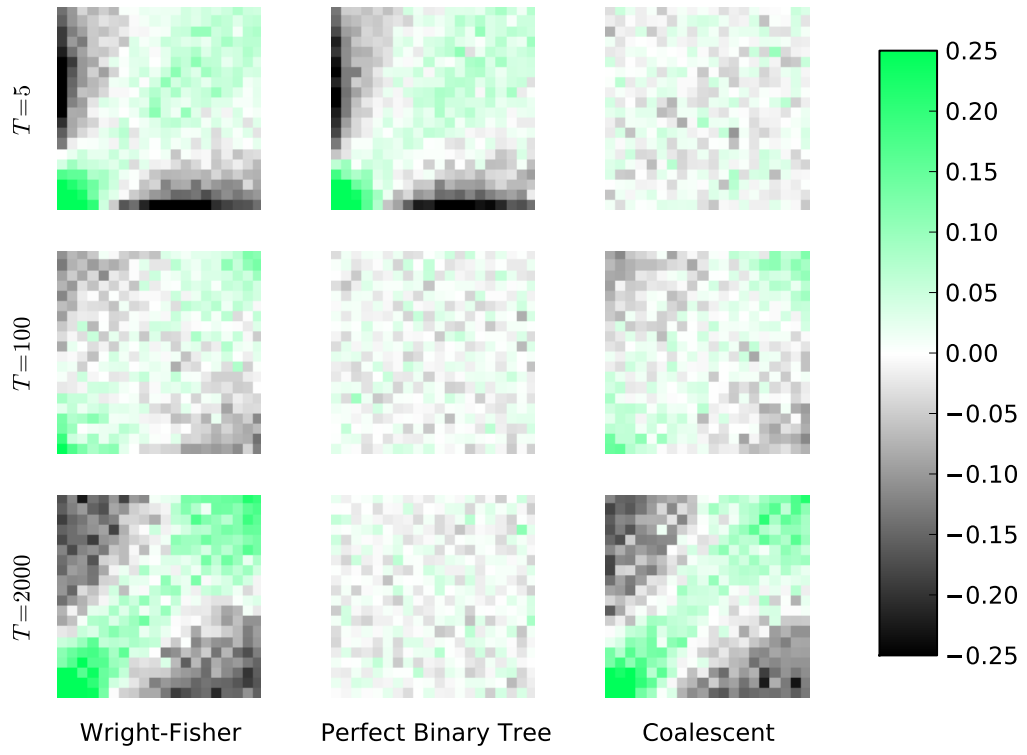

Figure 2.2: **The correlation of the lengths of consecutive admixture tracts for the WF with $2N = 1000$ (red), PBT (green) and coalescent (blue) models.** In all cases the admixture fraction is $m = .95$. Admixture tract lengths were transformed into the unit interval by their empirical quantiles, so uncorrelated lengths would produce an entirely white square. The simulations were run with a population size of $2N = 2000$.

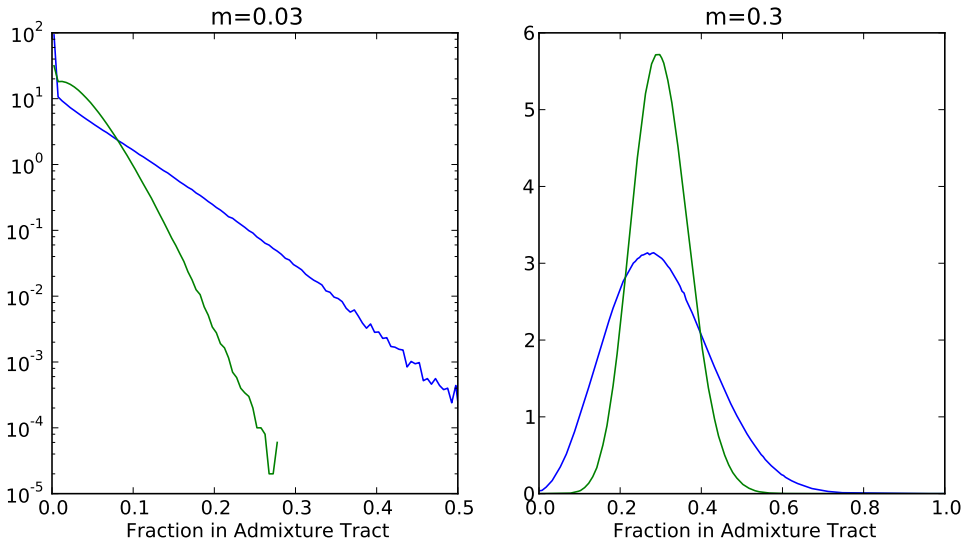

Figure 2.3: **Distributions of the fraction of 1cM windows that are parts of admixture tracts, for two values of m.** Parameters for the two simulations were otherwise the same, with $N = 5\times10^3$ and $T = 2\times10^4$. The distribution under the SMC' model is in green and the distribution under the coalescent and Wright-Fisher models is in blue. Note that the left graph is plotted on a log scale.

most of the genetic material are far from exponentially distributed. However, the effect rapidly diminishes as T increases.

2.4.3 Markovian Models

The MWF, SMC, and SMC' models all generate admixture tracts with exponentially distributed lengths. In these models, admixture tracts follow a geometric mixture of *iid* exponential random variables. In each of these Markovian models, the ancestry of a recombination segment only depends on the ancestry of the recombination segment to its left. As a result, the number of recombination segments that make up a admixture tract will be a geometric random variable. The geometric mixture of *iid* exponential random variables results in another exponential. Under the MWF model, each recombination segment is inherited from a distinct ancestor in generation T. Each of these ancestors is from the admixing population with probability m, so admixture tracts lengths will be exponentially distributed with scale $[T(1-m)]^{-1}$, as previously discussed. In the SMC, the recombined lineage cannot coalesce back to the current marginal tree, so as in the Markovian WF model, each recombination segment will be descended from a distinct ancestor and admixture tracts lengths will again be exponentially distributed with scale $[T(1-m)]^{-1}$. In SMC', back coalescences to the current marginal tree are possible, and occur with probability $1-2N(1-e^{-\frac{T}{2N}})/T$. In this event, the recombination segment will be migrant if and only if the previous segment was. Therefore, the probability that the segment on the right of a recombination point is migrant, given that the segment on the left was, is

$$\left[1-\frac{2N}{T}\left(1-e^{-\frac{T}{2N}}\right)\right]+\left[\frac{2N}{T}\left(1-e^{-\frac{T}{2N}}\right)\right]m=1-\frac{2N}{T}(1-m)\left(1-e^{-\frac{T}{2N}}\right),$$

so admixture tract lengths will have an $\textsc{Exp}[2N(1-m)(1-e^{-\frac{T}{2N}})]$ distribution. When $2N \gg T$, this is the approximately the same distribution given by the other two models, but for fixed $2N$ and as $T \to \infty$, SMC' makes the more accurate prediction that the average tract length goes to the non-zero value of $[2N(1-m)]^{-1}$.

These models may fail to give accurate predictions both for both small and large values of T. These are two separate effects. When T is small they give inaccurate predictions for the same reasons as the coalescent. In particular, they do not accurately model the correlation due to a fixed number of ancestors in the pedigree and the possibility of back-recombination. For this reason, tracts length distributions do not fit well, especially for large values of m.

For large values of T they fail because they do not accurately model the effect of inbreeding. The MWF model and the SMC give identical predictions (Figure 2.4). When T is large, they underestimate the length of admixture tracts for small values of m. For large values of m they underestimate the variance in tract length. In either case, the fit of tract length distribution to that expected under the WF model, or the coalescent, is poor. In the coalescent and WF models, nonadjacent segments may be descendants of the same ancestor, an event which occurs with higher probability as T increases. The overall effect of

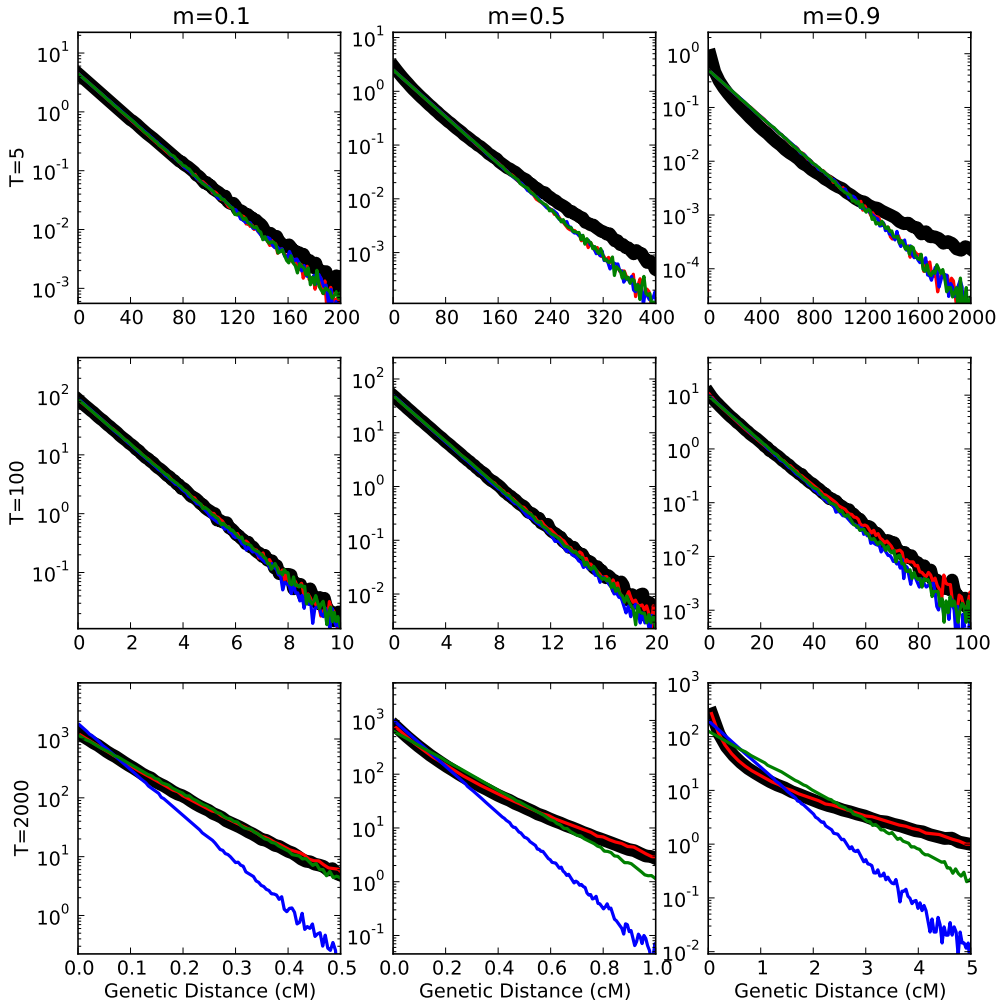

Figure 2.4: **Admixture tract length distributions for the MWF and SMC (both blue), SMC' (green), coalescent (red) models compared to the distribution under the WF model (thick black).** Note that the y-axis is shown on a logarithmic scale. The simulations were run with a population size of $2N = 2 \times 10^3$. For $T = 5$, the former three models give exponential distributions and do not match the WF distribution. For $T = 2000$ the coalescent and WF distributions are the same.

this is that the Markovian models are too likely to assign more distinct ancestors to a given length of chromosome, which increases the probability that some section was inherited from a non-migrant ancestor. The error for the SMC' is less than that of the SMC and Markovian Wright-Fisher model (Figure 2.4).

2.4.4 Perfect Binary Tree

In the Methods section, we derived a genealogical model that can be used to study tract length distributions when T is small. This process captures the correlation structure and admixture tract length distribution of the full WF model for small T (Figures 2.2 and 2.5), something that the other approximative models explored here fail to do. However, the model does not accurately describe the dynamics when T is large, as it assumes that all ancestors from generation T are distinct. For $T > \log_2 N$, this is not possible, and some ancestors must necessarily be the same.

This is consistent with the result of Baird et al. (2003), which found that asymptotically for large T, the probability that an individual inherits multiple blocks from one ancestor goes to zero. In this limit, where every recombination segment is inherited from a distinct ancestor, admixture tracts lengths will be *idd* exponential, as in the case of the Markov models.

2.4.5 Admixture Tracts as distances between junctions

We further compare our results with the results of Baird et al. (2003) to illustrate the effect of considering multiple ancestors of an individual and the effect of assumptions regarding crossover interference. Baird et al. (2003) consider the distribution of the lengths of genetic material inherited from one individual, in a branching-process model with complete interference, i.e. assuming at most one recombination event on a chromosome each generation. The found that the density, in z, for this distribution is given by

$$\frac{(1-z)^{T-1}\left(2T + T(T-1)\frac{y-z}{1-z}\right)}{1+yT},$$

where y is the recombination probability and T is the number of generations. When m is small, e.g. 0.01, most admixture tracts will be inherited from just one migrant ancestor. In this scenario, the Baird distribution is comparable to the admixture tract length distribution (Figure 2.6).

When $T = 5$, the Baird distribution differs from the WF and PBT models because it uses a different model of interference. Under its assumption of complete interference, no tract can span more than a map distance of y, whereas the other two models have no such maximum. In the bottom row, where $T = 2000$, both the Baird distribution and the PBT model fail to account for the back-coalescence of different fragments, and consequently predict tracts that are shorter than under the WF model. However, there are no effects with regards to their different assumptions about recombination interference. For $T = 100$, when the effects of

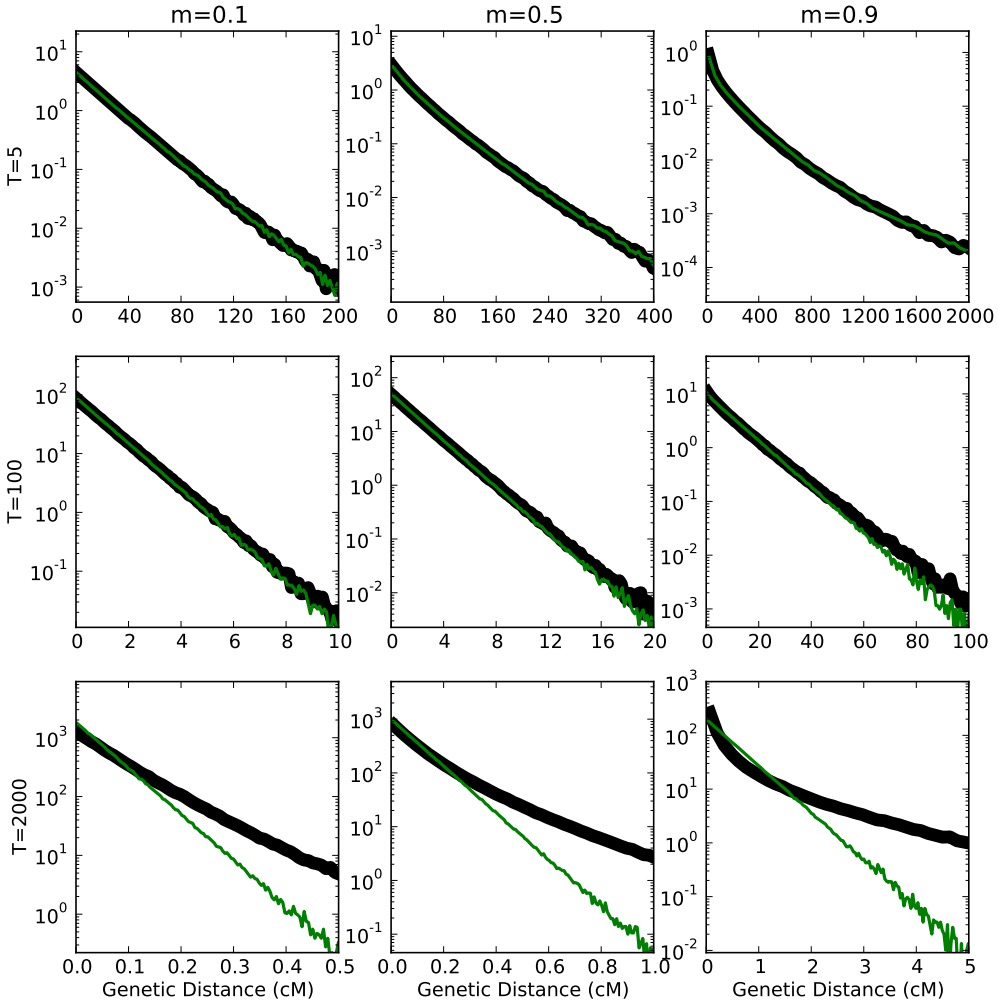

Figure 2.5: **Admixture tract length distributions for the PBT model (green) and the WF model (thick black).** The simulations were run with a population size of $2N = 2 \times 10^3$. Note that the y-axis is shown on a logarithmic scale. For $T = 5$, the PBT model matches the WF model closely, while for $T = 2000$, it does not, and has an exponential distribution instead.

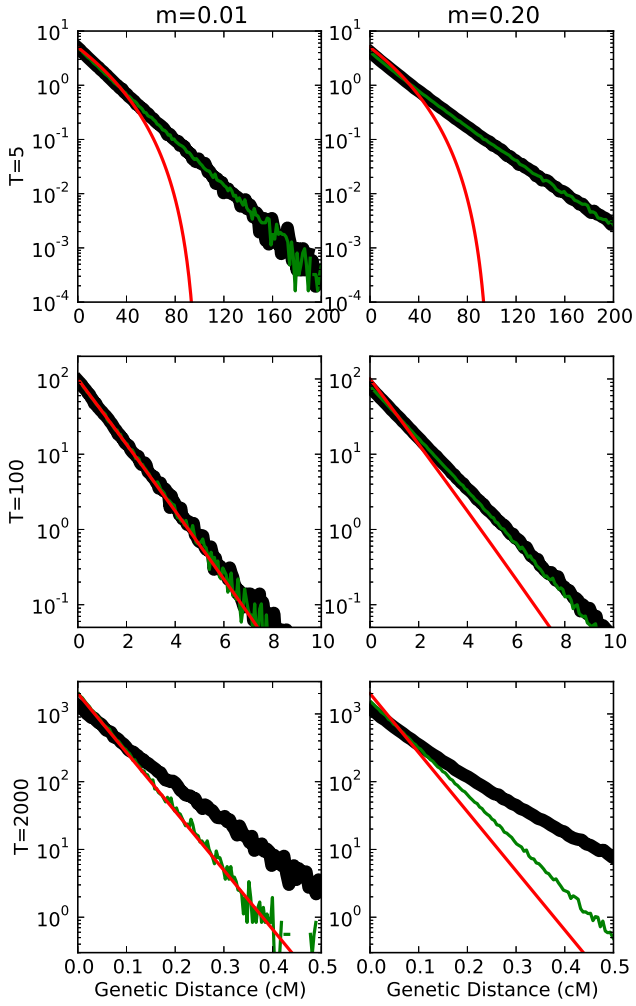

Figure 2.6: **Tract length distributions for the Baird distribution (red), PBT model (green) and the WF model (thick black).** The WF simulations were run with a population size of $2N = 2 \times 10^3$. Note that the y-axis is shown on a logarithmic scale. When m is small and at intermediate time scales, all three models agree.

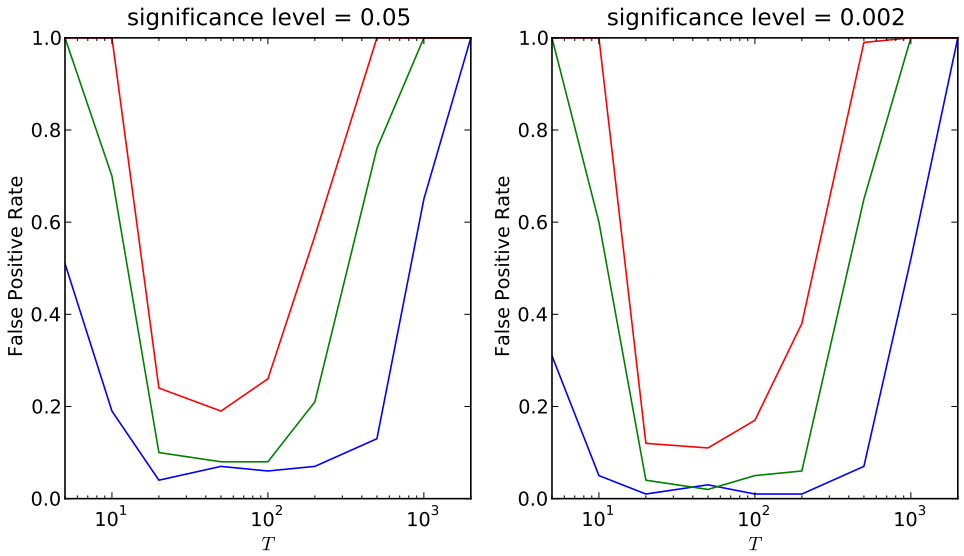

Figure 2.7: **Probability of erroneously inferring two pulses of admixture as a function of T, when using a MWF or SMC' null model.** The red, green, and blue lines correspond to $m = 0.5$, 0.3, and 0.1. The left plot is for a likelihood-ratio test with $\alpha = 0.05$ and the right plot is with $\alpha = 0.002$.

back-coalescence are negligible, all three models predict the same distribution, despite their different assumptions.

When m is not small, the Baird distribution fits less well, which is shown in the right column. This is mainly because each admixture tract is now more likely to be composed to genetic material inherited from multiple migrant ancestors.

2.4.6 Likelihood ratio test of the number of admixture pules

To determine the effect of wrongly assuming *iid* exponential tract lengths for inferences for real data, we implemented a likelihood ratio test and tested the null hypothesis of one admixture pulse, against the alternative of two admixture pulses, on data simulated under the null hypothesis. The false positive rate, defined as a fraction of these log-likelihood ratios which exceeded the critical value (obtained using simulations), was plotted as a function of T, and is shown in Figure 2.7. Notice that there is a strong excess of false positives, particularly when T is large or small. The false positive rate is less for intermediate values. This is explained by the observations from the previous sections, showing that the assumption of *iid* exponential tract lengths is particularly poor when T is very small (due to finite number of ancestors in the pedigree) or larger than N (due to inbreeding).

2.5 Discussion

We have found that under many scenarios, the Wright-Fisher model produces admixture tracts whose lengths are not well approximated as independent, exponential random variables. There are two major effects that are important to distinguish: the effect of a finite number of ancestors in the pedigree for small values of T and the effect of inbreeding for large values of T. Both of these effects cause deviations from the *idd* exponential assumption.

When using an HMM for ancestry deconvolution, the Markov model provides a prior on tract lengths. If there if signal regarding local ancestry in the data, then misspecification of this prior may not matter a great deal. However, for parametric population genetic data analysis, i.e. estimating the number of timing of admixture events, it may be desirable to consider possible biases incurred due to assumptions regarding exponential tract lengths. One way to verify inferences of multiple admixture pulses would be to compare the simulated tract length distribution under the WF model to the data.

The magnitude and direction of the estimation bias will depend on the model and the values of m and T. For small values of T, Figure 2.4 shows that the Markov models underestimate the number of long tracts. Consequently, estimates of T based on the number these longer tracts will be downwardly biased.

The biases can be avoided by using the Wright-Fisher, instead of a Markov, model to construct a prior for the local ancestry distribution. However, there are no known computationally efficient algorithms for integrating over this prior. However, efficient inference under the perfect binary tree model may be possible, because of the conditional independence given by equation 2.1. When T is small, this would be a good approximation to inference under the Wright-Fisher model. As the simulations show, when $20 < T \ll 2N$, all of the models produce approximately the same tract length distributions, so in this region of the parameter space, there will be minimal bias from using a Markov model.

The deviations from a Markov model explored here, may also affect methods that do not directly attempt to estimate admixture tract distributions. For example, ROLLOFF (Moorjani et al. 2011) assumes that the probability that two sites a distance r apart are linked after T generations, is given by $\exp(-rT)$, and uses this to make a prediction about the value of a correlation coefficient. Under the PBT model, this probability is $((1+\exp(-2r))/2)^T$, and under the WF model, this probablity is $(1-1/N)^T((1+\exp(-2r))/2)^T$. For some values of N, r, and T, these probabilities are approximately equal, but for others they are not. This suggest that further analyses might be warranted on the statistical properties of methods such as ROLLOFF (Moorjani et al. 2011).

Throughout this chapter, we have assumed that admixture occurred in a single generation. This is a highly restrictive and, in most cases, unrealistic assumption. In real data analysis, the effects of such assumptions should be carefully considered. However, the basic conclusions regarding distributions of tract length as functions of T are still valid. Our results can be extended to more complicated scenarios of multiple admixture events, or continuous gene-flow, by integrating over admixture times as in (Pool & Nielsen 2009). For the PBT model, continuous gene-flow, as well as overlapping generations, results in pedigrees which

are still binary trees, but of uneven depth. Consequently, this same technique will also allow us to relax the assumption of non-overlapping generations.

In our mathematical analysis and simulations, we have assumed that recombination events occurs according to a Poisson process and have ignored the possibility of crossover interference. For large values of T this approximation may be quite accurate, but for small values of T, crossover interference could potentially have a strong effect on the results, as illustrated in Figure 2.6. However, the transition rates of the ancestor copying process are simple functions of the mapping function induced by the model of crossover interference. The binary tree process under other models of crossover interference with known mapping functions, would typically still be mathematically tractable. Future methods for ancestry deconvolution and parametric admixture inference should seek to incorporate such mapping functions in addition to the non-Markovian properties of the ancestry process which has been the main focus of topic of this chapter.

2.6 Appendix

Most of these proofs are by induction on the length of the dyadic interval(s) in question. Towards this end, we will couple the two sides of equation 2.1 by introducing independent ancestry-copying processes S_x and D_x and letting

$$N_x \equiv \tfrac{1}{2}R_x S_x + \tfrac{1}{2}(1 - R_x)(1 + D_x). \tag{2.2}$$

By equation 2.1, N_x is also an ancestry-copying process.

Proof of theorem 2.2.1

The theorem is trivially true in the case when this length is 1, i.e. $A = I_{0,0}$.

Suppose the theorem holds for dyadic intervals with length greater than or equal to 2^{-j} and let A be a dyadic interval with size 2^{-j-1}. Without loss of generality, assume that $A \subseteq [0, \frac{1}{2})$. Note that $|2A| = 2^{-j}$, so by the inductive hypothesis, $S_x \mathbf{1}\{S_x \in 2A\}$ is conditionally independent of $S_x \mathbf{1}\{S_x \notin 2A\}$ given $\mathbf{1}\{S_x \in 2A\}$. We will use notation

$$S_x \mathbf{1}\{S_x \in 2A\} \perp S_x \mathbf{1}\{S_x \notin 2A\} \mid \mathbf{1}\{S_x \in 2A\}$$

to denote this. Since R_x is independent of S_x, it follows that

$$S_x \mathbf{1}\{R_x = 1, S_x \in 2A\} \perp S_x \mathbf{1}\{R_x = 1, S_x \notin 2A\} \mid \mathbf{1}\{R_x = 1, S_x \in 2A\}.$$

Finally, since $\mathbf{1}\{R_x = 0\} = \mathbf{1}\{R_x = 1, S_x \in 2A\} + \mathbf{1}\{R_x = 1, S_x \notin 2A\}$ and D_x is independent of everything in the above expression,

$$S_x \mathbf{1}\{R_x = 1, S_x \in 2A\} \perp S_x \mathbf{1}\{R_x = 1, S_x \notin 2A\} + \mathbf{1}\{R_x = 0\}(1 + D_x) \mid \mathbf{1}\{R_x = 1, S_x \in 2A\}.$$

By the definition of N_x, $N_x \in A \Leftrightarrow R_x = 1, S_x \in 2A$, so the theorem holds for dyadic intervals of length 2^{-j-1}, and consequently all dyadic intervals.

Proof of theorem 2.2.2

By equation 2.1, the rate at which N_x leaves $I_{1,0}$ or $I_{1,1}$ is this same as the rate at which R_x to switches from 1 to 0 or 0 to 1, respectively. This latter rate is equal to one, so the theorem holds for $j = 1$.

Assume that the theorem holds for all dyadic intervals with length 2^{-j}. Let I be a dyadic interval with length 2^{-j-1}. Note that $\mathcal{N}_0 \subset \sigma(\mathcal{R}_0, \mathcal{S}_0, \mathcal{D}_0)$ and without loss of generality, assume that $I \subset [0, 1/2)$, so that

$$\tfrac{1}{2}R_x S_x + \tfrac{1}{2}(1 - R_x)(1 + D_x) \in I \Leftrightarrow R_x = 1, S_x \in 2I.$$

We can use the law of total probability to find that

$$\begin{aligned}
n_I &= \lim_{x \downarrow 0} \frac{1 - \mathbb{P}_I(N_x \in I|\mathcal{N}_0)}{x} \\
&= \lim_{x \downarrow 0} \frac{1 - \mathbb{E}(\mathbb{P}(R_x = 1, S_x \in 2I|R_0 = 1, S_0 \in 2I, \mathcal{R}_0, \mathcal{S}_0, \mathcal{D}_0)|\mathcal{N}_0)}{x} \\
&= \lim_{x \downarrow 0} \frac{1 - \mathbb{E}(\mathbb{P}(R_x = 1|R_0 = 1)\mathbb{P}(S_x \in 2I|S_0 \in 2I, \mathcal{S}_0)|\mathcal{N}_0)}{x} \\
&= \lim_{x \downarrow 0} \frac{1 - \left(\frac{1}{2} + \frac{1}{2}e^{-2x}\right)\mathbb{E}(\mathbb{P}(S_x \in 2I|S_0 \in 2I, \mathcal{S}_0)|\mathcal{N}_0)}{x} \\
&= \lim_{x \downarrow 0} \frac{\frac{1}{2} - \frac{1}{2}e^{-2x}}{x} + \lim_{x \downarrow 0} \left(\frac{1}{2} + \frac{1}{2}e^{-2x}\right) \frac{1 - \mathbb{E}(\mathbb{P}(S_x \in 2I|S_0 \in 2I, \mathcal{S}_0)|\mathcal{N}_0)}{x} \\
&= 1 + \mathbb{E}\left(\lim_{x \downarrow 0} \left(\frac{1}{2} + \frac{1}{2}e^{-2x}\right) \frac{1 - \mathbb{P}(S_x \in 2I|S_0 \in 2I, \mathcal{S}_0)}{x} \middle| \mathcal{N}_0\right) \\
&= 1 + j.
\end{aligned}$$

where the interchange of limits follows from the dominated convergence theorem and the inductive hypothesis that the limit n_{2I} is equal to j.

Proof of theorem 2.2.3

We show this by induction on the length of J. By equation 2.1, rate at which N_x enters J is the rate at which R_x to switches from 1 to 0 or 0 to 1, which is 1. For $|J| = \frac{1}{2}$, $P(I, J) = \emptyset$, so $n_{I,J} = 1$ and the theorem holds.

To complete the proof by induction, we will need a lemma:

Lemma 2.6.1 *For a dyadic interval I,*

$$\mathbb{P}(N_x \in I|\mathcal{N}_0, N_0 \in I', N_x \in I') = \frac{1}{2} + \left(\mathbf{1}\{N_0 \in I\} - \frac{1}{2}\right)\exp(-2x).$$

We will prove both claims by induction on the length of the dyadic interval I. For $I = [0, \frac{1}{2})$, by equation 2.1, the left-hand side reduces to $\mathbb{P}(R_x = 1|R_0)$, which is equal to the right-hand side. The case of $I = [\frac{1}{2}, 1)$ is analogous, so the lemma is true for dyadic intervals of length $\frac{1}{2}$.

Assume that the lemma holds for dyadic intervals of length 2^{-j} and let I be a dyadic interval with length 2^{-j-1}. Without loss of generality, assume that $I \subset [0, \frac{1}{2})$, so that by equation 2.1,

$$N_x \in I \quad \Leftrightarrow \quad R_x = 1, S_x \in 2I.$$

Additionally, since $I' \subseteq [0, \frac{1}{2})$, we also have that

$$N_x \in I' \quad \Leftrightarrow \quad R_x = 1, S_x \in 2I'.$$

Therefore,

$$\begin{aligned}\mathbb{P}(N_x \in I|\mathcal{N}_0, N_0 \in I', N_x \in I') &= \mathbb{P}(R_x = 1, S_x \in 2I|\mathcal{N}_0, S_0 \in 2I, R_0 = 1, S_x \in 2I', R_x = 1) \\ &= \mathbb{P}(S_x \in 2I|\mathcal{N}_0, S_0 \in 2I, S_x \in 2I', R_0 = 1) \\ &= \mathbb{E}\left(\mathbb{P}(S_x \in 2I|\mathcal{S}_0, S_0 \in 2I, S_x \in 2I')|\mathcal{N}_0, R_0 = 1\right).\end{aligned}$$

Since $2I$ has length 2^{-j} and S_x has the same distribution as N_x, the inductive hypothesis implies that

$$\mathbb{P}(S_x \in 2I|\mathcal{S}_0, S_0 \in 2I, S_x \in 2I') = \frac{1}{2} + \left(\mathbf{1}\{S_0 \in 2I\} - \frac{1}{2}\right)\exp(-2x).$$

Furthermore, since we are conditioning on $R_0 = 1$, $\{S_0 \in 2I\} = \{N_0 \in I\} \in \mathcal{N}_0$. As a result, the conditional expectation evaluates to

$$\mathbb{P}(N_x \in I|\mathcal{N}_0, N_0 \in I', N_x \in I') = \frac{1}{2} + \left(\mathbf{1}\{N_0 \in I\} - \frac{1}{2}\right)\exp(-2x),$$

so the lemma will hold for dyadic intervals of length 2^{-j-1}, and consequently, all dyadic intervals with length less than 1. Assume that the rate at which N_x transitions from any dyadic interval to a disjoint dyadic intervals of length 2^{-j} is as the theorem states and let J be a dyadic interval with length 2^{-j-1}. To each dyadic interval I, we associate the random variable

$$T_I = \sup\{x < 0 : N_x \in I\}.$$

Note that $\max(T_I, T_{I^*}) = T_{I'}$ and $N_{T_{I'}} \in I \Leftrightarrow T_J > T_{J^*}\}$, so by the lemma,

$$\begin{aligned}\mathbb{P}(N_x \in I|\mathcal{N}_{T_{I'}}, N_x \in I') &= \frac{1}{2} + \left(\mathbf{1}\{N_{T_{I'}} \in I\} - \frac{1}{2})\right) \exp(2(T_{I'} - x)) \\ &= \frac{1}{2} + \left(\mathbf{1}\{T_J > T_{J^*}\} - \frac{1}{2})\right) \exp(2(T_{I'} - x)).\end{aligned}$$

Additionally, for $T_I < x < 0$, $N_x \notin I$, so by theorem 2.2.1, the left-hand side also equals $\mathbb{P}(N_x \in I|\mathcal{N}_0, N_x \in I')$. So for J, a dyadic interval of size 2^{-j-1},

$$\begin{aligned}n_{I,J} &= \lim_{x\downarrow 0} \frac{\mathbb{P}_I(N_x \in J|\mathcal{N}_0)}{x} \\ &= \lim_{x\downarrow 0} \frac{\mathbb{P}_I(N_x \in J|\mathcal{N}_0, N_x \in J')\mathbb{P}_I(N_x \in J'|\mathcal{N}_0)}{x} \\ &= \lim_{x\downarrow 0} \mathbb{P}(N_x \in J|\mathcal{N}_0, N_x \in J') \lim_{x\downarrow 0} \frac{\mathbb{P}_I(N_x \in J'|\mathcal{N}_0)}{x} \\ &= \left(\frac{1}{2} + \left(\frac{1}{2} - \mathbf{1}\{T_J > T_{J^*}\}\right) \exp(-2T_{J'})\right) \prod_{i\in P(I,J')} \frac{1}{2} + \left(\mathbf{1}\{T_i > T_{i^*}\} - \frac{1}{2}\right) \exp(-2T_{i'}) \\ &= \prod_{i\in P(I,J)} \frac{1}{2} + \left(\mathbf{1}\{T_i > T_{i^*}\} - \frac{1}{2}\right) \exp(-2T_{i'}).\end{aligned}$$

Chapter 3

Admixture Proportion Moments

3.1 Introduction

It is common in population genetic analyses to consider individuals as belonging fractionally to two or more discrete source populations. The proportion of an individual's genome that belongs to a population is called that individual's 'admixture fraction' or 'admixture proportion'. Programs such as `Structure` (Pritchard et al. 2000), `Eigenstrat` (Price et al. 2006), `Frappe` (Tang et al. 2005), or `Admixture` (Alexander et al. 2009) can jointly estimate these admixture fractions for multiple individuals in a sample, along with the corresponding allele frequencies in each of the source populations. These admixture fractions are often presented in a 'structure plot,' an example of which is shown in figure 3.1. We will henceforth refer to these methods as 'structure analyses'.

This approach has proven highly useful for understanding genetic relationships in many different species, e.g. humans (Rosenberg et al. 2002), cats (Menotti-Raymond et al. 2008), or pandas (Zhang et al. 2007). Other analyses reconstruct admixture tracts for each genome in the sample, by inferring the local ancestry of every position, or window, in each sampled genome (Tang et al. 2006; Maples et al. 2013). In this context, the admixture fraction for a genome is the fraction of its total length that is inherited from a particular source population.

Although structure analyses are not tied to any particular mechanistic model of population history and demography, the admixture fractions and admixture tracts are commonly interpreted to be the result of past admixture events in which modern populations were formed by admixture (or introgression) between ancestral source populations. The distribution of admixture tract lengths has been related to specific mechanistic models of admixture (Falush et al. 2003; Tang et al. 2006; Pool & Nielsen 2009), and has been used to estimate times of admixture (Gravel 2012). However, the admixture proportions themselves also contain information regarding admixture times. Following an admixture event, the variance in admixture proportions within a population will be high, but will thereafter decrease, and will eventually converge to zero in the limit of large genomes. The variance in admixture fractions among individuals contains substantial information about the time since admixture that can be used in addition to the tract length distribution. In some cases, this may be

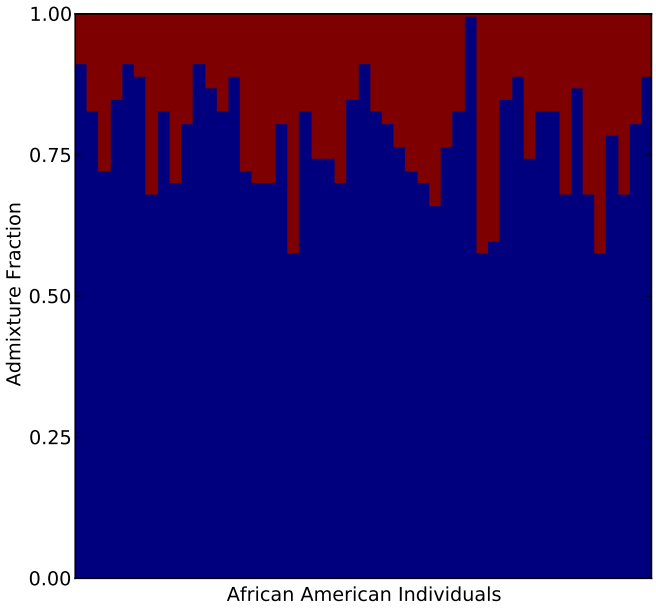

Figure 3.1: **Admixture fractions for 49 African American individuals in the HapMap 3 data.** Source population allele frequencies were estimated using 113 Yoruban and 111 European individuals.

more robust than inferences based on tract lengths, because the length distribution of tracts is often difficult to infer, and is often not modeled accurately by the hidden Markov model (HMM) methods used to infer tract lengths (Liang & Nielsen 2014a). Even in cases where tract lengths can be accurately inferred, studies aimed at estimating admixture times should benefit from using both variance in admixture proportions among individuals and overall admixture tract lengths distributions.

Verdu & Rosenberg (2011) developed a method for computing moments of admixture proportions in a model in which admixed population is formed as a mixture between multiple source populations, allowing for arbitrary gene-flow from the source populations over a number of generations (g). They establish recursions for the moments of the admixture fractions and use these equations to determine how the mean and the variance changes through time in particular admixture scenarios. These moments are expectations for *single* individual's admixture fraction and are averaged over the possibile genealogical histories of the population. As a result, they can be difficult to relate to data because replicates from multiple identical populations rarely are available. In this chapter, we consider a different problem, the problem of calculating sample moments for admixture proportions obtained from individuals in one population.

We extend the model model in Verdu & Rosenberg (2011) to incorporate the effects of recombination and genetic drift by adding a a random union of zygotes component. Recombination is important because even if one half of a chromosome's ancestors are from the first source population, it is unlikely that exactly one half of that chromosome's genetic material is inherited from that population. Genetic drift is important because the individuals in a sample might share ancestors and, therefore, have more similar admixture fractions than expected by chance in a model without drift. The results developed in this chapter should be directly applicable for quantifying the results of a structure analysis.

3.2 General Mechanistic Model

We start by considering admixture fractions in haploid genomes. These haploid admixture fractions can later be paired up to create diploid admixture fractions. The admixture fraction of a (haploid) genome H_i, is the proportion of H_i that is inherited from a particular source population. For notational simplicity, we only consider gene-flow only from one population into another. We will later discuss how to extend this model to multiple admixing source populations. We use the same mechanistic admixture model of Verdu & Rosenberg (2011), and will use its notation where possible. Finally, we use the random union of zygotes model, with a diploid population size of N ($2N$ chromosomes), for genetic drift and recombination, and assume a sample size of n chromosomes from a single population.

In this model, a hybrid population of N diploid individuals forms in generation 1 from two previously isolated source populations. In this first generation, individuals in the hybrid population are from the first source population with probability s_0 or from the second source population with probability $1 - s_0$. In generation $g + 1$, each chromosome is, independently,

from the first source population with introgression probability s_g, or from the hybrid population with probability $1 - s_g$. Chromosomes inherited from the hybrid population are the product of the recombination of the two chromosomes of one individual (zygote), chosen uniformly at random. Finally, these $2N$ chromosomes are paired up to form the N individuals in generation $g + 1$.

Finally, we let the stochastic process $A(\ell)$ represent the local ancestry along a chromosome as a function of ℓ, the physical position:

$$A(\ell) = \begin{cases} 0 & : \ \ell \text{ is descended from first source population} \\ 1 & : \ \ell \text{ is descended from second source population} \end{cases}.$$

The fraction of the chromosome descended from the second source population is given by

$$H = \frac{1}{L} \int_0^L A(\ell) d\ell,$$

where L is the total length of the chromosome.

Assume that g generations after the start of admixture we have randomly sampled n chromosomes from the hybrid population and determined their corresponding admixture fractions, $H_{1(g)}, H_{2(g)}, \ldots, H_{n(g)}$. We are interested in the joint distribution of these n random variables. When $n = 1$ and as $L \to \infty$, this is the admixture fraction considered by Verdu & Rosenberg (2011).

Because the n chromosomes have possibly overlapping geneologies, the admixture fractions are not independent. However, the joint distribution of the admixture fractions does not depend on their ordering, so they are exchangeable. As a result, they can be viewed as being identically and independently (*iid*) drawn from a random distribution $\mathcal{G}$. This random distribution can be interpreted as a function of the random genealogy of the entire hybrid population up to g generations in the past. When g is small, the genealogies of the n samples will be unlikely to differ from n non-overlapping binary trees, so $\mathcal{G}$ will be approximately constant. If g is large however, these genealogies are likely to overlap, and this will no longer be true.

Verdu & Rosenberg (2011) focus on moments of $H_{1(g)}$, in particular on the mean and variance. However, because the admixture fractions are not independent, even as $n \to \infty$, the sample mean and sample variance will converge to the mean and variance of $\mathcal{G}$, which are random quantities. For example,

$$\mathbb{E}(H_{1(g)}) \neq \mathbb{E}(H_{1(g)}|\mathcal{M}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H_{i(g)}$$

$$\text{var}\left(H_{1(g)}\right) \neq \text{var}(H_{1(g)}|\mathcal{M}) = \lim_{n\to\infty} \frac{1}{n-1} \sum_{i=1}^{n} \left(H_{i(g)} - \frac{1}{n} \sum_{j=0}^{n} H_{j(g)} \right)^2,$$

and similarly for higher-order moments. The moments of the admixture factions have two components: randomness from sampling the population genealogy, and randomness

from the sampling of chromosomes. The expressions to left account for both, while the expressions to the right only account for the latter. Variances among individuals within one population correspond to $\text{var}(H_{1(g)}|\mathcal{G})$, while variances over replicate populations correspond to $\text{var}(H_{1(g)})$. This latter value will be larger than the expected sample variance calculated from multiple individuals sampled from the same population, and will rarely be useful for inference purposes.

In the following sections, we will show how the constants on the left-hand side, as well as expectations of the random variables on the right-hand side, can be derived for mechanistic models of introgression. By comparing these expectations to the observed admixture parameters from a sample, we will be able to construct a method of moments estimator for the parameters of the model.

Let k_1 be the sample mean:

$$k_1 \equiv \frac{1}{n}\sum_{i=1}^{n} H_{i(g)}.$$

We can express its expectation in terms of the 1-point correlation function of A:

$$\begin{aligned}\mathbb{E}(k_1) &= \mathbb{E}(H_{1(g)}) \\ &= \frac{1}{L}\int_0^L \mathbb{P}\{A_{1(g)}(\ell) = 1\} d\ell \\ &= \mathbb{P}\{A_{1(g)}(0) = 1\}.\end{aligned}$$

Similarly, let k_2 be the unbiased estimator of the sample variance:

$$k_2 \equiv \frac{1}{n-1}\sum_{i=1}^{n} \left(H_{i(g)} - k_1\right)^2.$$

Its expectation is given by

$$\begin{aligned}\mathbb{E}(k_2) &= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(H_{i,g}^2) - \frac{1}{n(n-1)}\sum_{i,j=1}^{n}\mathbb{E}(H_{i,g}H_{j,g}) \\ &= \mathbb{E}(H_{1,g}^2) - \mathbb{E}(H_{1,g}H_{2,g}).\end{aligned}$$

These expectations can be written in terms of two-point correlation functions of A:

$$\begin{aligned}\mathbb{E}(H_{1(g)}^2) &= \frac{1}{L^2}\mathbb{E}\left(\int_0^L A_{1(g)}(\ell)d\ell \int_0^L A_{1(g)}(\ell)d\ell\right) \\ &= \frac{1}{L^2}\int_0^L\int_0^L \mathbb{E}\left(A_{1(g)}(\ell)A_{1(g)}(\ell')\right)d\ell d\ell' \\ &= \frac{1}{L^2}\int_0^L\int_0^L \mathbb{P}\left\{A_{1(g)}(\ell)=1, A_{1(g)}(\ell')=1\right\}d\ell d\ell'.\end{aligned}$$

Similarly,

$$\mathbb{E}(H_{1(g)}H_{2(g)}) = \frac{1}{L^2}\int_0^L\int_0^L \mathbb{P}\left\{A_{1(g)}(\ell)=1, A_{2(g)}(\ell')=1\right\}d\ell d\ell'.$$

Writing these two correlation functions as

$$\mathbf{v}_{2(g)} = \begin{pmatrix} \mathbb{P}\left\{A_{1(g)}(\ell)=1, A_{1(g)}(\ell')=1\right\} \\ \mathbb{P}\left\{A_{1(g)}(\ell)=1, A_{2(g)}(\ell')=1\right\}\end{pmatrix},$$

we find that

$$\mathbb{E}(k_2) = \frac{1}{L^2}\int_0^L\int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix}\mathbf{v}_{2(g)}d\ell d\ell'. \tag{3.1}$$

In general, the i^{th} k-statistic is an unbiased estimator of the i^{th} cumulant of $\mathcal{G}$, and its expectation can be written as an integral over $[0,L]^i$ of a linear combinations of i-point correlation functions. For example,

$$\begin{aligned}\mathbb{E}(k_3) &= \frac{1}{L^3}\int_0^L\int_0^L\int_0^L \begin{pmatrix} 1 & -1 & -1 & -1 & 2 \end{pmatrix}\mathbf{v}_{3(g)}d\ell d\ell' d\ell'' \\ \mathbb{E}(k_4) &= \frac{1}{L^4}\int_{[0,L]^4}\Big(1 \underbrace{-1}_{4\text{ times}} \underbrace{-1}_{3\text{ times}} \underbrace{2}_{6\text{ times}} 6\Big)\mathbf{v}_{4(g)}d\ell d\ell' d\ell'' d\ell''' \\ &\cdots\end{aligned}$$

Remarkably, the linear combinations required to compute the expectations of the k-statistics correspond exactly to the higher-order disequilibria as defined by Bennett (1952). Furthermore, if instead the we choose to compute the expectations of the h-statistics, which estimate the central moments, the linear combinations would correspond to the higher-order disequilibria as defined by Slatkin (1972).

We next find the recurrence relations these correlation functions satisfy and solve them in the some special cases. In particular we will consider the case of a single admixture event g generations ago and the case of constant gene-flow starting g generations ago.

3.2.1 A Single Admixture Event

We start with a simple case, where introgression only occurs in the founding generation, i.e. $s_g = 0$ for $g > 0$. Using the random union of zygotes model, we can compute $\mathbf{v}_{2(g)}$ in terms of the probabilities from the previous generation:

If two sites at ℓ and ℓ' are on the same chromosome in generation $g+1$, then they were inherited from one chromosome from generation g with probability $[\ell\ell']$ and from two chromosomes from generation g with probability $[\ell|\ell']$. If they are on different chromosomes, then the probability that they are descended from one chromosome in generation g is $\frac{1}{2N}[\ell\ell']$ and the probability that they are descended from two chromosomes is $\frac{1}{2N}[\ell|\ell'] + \left(1 - \frac{1}{2N}\right)$ In matrix notation,

$$\mathbf{v}_{2(g+1)} = (\mathbf{L}_2\mathbf{U}_2)\,\mathbf{v}_{2(g)} = (\mathbf{L}_2\mathbf{U}_2)^g\,\mathbf{v}_{2(0)},,$$

where the the recombination and drift matrices are given by

$$\mathbf{L}_2 = \begin{pmatrix} 1 & 0 \\ \frac{1}{2N} & 1 - \frac{1}{2N} \end{pmatrix}$$
$$\mathbf{U}_2 = \begin{pmatrix} [\ell\ell'] & [\ell|\ell'] \\ 0 & 1 \end{pmatrix}.$$

This is the the same matrix equation (Wright 1933 and Hill and Robertson 1966) derived for the decay of two-locus linkage disequilibrium. The 'alleles' we consider are the local ancestry at ℓ and ℓ'. To the extent possible, our notation will follow (Hill 1974), whose results for measures of multi-locus linkage disequilibria we use. The matrices $\mathbf{L}_2$ and $\mathbf{U}_2$ share $(1 \;\; -1)$ as a left-eigenvector, with corresponding eigenvalues $1 - \frac{1}{2N}$ and $[\ell\ell']$. As a result,

$$\begin{aligned}\mathbb{E}(k_2) &= \frac{1}{L^2}\int_0^L\int_0^L \begin{pmatrix} 1 & -1 \end{pmatrix} \cdot (\mathbf{L}_2\mathbf{U}_2)^g\,\mathbf{v}_{2(0)} d\ell d\ell' \\ &= \frac{1}{L^2}\left(1 - \frac{1}{2N}\right)^g \left(s_0 - s_0^2\right)\int_0^L\int_0^L [\ell\ell']^g d\ell d\ell'. \end{aligned} \tag{3.2}$$

For a model using the Haldane map function, $[\ell|\ell'] = \frac{1-\exp(-2|\ell-\ell'|)}{2}$, this equation becomes

$$\begin{aligned}\mathbb{E}(k_2) &= \frac{1}{L^2}\left(1 - \frac{1}{2N}\right)^g \left(s_0 - s_0^2\right)\int_0^L\int_0^L \left(\frac{1+\exp(-2|\ell-\ell'|)}{2}\right)^g d\ell d\ell' \\ &= \frac{2}{L^2}\left(1 - \frac{1}{2N}\right)^g \left(s_0 - s_0^2\right)\int_0^L (L-\ell)\left(\frac{1+\exp(-2\ell)}{2}\right)^g d\ell d\ell', \end{aligned}$$

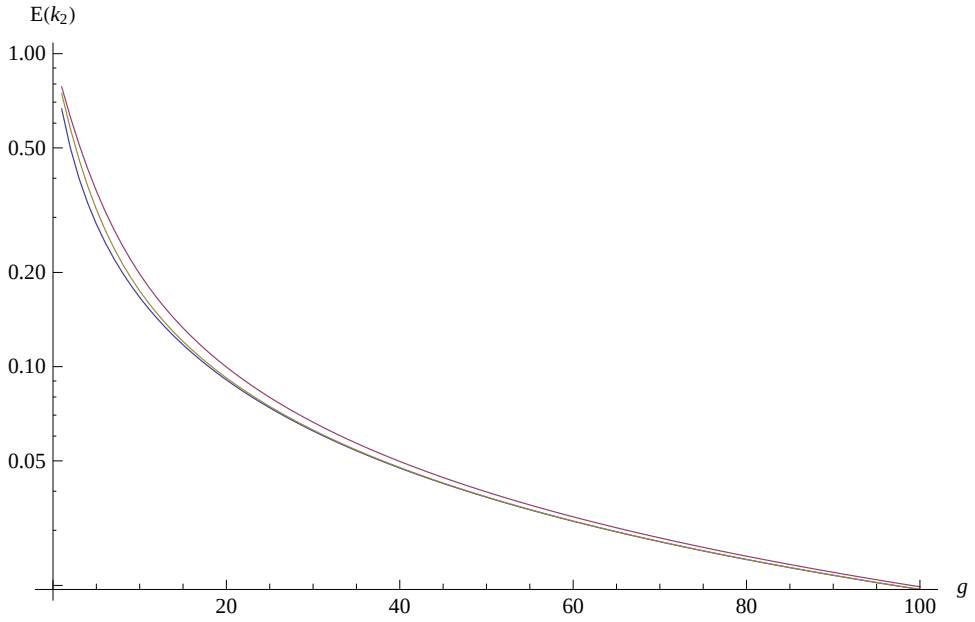

Figure 3.2: **The expected sample variance given by equation 3.1 plotted on a logarithmic scale, for a three different map functions.** We used a map distance of $L = 1$ Morgan and $N = 10^4$. The Haldane map function $(1/2 - e^{-2x}/2)$ is in red, the Kosambi map function $(\tanh(2x)/2)$ is in yellow, and the complete interence map function (x) is in blue. For all values of g, the expectations are ordered in the same order as the map functions, but the difference between the three disappears by $g = 100$.

while for a model of complete crossover inteference on a chromosome of length 1 Morgan, we can get a closed form solution:

$$\begin{aligned}\mathbb{E}(k_2) &= \left(1 - \frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \int_0^1 \int_0^1 (1 - |\ell - \ell'|)^g \, d\ell d\ell' \\ &= \left(1 - \frac{1}{2N}\right)^g \left(s_0 - s_0^2\right) \frac{2}{2+g}.\end{aligned}$$

For predicting the expected sample variance, the difference between these two models is not large, as shown in figure 3.2. For the simulations and inference in this chapter, we will ignore crossover interference, and use the Haldane map function. However, none of the mathematical results of this chapter will require this assumption.

For computing higher-order correlation functions, we find a similar equation

$$\mathbf{v}_{i(g)} = (\mathbf{L}_i \mathbf{U}_i)^g \, \mathbf{v}_{i(0)}. \tag{3.3}$$

Bennett's coefficients for higher-order linkage are left-eigenvectors of the recombination

matrix $\mathbf{U}_i$. For $i = 3$, it is also a left-eigenvector of the drift matrix, so we immediately get that

$$\mathbb{E}(k_3) = \frac{s_0(1-s_0)(2-s_0)}{L^3}\left(1-\frac{1}{2N}\right)^T\left(1-\frac{2}{2N}\right)^T\int_{[0,L]^3}[\ell\ell'\ell'']^G d\ell d\ell' d\ell''.$$

For $i \geq 4$, this is no longer true, but the results of (Hill 1974) can be used to compute $\mathbf{v}_i(g)$ without having to exponentiate the entire drift and recombination matrices. For example, for k_4, the drift and recombination matrices are 15×15, but using the technique in (Hill 1974), we only need to exponentiate a 4×4 matrix to compute $\mathbb{E}(k_4)$.

3.2.2 Varying Migration

If $s_g > 0$ for $s \geq 1$, we obtain a modified version of Equation 3.3:

$$\mathbf{v}_{i(g)} = \mathbf{L}_i\mathbf{D}_{i(g)}\mathbf{U}_i\mathbf{v}_{i(g-1)}, \tag{3.4}$$

where the diagonal matrix $\mathbf{D}_{i(g)}$ has entries giving the probabilities the set of chromosomes, p, in a correlation function are all from the hybrid population in the previous generation:

$$d_{p,p(g)} = (1-s_g)^{|p|}.$$

Note that if $s_{(g)}$ is fixed, then equation (3.4) is linear, and can be solved using a Laplace transform.

3.3 Inference of Admixture Times

The equations in the previous section can be used to develop a method of moments-estimators for admixture parameters by numerically solving the admixture parameters in terms of the expectations for the k-statistics. Substituting in the observed values for the k-statistics gives estimates for the admixture parameter(s).

However, with real data, we only have estimates of the admixture fractions, so some of the variability seen in the distribution of admixture fractions will be due to estimation variability. To account for this, we assume that the estimations errors are additive and *iid*:

$$\hat{H}_{i(g)} = H_{i(g)} + \epsilon_i.$$

Because cumulants are additive,

$$\begin{aligned}\mathbb{E}(k_n) &= \mathbb{E}\left(\kappa_n(H_{i(g)} + \epsilon_i|\mathcal{G})\right)\\ &= \mathbb{E}\left(\kappa_n(H_{i(g)}|\mathcal{G})\right) + \kappa_n(\epsilon_i).\end{aligned}$$

The expectations we have computed are just the term of this sum. To correct for the variability in the estimates, we need to subtract off the second term. We use a block bootstrap to estimate these effects.

One additional complication arises in dealing with genotyping data. We have assumed that we have the ancestry fractions for each haplotype in the sample, but with genotyping data, we instead have their pairwise means: $(H_{1(g)} + H_{2(g)})/2 \ldots$. This is results in a decrease in the expectations of the k-statitics. Conditional on the random distribution $\mathcal{G}$, $H_{1(g)}, H_{2(g)}, \ldots$ are *iid* drawn from $\mathcal{G}$. Cumulants are additive, so we use the law of total expectation to find that

$$\begin{aligned}
\kappa_i\left(\frac{H_{1(g)}+H_{2(g)}}{2}\right) &= \mathbb{E}\left(\kappa_i\left(\frac{H_{1(g)}+H_{2(g)}}{2}\middle|\mathcal{G}\right)\right)\\
&= \mathbb{E}\left(\kappa_i\left(\frac{H_{1(g)}}{2}\middle|\mathcal{G}\right)+\kappa_i\left(\frac{H_{2(g)}}{2}\middle|\mathcal{G}\right)\right)\\
&= 2^{-i+1}\mathbb{E}\left(\kappa_i\left(H_{1(g)}\middle|\mathcal{G}\right)\right)\\
&= 2^{-i+1}\kappa_i\left(H_{1(g)}\right).
\end{aligned}$$

3.3.1 Comparison to Verdu and Rosenberg

The recursion equations given by Verdu & Rosenberg (2011) are different from the ones we have derived. This is partly because we have accounted for the effects of genetic drift and recombination, but also because we are computing the moments of slightly different quantities.

In figure 3.3, we have shown the admixture fractions for five replicate populations 5, 50, and 500 generations after an admixture pulse. The variance that (Verdu & Rosenberg 2011) compute variance over all the replicate populations, while the variance we have computed in this chapter is the expectation of the variance within a single population. When g is small, these similar, but when g is large, the variance within a population goes to zero, but the variance across the replicate populations does not. This effect is shown in Figure 3.4. Initially, both quantities decline exponentially in g, but after $2^g > nLg$, the variance we predict begins to decline linearly instead. This is because variance is inversely proportional to the number of genetic ancestors of the sample. When g is small, the number of genetic ancestors is approximately 2^g. However, the approximate number of recombination events in the sample is approximately bounded by nLg, so when this quantity is smaller than 2^g, it provides a better approximation for the number of genetic ancestors. In this regime, the variance will decline linearly in g.

It is also possible to compute the variance over all population replicates under our model, which allows a direct comparison to Verdu & Rosenberg (2011). In the case of one pulse of admixture, we can now solve equations 3.1 for $\mathbb{P}\left\{A_{1,g}(\ell) = 1, A_{1,g}(\ell') = 1\right\}$ to get

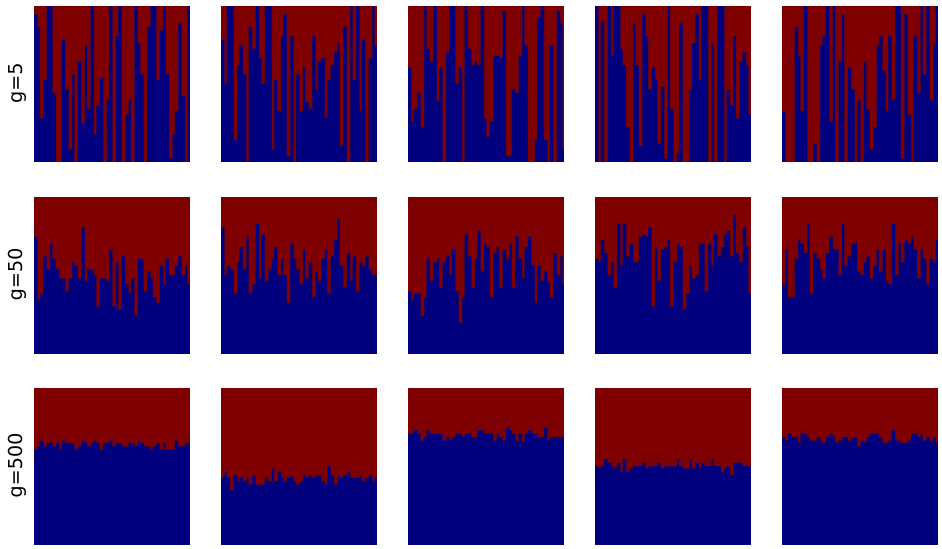

Figure 3.3: **The admixture fractions of five replicate populations (each column) 5, 50, and 500 generations after an admixture pulse.** As the admixture event grows more ancient, the variability within a replicate population decreases, but some variability is still maintained across the populations.

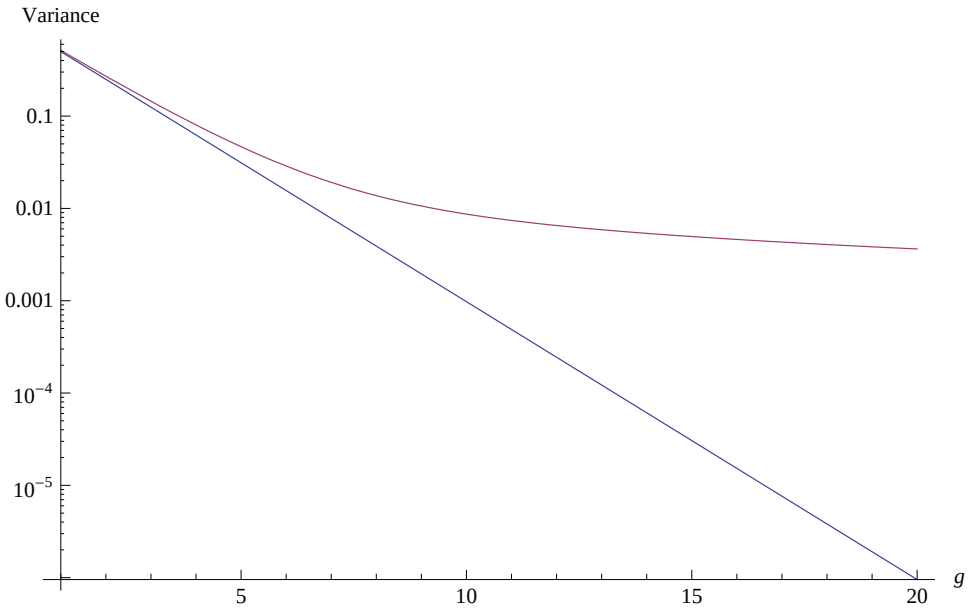

Figure 3.4: **The variance predicted by Verdu & Rosenberg (2011) and equation 3.5, plotted on a logarithmic scale.** The variance we predict (red) is always larger, but the two a very similar when g is small.

$$\begin{aligned}
\text{var}(H_{1(g)}) &= \mathbb{E}(H_{1,g}^2) - s_0^2 \\
&= \frac{1}{L^2}\int_0^L \int_0^L \mathbb{P}\left\{A_{1,g}(\ell) = 1, A_{1,g}(\ell') = 1\right\} d\ell d\ell' - s_0^2 \\
&= \frac{1}{L^2}\left(s_0 - s_0^2\right)\int_0^L \int_0^L 1 - (1 - [\ell\ell']) \frac{1 - [\ell\ell']^g \left(1 - \frac{1}{2N}\right)^g}{1 - [\ell\ell']\left(1 - \frac{1}{2N}\right)} d\ell d\ell. \qquad (3.5)
\end{aligned}$$

This variance and the expectation of the second k-statistic have the same limit as $N \to \infty$, but for finite N, the variance is larger. This is because

$$\text{var}(H_{1(g)}) = \text{var}\left[\mathbb{E}(H_{1(g)}|\mathcal{G})\right] + \mathbb{E}\left[\text{var}(H_{1(g)}|\mathcal{G})\right] = \text{var}[k_1] + \mathbb{E}[k_2].$$

The first variance is small when N is large, but is always non-negative. The difference between this equation and equation 3.1 only becomes significant on a coalescent time scale. In the absence of genetic drift, the admixture fractions are approximately independent, becuase the samples do not share ancestors.

	Observed	Bootstrap	Corrected
k_1	0.777	-2.22×10^{-15}	0.777
k_2	9.00×10^{-3}	2.59×10^{-4}	8.75×10^{-3}
k_3	2.98×10^{-4}	1.60×10^{-5}	2.82×10^{-4}
k_4	-3.99×10^{-5}	-1.41×10^{-6}	-3.85×10^{-5}

Table 3.1: **k-statistics for ASW admixture fractions from HapMap 3 project.**

3.3.2 Application to African American Data

We applied this method to a subset of the ASW, CEU, and YRI data from the HapMap 3 project (3 Consortium et al. 2010). After excluding children from trios, there were the genotypes for 49 ASW, 113 YRI, and 112 CEU individuals. We estimated the admixture fractions using the supervised learning mode of `Admixture`, with the CEU and YRI individuals assigned to separate clusters. The sampling distribution of the admixture fractions was estimated using the block bootstrap with 10^4 replicates and 2678 blocks, giving a block size of approximately 10 CM. The admixture fractions for the 49 ASW samples are shown in Figure 3.1 and the observed k-statistics are given in table 3.1.

We assumed a 3-parameter model of constant admixture. For $g_{start} \leq g \leq g_{stop}$, $s_g = s$ with $s_g = 0$ elsewhere. By matching the block-bootstrap corrected k_2 and k_3 to the predictions of equation 3.1, we obtained a point estimates of

$$\begin{aligned} \hat{s} &= 0.0277 \\ \hat{g}_{start} &= 2 \\ \hat{g}_{stop} &= 11. \end{aligned}$$

We obtained confidence intervals, shown in Figure 3.5, by simulation. For each cell in the grid, we simulated 10^3 replicates under the corresponding g_{start} and g_{stop}, with $s = 1 - k_1^{1/(g_{stop} - g_{start} + 1)}$. For each replicate, we computed the k_2, k_3, and k_4 statistics. A cell was then included in the confidence interval if and only if the corrected k_2, k_3, and k_4 statistics from the HapMap data fall inside a centered interval containing 98.7% of the probability mass of the simulated distribution. This mass was chosen so that under the Bonferroni correction for three tests, there is at least a 95% chance of including the true parameter values in the confidence region.

The point estimates for g_{start} and g_{stop} correspond to the values for which the observed k-statistics are closest to their simulated medians.

3.4 Discussion

We have extended the mechanistic model of Verdu & Rosenberg (2011) to account for recombination and genetic drift. Doing so allows us to apply the predictions of this model

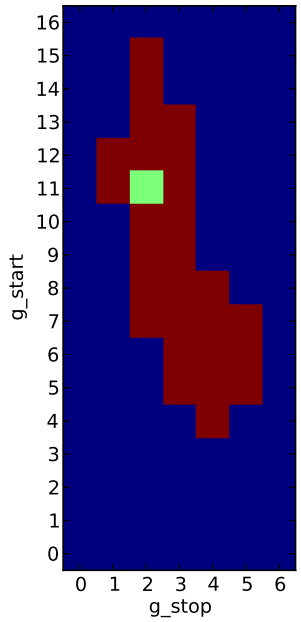

Figure 3.5: **95% confidence region for a model with constant admixture from generations g_{start} to g_{stop}.** The point estimate of $g_{start} = 11$ and $g_{stop} = 2$ generations ago is colored green.

to data. This mechanistic model allows for a large number of parameters. For the purposes of inference, it seems that imposing constraints, i.e. a small number of pulses or constant admixture, will be needed to narrow the search space.

In this chapter, we have assumed that admixture only comes from one source population, this need not be the case. To account for admixture from multiple source populations, equation 3.1 must be modified to account for the probability that haplotypes trace their descent to multiple source populations. Algorithmically, this is feasible, but the notation is cumbersome. The resulting equations are given in the appendix, along with the equations for computing expectations of higher-order k-statistics.

Applications of the method to African-American HapMap data provides estimates of the time since admixture between people of Europe and and African descent in America. Notice that the confidence set for the admixture parameters does not include values of $g_stop = 0$. We interpret this as evidence that admixture rates have declined the last few generations. The point estimate of time gene-flow stopped is $g_stop = 2$. This probably reflects a more gradual reduction in gene-flow within the last 5 generations or so, rather than a discrete stop in gene-flow 2 generations ago. The discreteness is enforced by the model. Also notice that admixture before 15 generations ago can be rejected. With a generation time of 25-30 years, this corresponds to 325-400 years, and is in good accordance with the historical record. The point estimate of the time of first admixture is 11 generations, or approx. 275-330 years ago.

Structure analyses have become one of the most commonly applied tools in population genomic analyses. The theory developed in this chapter allows users of structure analyses to interpret their data in the context of a model of admixture between populations, and should find use in many studies aimed at understanding the history of populations.

3.5 Appendix

These are the matrices for computing $\mathbb{E}(k_3)$. The matrices for computing $\mathbb{E}(k_4)$ are 15×15 and not given here, but can be found in (Hill 1974).

$$\mathbf{v}_{3(g)} = \begin{pmatrix} \mathbb{P}\{A_{1(g)}(\ell) = A_{1(g)}(\ell') = A_{1(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{1(g)}(\ell') = A_{2(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{2(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{1(g)}(\ell'') = 1\} \\ \mathbb{P}\{A_{1(g)}(\ell) = A_{2(g)}(\ell') = A_{3(g)}(\ell'') = 1\} \end{pmatrix}$$

$$\mathbf{U}_3 = \begin{pmatrix} [\ell\ell'\ell''] & [\ell\ell'|\ell''] & [\ell|\ell'\ell''] & [\ell\ell''|\ell'] & 0 \\ 0 & [\ell\ell'] & 0 & 0 & [\ell|\ell'] \\ 0 & 0 & [\ell'\ell''] & 0 & [\ell|\ell''] \\ 0 & 0 & 0 & [\ell\ell''] & [\ell'|\ell''] \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{L}_3 = \frac{1}{4N^2} \begin{pmatrix} 4N^2 & 0 & 0 & 0 & 0 \\ 2N & 2N-1 & 0 & 0 & 0 \\ 2N & 0 & 2N-1 & 0 & 0 \\ 2N & 0 & 0 & 2N-1 & 0 \\ 1 & 2N-1 & 2N-1 & 2N-1 & (2N-1)(2N-2) \end{pmatrix}$$

$$\mathbf{D}_{3(g)} = \begin{pmatrix} 1-s_g & 0 & 0 & 0 & 0 \\ 0 & (1-s_g)^2 & 0 & 0 & 0 \\ 0 & 0 & (1-s_g)^2 & 0 & 0 \\ 0 & 0 & 0 & (1-s_g)^2 & 0 \\ 0 & 0 & 0 & 0 & (1-s_g)^3 \end{pmatrix}$$

When there is migration from both source populations, the recursion relations for the i-point correlation functions will depend on $i-1$-point, $i-2$-point, ... correlations functions as well. As as example, consider the case of $\mathbf{v}_{2(g)}$. Let the introgression probability from the second source population be given by t_g. The recursion equation for $\mathbf{v}_{2(g)}$ now also depends on $\mathbf{v}_{1(g)}$.

$$\begin{aligned} \mathbf{v}_{2(g+1)} &= \mathbf{L}_2 \begin{pmatrix} 1-s_g-t_g & 0 \\ 0 & (1-s_g-t_g)^2 \end{pmatrix} \mathbf{U}_2 \mathbf{v}_{2(g)} + \begin{pmatrix} t_g \\ t_g^2 + 2t_g \mathbb{P}\{A_{1(g)}(\ell) = 1\} \end{pmatrix} \\ &= \mathbf{L}_2 \begin{pmatrix} 1-s_g-t_g & 0 \\ 0 & (1-s_g-t_g)^2 \end{pmatrix} \mathbf{U}_2 \mathbf{v}_{2(g)} + \begin{pmatrix} t_g \\ t_g^2 + 2t_g \mathbf{v}_{1(g)} \end{pmatrix}. \end{aligned}$$

Similarly, the recursion equation for $\mathbf{v}_{3(g)}$ depends on $\mathbf{v}_{2(g)}$ and $\mathbf{v}_{1(g)}$.

Chapter 4

Weighted Three-Locus Linkage Disequilibrium

4.1 Introduction

There are many methods for inferring the presence of admixture, e.g. with f-statistics Reich et al. (2009) or by estimating admixture proportions with programs such as Structure Pritchard et al. (2000) or Admixture Alexander et al. (2009). However, there has been less research on estimating admixture times, possibly because such methods require data which was unavailable until the advent of high-throughput next generation sequencing. Some of these methods use the inferred local ancestry of sequences to construct admixture tract length distributions. Over time, recombination is expected to decrease the average lengths of admixture tracts. This was tract length distribution first worked out in the context of junctions Fisher (1949) and later extended to randomly mating populations by Stam (1980). Baird et al. (2003) first discussed the lengths of tracts descended from a single ancestor. These results informed later analyses of admixture tract length distribution, such as Pool & Nielsen (2009), Gravel (2012), and Liang & Nielsen (2014a). Gravel (2012) also implemented the software program TRACTS, which estimates admixture histories by fitting the tract length distribution, obtained by local ancestry inference, to a exponential approximation.

Another approach, which we will follow in this chapter, is based on the decay of ancestral linkage disequilibrium (LD). In a well-mixed, genetically isolated human populations, linkage disequilibrium decays to zero on a scale of tenths of centiMorgans. However, when an admixed population is founded, it begins with large of amount of linkage disequilibrium, which is a result of the allele frequency differences between the source populations. This occurs even if the LD in the source populations themselves is negliable. The linkage disequilibrium in the admixed population then fluctuates in the generations after its founding, decreasing as a result of drift and recombination, or increasing becuase of additional waves of mgiration. From the LD present in a modern day admixed population, it is possible to make inferences about the population's admixture history. This technique was first in the program ROLLOFF Moorjani et al. (2011) and was later extended by ALDER Loh et al.

(2013).

These two methods use the fact that if an admixed population takes in no additional migrants after the founding generation, the LD present in the population is expected to decay exponentially as a function of distance. The rate constant of this exponential decay is proportional to the age of the founding admixture pulse and so can be used as an estimator. ROLLOFF and ALDER are well suited for inferring the time of the admixture event when the population's admixture history can be approximated as a single pulse. However, it can be important to estimate parameters for admixture histories involving multiple pulses, such as estimating the date of Native American admixture in Rapa Nui Moreno-Mayar et al. (2014) or determining migration patterns in the Americas Gravel et al. (2013). In these instances the expected decay of LD will become a mixture of exponentials. ROLLOFF and ALDER have limited resolution, as they can usually only infer the date of the most recent migration wave Moorjani et al. (2011), or reject the hypothesis of a single pulse admixture Loh et al. (2013).

ROLLOFF and ALDER use the information contained in pairs of sites by looking at the two-locus linkage disequilibrium between them. We use the information in triples of sites by considering three-locus LD. There are two ways of measuring the linkage between n loci. Two-locus linkage disequilibrium decreases geometrically each generation as a result of recombination. Bennett (1952) defines n-locus linkage in a way that this property is maintained. Another property of two-locus LD is that it is equal to the covariance in the allele frequencies between the two sites. Slatkin (1972) defines n-locus LD analogously. For two and three loci, these two definitions coincide, but for four or more loci, they do not.

In this chapter, we will use Bennett and Slatkin's definition of three-locus LD to look at the decay of weighted LD for three sites as a function of the genetic distance between them. We derive an equation that describes the decay of three-locus LD under an admixture history with multiple waves of migration. We then compare the results of coalescent simulations to this equation, and develop some guidelines for when admixture histories more complex than a single pulse can be resolved. Finally, we compute the our method for the Columbian and Mexican samples in the 1000 Genomes data set, using the Yoruba samples as a reference. Fitting a two-pulse model to data, we estimate admixture histories for the two populations which are qualitatively consistent with the results reported in Gravel et al. (2013).

4.2 Model

We use the same random union of gametes admixture model as in Liang & Nielsen (2014b), which is itself an extension of the mechanistic admixture model formulated by Verdu & Rosenberg (2011). In this model, two (or more) source populations contribute migrants to form an admixed population consisting of $2N$ haploid individuals. Each generation in the admixed population is formed through the recombination of randomly selected individuals from the previous generation, with some individuals potentially replaced by migrants from the source populations. For simplicity, we consider a model with only two source pop-

ulations. Furthermore, the first source population only contributes migrants in the founding generation, T. The second source population contributes migrants in the founding generation and possibly in one or more generations thereafter. In generation i, for $i = T-1, \ldots, 0$ (before the present), a fraction m_i of the admixed population is replaced by individuals from the second source population.

4.3 Linkage Disequilibrium and Local Ancestry

ROLLOFF and ALDER use the standard two-locus measure of LD between a SNP at positions x and another SNP at position y, which is a genetic distance d to the right,

$$D_2(d) = \mathrm{cov}(H_x, H_y), \tag{4.1}$$

where H_x and H_y represent the haplotype or genotypes of an admixed chromosome at positions x and y. In the case of haplotype data, $H_{i,x} = 1$ if the i^{th} sample is carying the derived allele at the SNP at position x, or is 0 otherwise. Alternatively, for genotype data, $H_{i,x}$ take on values from $\{0, 1/2, 1\}$ depending on the number of copies of the derived allele the i^{th} sample is carrying at the SNP position x. We consider an additional site at position z, which is located a further genetic distance d' to the right of y. The three-loci LD, as defined by as defined by Bennett (1952) and Slatkin (1972), is given by

$$D_3(d, d') = \mathrm{cov}(H_x, H_y, H_z) = \mathbb{E}[(H_x - \mathbb{E}H_x)(H_y - \mathbb{E}H_y)(H_z - \mathbb{E}H_z)]. \tag{4.2}$$

The LD in an admixed population depends on the genetic differentiation between the source populations and and its admixture history. Let A_x represent the local ancestry at position x, with $A_x = 1$ if x is inherited from an ancestor in the first source population, and $A_x = 0$ if x is inherited from the second source population. We can compute the expectation of D_3 in terms of the three-point covariance function of A_x and so seperate out the effects of allele frequencies and local ancestry. We make the assumption that the alleles in the source populations are independent, so that

$$\mathrm{cov}\left(H_x, H_y, H_z\right) = \mathrm{cov}\left(\mathbb{E}[H_x|A_x], \mathbb{E}[H_y|A_y], \mathbb{E}[H_z|A_z]\right).$$

The background LD in unadmixed human populations decays to zero on a scale of tenths of centiMorgans, so this approximation is appropriate when d and d' are both larger than 0.5 cM. The conditional expectations above are the allele frequencies at each site in the admixed population, conditional on the local ancestry. These are given by $\mathbb{E}[H_x|A_x] = F_x + \delta A_x$, where F_x is the allele frequency of locus x in the first source population and δ_x is the difference of the allele frequencies of locus x in the two source populations. Equation 4.2 becomes

$$\begin{aligned} D_3(d, d') &= \mathrm{cov}\left(f_x + \delta_x A_x, f_y + \delta_y A_y, f_z + \delta_z A_z\right) \\ &= \delta_x \delta_y \delta_z \mathrm{cov}(A_x, A_y, A_z). \end{aligned} \tag{4.3}$$

A similar argument shows that $D_2(d)$ is proportional to the two-point covariance function of the local ancestry.

4.3.1 Local Ancestry Covariance Functions

If we take genetic drift into account, the three-point covariance function is random. To compute its expectation, we multiply out the covariance in equation 4.2 to get

$$\mathbb{E}[\mathrm{cov}(A_x, A_y, A_z)] = \mathbb{E}[A_x A_y A_z] - \mathbb{E}[A_x A_y]\mathbb{E}[A_z] - \mathbb{E}[A_x A_z]\mathbb{E}[A_y] - \mathbb{E}[A_y A_z]\mathbb{E}[A_x] + 2\mathbb{E}[A_x]\mathbb{E}[A_y]\mathbb{E}[A_z].$$

Each one of these expectations on the right-hand side is the probability that one or more sites is inherited from an ancestor from first source population. We organize these products of probabilities in a column vector:

$$\mathbf{v}_3 = \begin{pmatrix} \mathbb{P}\{A_x = A_y = A_z = 1\} \\ \mathbb{P}\{A_y = A_z = 0\}\mathbb{P}\{A_x = 0\} \\ \mathbb{P}\{A_x = A_z = 0\}\mathbb{P}\{A_y = 0\} \\ \mathbb{P}\{A_x = A_y = 0\}\mathbb{P}\{A_z = 0\} \\ \mathbb{P}\{A_x = 0\}\mathbb{P}\{A_y = 0\}\mathbb{P}\{A_z = 0\} \end{pmatrix},$$

so that $\mathrm{cov}(A_x, A_y, A_z) = (1, -1, -1, -1, 2)\mathbf{v}_3$. There is one entry in $\mathbf{v}_3$ for each of the five ways in which the three markers at positions x, y, and z can arranged on one or more chromosomes. In the founding generation T, this column vector is given by $\mathbf{v}_{3(T)} = (1 - m_T, (1-m_T)^2, (1-m_T)^2, (1-m_T)^2, (1-m_T)^3)'$. The probabilities for subsequent generations can be found by left-multiplying drift, recombination, and migration matrices:

$$\mathbf{v}_{3(i)} = \mathbf{D}_i \mathbf{L} \mathbf{U} \mathbf{v}_{3(i-1)},$$

The matrics $\mathbf{D}_i$, $\mathbf{L}$, and $\mathbf{U}$ account for the effects of migration, drift, and recombination, respectively. The migration matrix is a diagonal matrix given by

$$\mathbf{D}_i = \mathrm{diag}(1 - m_i, (1 - m_i)^2, (1 - m_i)^2, (1 - m_i)^2, (1 - m_i)^3).$$

Its entries are the probabilities that one, two, or three chromosomes in the admixed population will not be replaced by chromosomes from the second source population in generation i. The lower triangular drift matrix

$$\mathbf{L} = \frac{1}{4N^2} \begin{pmatrix} 4N^2 & 0 & 0 & 0 & 0 \\ 2N & 2N-1 & 0 & 0 & 0 \\ 2N & 0 & 2N-1 & 0 & 0 \\ 2N & 0 & 0 & 2N-1 & 0 \\ 1 & 2N-1 & 2N-1 & 2N-1 & (2N-1)(2N-2) \end{pmatrix}$$

gives the standard Wright-Fisher drift transition probabilities between the states as a function of the population size $2N$. Finally, the upper triangular recombination matrix is determined by the recombination rates between the three sites:

$$\mathbf{U} = \begin{pmatrix} e^{-d-d'} & (1-e^{-d})e^{d'} & (1-e^{-d})(1-e^{-d'}) & e^{-d}(1-e^{-d'}) & 0 \\ 0 & e^{-d'} & 0 & 0 & 1-e^{-d'} \\ 0 & 0 & 1-e^{-d}-e^{-d'}+2e^{-d-d'} & 0 & e^{-d}+e^{-d'}-2e^{-d-d'} \\ 0 & 0 & 0 & e^{-d} & 1-e^{-d} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The covariance function is then given by

$$\mathbb{E}[\text{cov}(A_x, A_y, A_z)] = (1, -1, -1, -1, 2) \left(\prod_{i=0}^{T-1} \mathbf{D}_i \mathbf{L} \mathbf{U} \right) \mathbf{v}_{3(0)}. \tag{4.4}$$

We can obain an analogous equation for $\text{cov}(A_x, A_y)$, involving the migration, drift, and recombination matrices for two loci:

$$\mathbb{E}[\text{cov}(A_x, A_y)] = (1, -1) \left(\prod_{i=0}^{T-1} \mathbf{D}_i \mathbf{L} \mathbf{U} \right) \mathbf{v}_{2(0)}.$$

In some cases, equation 4.4 simplifies further. In a one-pulse migration model, in which $m_T = M$ and is there after 0, the $\mathbf{D}_i$'s become identity matrices, and we get the closed from expression

$$\mathbb{E}[\text{cov}(A_x, A_y, A_z)] = M(1-M)(1-2M) \left(1 - \frac{1}{2N}\right)^T \left(1 - \frac{2}{2N}\right)^T e^{-T(d+d')}.$$

This is because $(1, -1, -1, -1, 2)$ is a left eigenvector of both $\mathbf{L}$ and $\mathbf{U}$, with corresponding eigenvectors $(1-1/2N)(1-2/2N)$ and $\exp(-d-d')$. Note that when $M = 0$, the covariance function will be identically 0. Another case is a two pulse model in which we ingore the effects of genetic drift. In this model, admixture only occurs T and T_2 generations before the present, so that $m_T = M_1, m_{T'} = M_2$, and all other m_i's are 0. Making the substitution $T_1 = T - T_2$, the right hand side of equation 4.4 becomes

$$\begin{aligned}(1-M_1)(1-M_2)e^{-T_2(d+d')} \Big[& M_2(1-M_1)^2 - 2M_2^2(1-M_1)^2 + M_1(1-2M_1)e^{-T_1(d+d')} \\ & -M_1M_2(1-M_1)\left(e^{-M_1 d} + e^{-M_1 d'} + \left(1 - e^{-d} - e^{-d'} + 2e^{-d-d'}\right)^{T_1}\right)\Big].\end{aligned} \tag{4.5}$$

The corresponding expression for the two-point covariance function is given by

$$(1-M_1)(1-M_2)e^{-T_2 d}\left(M_2 - M_1M_2 + M_1 e^{-T_1 d}\right), \tag{4.6}$$

which is a mixture of two exponentials. The relative complexity of equation 4.5 is actually a feature, as it makes detecting the presence of the second pulse of admixture easier.

4.4 Weighted Linkage Disequilibrium

As Loh et al. (2013) noted, we cannot use the LD in the admixed population directly, because the allele frequency differences in the source populations can be of either sign. However, if we compute expectation of the product of the LD with the product of the allele frequency differences, using equation 4.3 we obtain

$$\mathbb{E}\left[\delta_x \delta_y \delta_z D_3(d, d')\right] = \mathbb{E}[\delta_x^2 \delta_y^2 \delta_z^2] \mathbb{E}[\text{cov}(A_x, A_y, A_z)],$$

because the local ancestry in the admixed sample is independent of the allele frequencies in the admixed population. This expectation of the weighted LD is non-zero, and can be estimated by aggregating over triples of SNPs which are seperated by distances of approximately d and d'. The LD term can be estimated from the admixed population, while the δ's can be estimated from reference populations which are closely related to the two source populations.

We arrange the data from the admixed samples in an $n \times S_n$ matrix $\mathbf{H}$, where n is the number of admixed haplotypes/genotypes, and S_n is the number of segregating sites in the sample. For ease of notation, we assume that the positions are given in units which make the unit interval equal to the desired bin resolution.

For a given d and d' the set of SNP triples we use in the estimator for the weighted LD is

$$S[d, d'] = \{x, y, z : d \leq x - y < d + 1 \text{ and } d' \leq y - z < d' + 1\}.$$

Let w_x be the difference in the empirical allele frequences in two reference populations and let f_x be empirical allele frequency in the admixed population. An unbiased estimator of the weighted LD is

$$\hat{a}[d, d'] = \frac{1}{|S[d, d']|} \sum_{x,y,z \in S[d,d']} \frac{n \sum_{i=1}^{n} w_a w_b w_c (H_{i,x} - f_x)(H_{i,y} - f_y)(H_{i,z} - f_z)}{(n-1)(n-2)}.$$

4.5 Algorithm

Directly computing $\hat{a}[d, d']$ over the set $d, d' \in \{0, 1, \ldots, P\}^2$ would be cubic in the number of segregating sites, but as is the case with ALDER, we can use using a fast Fourier transform (FFT) to approximate $\hat{a}$, giving an algorithm whose run-time is instead linear in the number of segregating sites. We first rearrange $\hat{a}$ to get

$$\hat{a}[d, d'] = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n} \sum_{x,y,z \in S[d,d']} \delta_x \delta_y \delta_z (H_{i,x} - f_x)(H_{i,y} - f_y)(H_{i,z} - f_z)}{\sum_{x,y,z \in S[d,d']} 1},$$

and define sequences $b_i[d]$ and $c[d]$ by binning the data and then doubling the length by padding with P zeros,

$$b_i[d] = \begin{cases} \sum_{x: d \leq \lfloor x \rfloor < d+1} \delta_x (H_{i,x} - f_x) & : 0 \leq d < P \\ 0 & : P \leq d < 2P \end{cases}$$

$$c[d] = \begin{cases} |\{x : d \leq \lfloor x \rfloor < d + 1\}| & : 0 \leq d < P \\ 0 & : P \leq d < 2P \end{cases}$$

We can approximate $|S[d, d']|$ and the n sums in the numberator of $\hat{a}[d, d']$ in terms of convolutions of these sequences:

$$|S[d, d']| \approx \sum_{w=0}^{P} c[w] c[w + d] c[w + d + d']$$

$$\sum_{x,y,z \in S[d,d']} \delta_x \delta_y \delta_z (H_{i,x} - f_x)(H_{i,y} - f_y)(H_{i,z} - f_z) \approx \sum_{w=0}^{P} b_i[w] b_i[w+d] b_i[w+d+d'].$$

These convolutions can be efficiently computed with an FFT, since under a two-dimensional discrete Fourier transform,

$$\sum_{w=0}^{P} b_i[w] b_i[w+d] b_i[w+d+d'] \leftrightarrow B_i[j] \bar{B}_i[k] B_i[k-j],$$

where B_i is the (one-dimensional) discrete Fourier transform of b, and $B_i[-j]$ is the j^{th} to last most element of B_i. Summing over i and taking the inverse discrete Fourier transform, we can approximate the discrete Fourier transform of numerator of $\hat{a}$. We use the same method applied to c to approximate the denominator of $\hat{a}$. Because the number of bins is generally much less than the number of segregating sites, the rate-limiting step of this algorithm is the binning step to form c and the b_i's, which is $O(S_n)$, rather than the FFTs, which are $O(P^2 \log(P))$.

When using only the admixed population itself as a reference population, the method described above will be biased if the same samples are used to estimate both the linkage disequilibria and the weights. We cannot efficiently compute a polyache statistics like Loh et al. (2013). At the cost of some power, we instead adopt the approach of Pickrell & Pritchard (2012) and separate the admixed population into two equal-sized groups. We then use one group to estimate the weights, and the other group to estimate linkage disequilibrium, and vice versa. This gives gives two unbiased estimates for the numerator of $\hat{a}$, which we then average.

4.5.1 Fitting the Two-Pulse Model

We fit equation 4.6 to the estimates of the weighted LD using non-linear least squares, with two modifications. We added a proportionality constant to account for the expected square allele frequency difference between the source populations. We also subtracted out an affine term in the weighted LD which is due to population substructure (Loh et al. 2013). We estimated this by computing the three-way covariance between triples of chromosomes. We use the jackknife to obtain confidence intervals for the resulting estimates by leaving out each chromosome in turn and refitting on the data for the remaining chromosomes.

4.6 Simulations

We used the program macs Chen et al. (2009) to generate two source populations which diverged 4000 generations ago and a coalescent simulation to generate an admixed population from the two source populations according to two-pulse and constant admixture models. We sampled 50 diploid individuals from the admixed and two source populations, each consisting of 20 chromosomes of length 1 Morgan. The effection population size was $2N = 1000$ for

the admixed population and two source populations. Using a two pulse model, we varied the migration probabilities and timings for each pulse to examine the accuracy of equation 4.6. We also simulated data for a model with a constant rate of admixture each generation, and compared this to the predictions made by equation 4.4.

4.7 Data Set

We computed the weighted LD for the Mexican and Columbian populations in the first phase of the 1000 Genomes data set. These consisted of 66 individuals from Los Angeles and 60 individuals from Medellin, respectively. We used the 88 Yoruba samples as the one reference population. We computed the weighted LD on the genotypes to avoid effects of phasing errors.

4.8 Discussion

4.8.1 Simulations

We find there is a generally a close match between our equations and the simulated data under both under two pulse admixture scenarios (figures 4.1 and 4.2) and constant admixture scenarios (figure 4.3). The exception is when the total admixture proportion $M_2 + M_1(1 - M_2)$ is close to 0.5. As the total admixture proportion increases above 0.5, the contours for equation 4.2 flip from being concave down to concave up. This transition can be seen by comparing the upper left side of figre 4.2 to its lower right. At this threshold, the contours of the estimated weighted LD depend on the actual admixture fractions of the samples, which may differ from the expectation as a result of genetic drift. This mismatch between theory and simulations is most evident in figure 4.2, for $m_1 = 0.1, m_2 = 0.4$ and $m_1 = 0.2, m_2 = 0.4$.

When there is continous admixture scenario, the shape of the weighted LD surface depends on both the duration and total amount of admixture. When the duration is short, the weighted LD surfaces are indistinguishable from teh weighted LD surfaces produced by one pulse of migration. As the duration increases, the contours of teh weighted LD surface become more curved. The contours are convcave up when the total proportion is greater than 50% and concave down when it is less. When the total proportion is exactly 50%, the amplitude of the weighted LD surface is much smaller than the smapling error.

For two pulse models, the effects of the second pulse of migration only become evident when temporal spacing between the pulses is large enough ($T1 > T2$). Otherwise, the resulting weighted LD surface cannot be distingiushed from the weighted LD surface produced by one pulse of admixture. As in the case of continuous admxiture the concavity of the surface contours is determined by the total admixture proportion.

These qualitative about the similarity between one pulse and two puls admixture scenarios are borne out by simulations of the estimation error, shown in figure 4.4. When the spacing

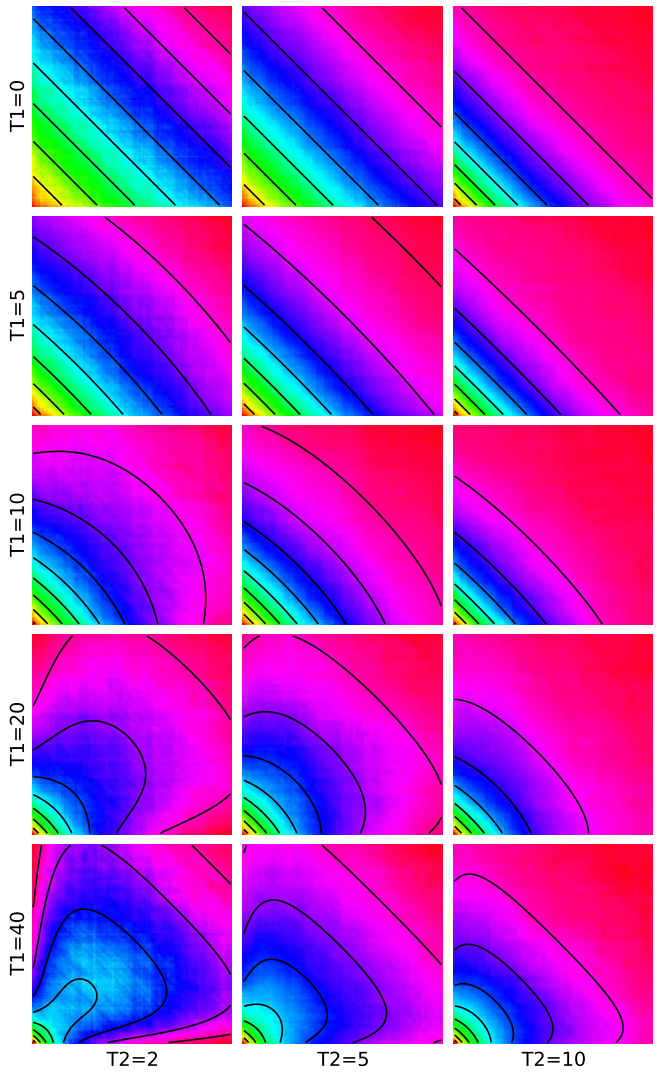

Figure 4.1: **Predicted weighted LD surfaces from simulations and theory for varying admixture times.** The heat maps are from simulations and the contours are plotted from equation 4.2. The two admixture probabilities were fixed at $m_1 = m_2 = .2$ and the the times of the two admixture pulses, T_1 and T_2, were varied. Each square covers the range 0.5 cM $< d, d' <$ 20 cM. When time of the more recent pulse is greater than half of that of the more ancient pulse, i.e. $2T_1 > T_1 + T_2$, the contours of the resulting weighted LD surface are straight, making it difficult to distinguish from the weighted LD surface produced by a one-pulse admixture scenario.

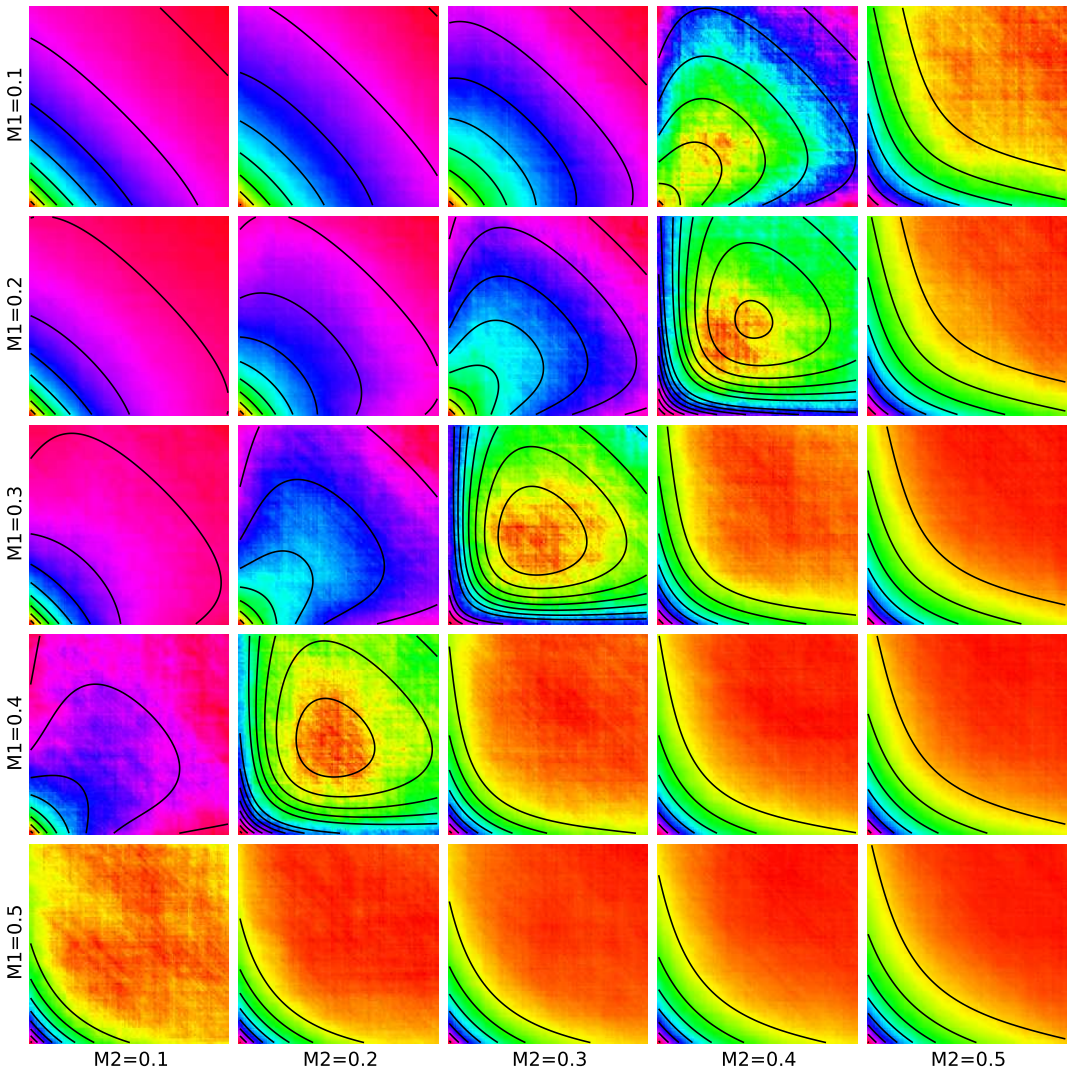

Figure 4.2: **Predicted weighted LD surfaces from simulations and theory.** The heat maps are from simulations and the contours are plotted from equation 4.2. The two admixture times were fixed at 2 and 12 generaations ago ($T_1 = 10$ and $T_2 = 2$) while the admixture probabilities were varied. Each square covers the range 0.5 cM $< d, d' <$ 20 cM. As the total admixture proportion $m_2 + m_1(1 - m_2)$ increases above 0.5, the concavity of the contours flips. Weighted LD surfaces for $m_1 > 0.5$ or $m_2 > 0.5$ are not shown, but are qualitatively similar to the surfaces on the lower and rightmost sides.

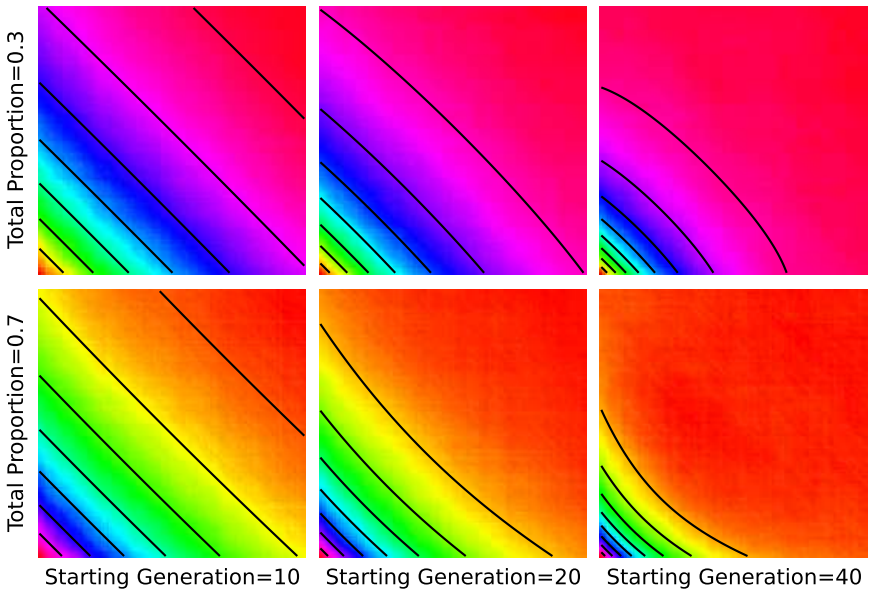

Figure 4.3: **Weighted LD surfaces produced by constant admixture.** The heatmaps are from simulations and the contours are from equation XX. In all six plots, admixture stopped 5 generations before the present. Each square covers the range 0.5 cM $< d, d' <$ 20 cM. We varied the time of the beginning of the admixture and the total admixture probability. The admixture probability for each generation was constant, and chosen so that the total admixture proportion was either 0.3 or 0.7. When the admixture is spread over 5 generations (the leftmost column), the resulting weighted LD surface is similiar to a one-pulse weighted LD surface. For longer durations, the weighted LD surfaces are similar to those produced by two pulses of admixture.

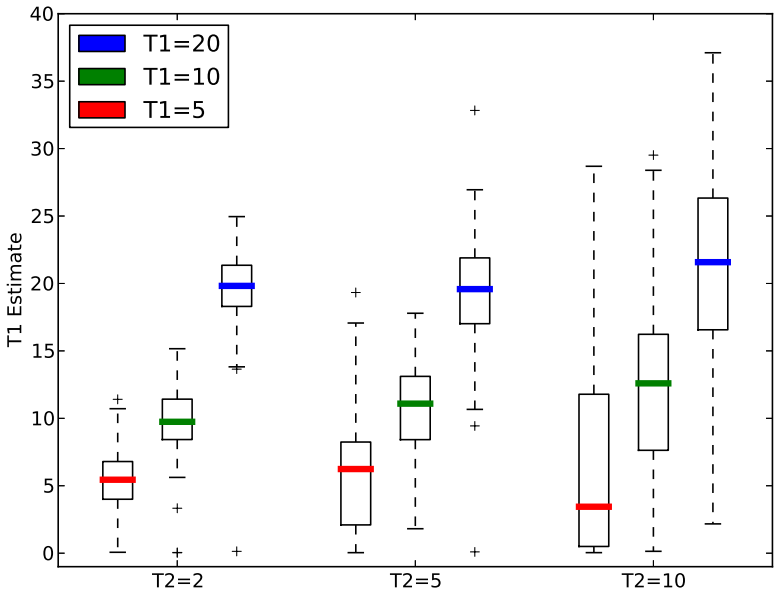

Figure 4.4: **Accuracy of estimates of T_1 as a function of other parameters.** Nine admixture scenarios, $T_1 \in \{5, 10, 20\}$ and $T_2 \in \{2, 5, 10\}$, were simulated 100 times each. The admixture probabalities were fixed at $M_1 = 0.3$ and $M_2 = 0.2$. The colored bars give the medians of estimates for each of these nine cases, the boxes delimit the interquartile range, and the whiskers extend out to 1.5 times the interquartile range. As the time between the two pulses of admixture increases, the error in the estimates decreases. Consistent with the simulations shown in figure 4.1, there is limited power to estimate the time of the more ancient admixture pulse when $T_2 > T_1$.

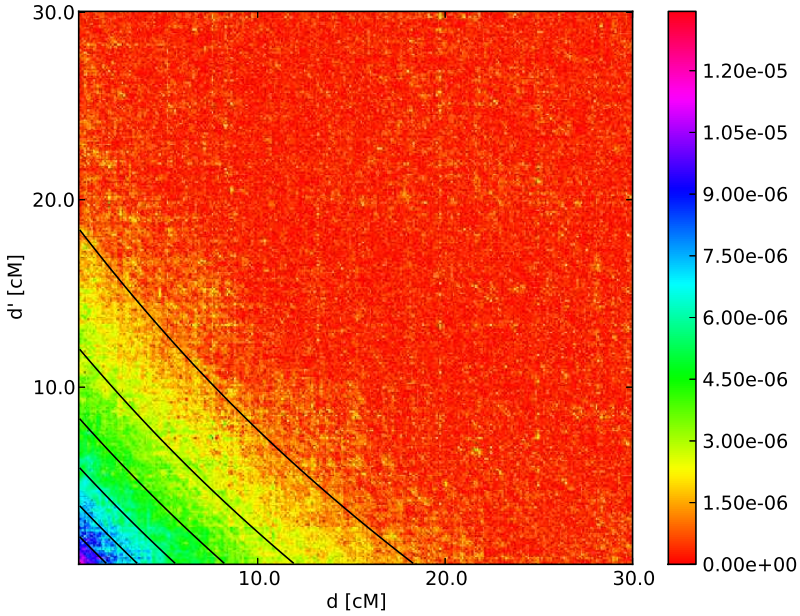

Figure 4.5: **Weighted LD surface for Mexican samples with Yoruba as reference.** The model with the best fit is two pulses from the non-Yoruba source population at $T_1 + T_2 = 12.3 \pm 3.3$ and $T_2 = 9.9 \pm 2.7$ generations ago. The jackknife confidence intervals for the times of these two pulses overlap.

between the two pulses is small relative to their age, the median of the estimates of the timing of the second pulse is close to the true value, but the interquartile range is large. Moreover, the best fit often lies on a boundary of the parameter space which is equivalent to a one pulse admixture model. When the spacing between the pulses is larger, the estimates for the timing of the older pulse before more precise.

4.8.2 1000 Genomes

Gravel et al. (2013) have previously analyzed the 1000 Genomes data that we computed weighted LD surfaces for. For the Mexican samples, they found a small but consistent amount of African ancestry, which appeared in the population 15 generations ago, with continuing contributions from European and Native American populations since that date, but no African migration. In fitting a two-pulse model to the Mexican weighted LD surface (figure 4.5), we estimated that the two pulses occured 12.3±3.3 and 9.9±2.7 generations ago. These confidence intervals overlap, and so we cannot reject a one-pulse admixture history. This is not quite consistent with the constant migration model that Gravel et al. (2013) found, but as we have seen from simulations, it is hard to distinguish a constant migration model from a one-pulse model when the duration of the migration is short.

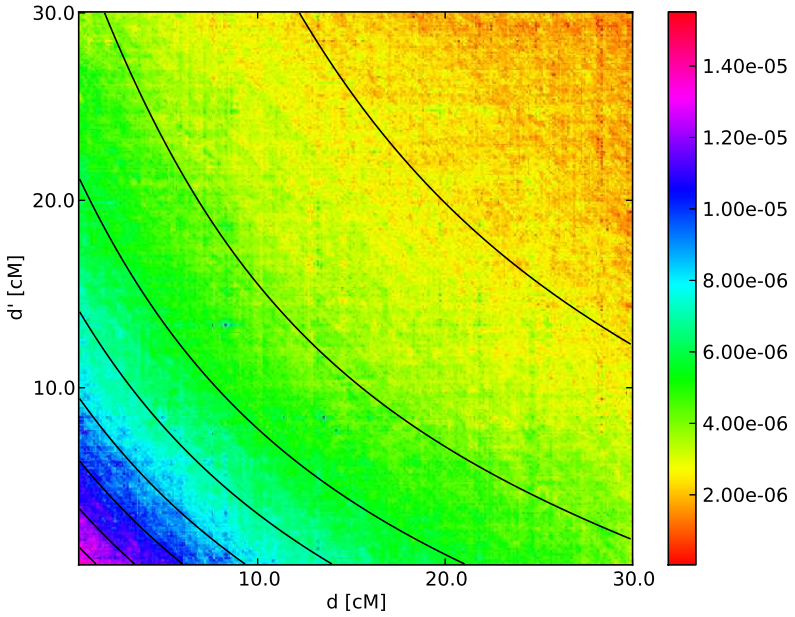

Figure 4.6: **Weighted LD surface for Columbian samples with Yoruba as reference** The two-pulse model that fits best is two pulses of non-Yoruba admixture at $T_1 + T_2 = 11.8 \pm 1.2$ and $T_2 = 2.64 \pm 0.08$ generations ago. The jackknife confidence intervals for the times of these two pulses do not overlap. The amplitude of this weighted LD surface is approximately ten times larger than that of the Mexican samples. This a result of larger proportion of Yoruba ancestry in the Columbian samples.

The weighted LD surface for the Columbian samples is shown in figure 4.6. From this, we estimated two pulses of non-Yoruba migration at 11.8 ± 1.2 and 2.64 ± 0.08 generations before the present. Gravel et al. (2013) also inferred two pulses of admixture, corresponding to 3 and 9 generations ago. The weighted LD surface of the Columbian samples has contours which are strongly concave up, in constrast to those of the Mexican samples.

4.8.3 Comparison to Existing Methods

Compared to existing weighted LD methods, our our method uses more information in the data because it compares triples of SNPs instead of pairs. This gives our method the ability to infer admixture histories more complex than a one-pulse model. However, this comes at the price of greater estimation variances. ALDER and ROLLOFF can make estimates from just tens of samples, while our method requires hundreds of samples. Part of this difference can be attributed to the fact that ALDER and ROLLOFF make inferences over a smaller class of models, but the main reason arises from the fact that the existing two models are estimating second moments of the data, while we are estimating third moments. The variance of these estimates are both inversely proportional to the sample size, but the constants for estimating third moments are larger. As data becomes more readily avaliable, this disadvantage should disappear.

Bibliography

3 Consortium, I. H., et al. 2010, Nature, 467, 52
Alexander, D. H., Novembre, J., & Lange, K. 2009, Genome Research, 19, 1655
Baird, S. J., Barton, N. H., & Etheridge, A. M. 2003, Theoretical Population Biology, 64, 451
Ball, F., & Stefanov, V. T. 2005, Mathematical Biosciences, 196, 215
Baran, Y., Pasaniuc, B., Sankararaman, S., et al. 2012, Bioinformatics, 28, 1359
Barton, N. H., & Bengtsson, B. O. 1986, Heredity, 57, 357
Barton, N. H., & Etheridge, A. M. 2011, Genetics, 188, 953
Bennett, J. 1952, Annals of Eugenics, 17, 311
Bickeböller, H., & Thompson, E. A. 1996a, Theoretical Population Biology, 50, 66
—. 1996b, Genetics, 143, 1043
Brisbin, A., Bryc, K., Byrnes, J., et al. 2012, Human Biology, 84, 343
Bryc, K., Auton, A., Nelson, M. R., et al. 2010, Proceedings of the National Academy of Sciences, 107, 786
Cannings, C. 2003, Human heredity, 56, 126
Chapman, N. H., & Thompson, E. A. 2002, Genetics, 162, 449
Chen, G. K., Marjoram, P., & Wall, J. D. 2009, Genome research, 19, 136
Dimitropoulou, P., & Cannings, C. 2003, Bioinformatics, 19, 790
Donnelly, K. P. 1983, Theoretical Population Biology, 23, 34
Falush, D., Stephens, M., & Pritchard, J. K. 2003, Genetics, 164, 1567
Fisher, R. A. 1949, The Theory of Inbreeding (Edinburgh, Scotland: Oliver and Boyd)
Gravel, S. 2012, Genetics, 191, 607
Gravel, S., Zakharia, F., Moreno-Estrada, A., et al. 2013, PLoS genetics, 9, e1004023
Griffiths, R. C., & Marjoram, P. 1996, Journal of Computational Biology, 3, 479
Guo, S.-W. 1994, American Journal of Human Genetics, 54, 1104
Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. 2009, PLoS Genetics, 5, e1000695
Henn, B. M., Botigué, L. R., Gravel, S., et al. 2012, PLoS Genetics, 8, e1002397
Hey, J. 2010, Molecular biology and evolution, 27, 905
Hey, J., & Nielsen, R. 2004, Genetics, 167, 747
Hill, W. G. 1974, Theoretical Population Biology, 5, 366
Hoggart, C. J., Parra, E. J., Shriver, M. D., et al. 2003, The American Journal of Human

Genetics, 72, 1492
Hudson, R. R. 1983, Theoretical Population Biology, 23, 183
—. 2002, Bioinformatics, 18, 337
Li, N., & Stephens, M. 2003, Genetics, 165, 2213
Liang, M., & Nielsen, R. 2014a, Genetics, genetics
—. 2014b, bioRxiv, 008078
Loh, P.-R., Lipson, M., Patterson, N., et al. 2013, Genetics, 193, 1233
Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. 2013, The American Journal of Human Genetics, 93, 278
Marjoram, P., & Wall, J. 2006, BMC Genetics, 7, 16
Martin, O. C., & Hospital, F. 2011, Genetics, 189, 645
McVean, G. A., & Cardin, N. J. 2005, Philosophical Transactions of the Royal Society B: Biological Sciences, 360, 1387
Menotti-Raymond, M., David, V. A., Pflueger, S. M., et al. 2008, Genomics, 91, 1
Moorjani, P., Patterson, N., Hirschhorn, J. N., et al. 2011, PLoS genetics, 7, e1001373
Moorjani, P., Thangaraj, K., Patterson, N., et al. 2013, The American Journal of Human Genetics, 93, 422
Moreno-Mayar, J. V., Rasmussen, S., Seguin-Orlando, A., et al. 2014, Current Biology
Paşaniuc, B., Sankararaman, S., Kimmel, G., & Halperin, E. 2009, Bioinformatics, 25, i213
Parra, E. J., Marcini, A., Akey, J., et al. 1998, The American Journal of Human Genetics, 63, 1839
Pickrell, J. K., & Pritchard, J. K. 2012, PLoS genetics, 8, e1002967
Pool, J. E., & Nielsen, R. 2009, Genetics, 181, 711
Price, A. L., Patterson, N. J., Plenge, R. M., et al. 2006, Nature Genetics, 38, 904
Price, A. L., Tandon, A., Patterson, N., et al. 2009, PLoS Genetics, 5, e1000519
Pritchard, J. K., Stephens, M., & Donnelly, P. 2000, Genetics, 155, 945
Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. 2009, Nature, 461, 489
Reich, D., Patterson, N., De Jager, P. L., et al. 2005, Nature Genetics, 37, 1113
Rodolphe, F., Martin, J., & Della-Chiesa, E. 2008, Theoretical Population Biology, 73, 289
Rosenberg, N. A., Pritchard, J. K., Weber, J. L., et al. 2002, Science, 298, 2381
Sankararaman, S., Kimmel, G., Halperin, E., & Jordan, M. I. 2008, Genome Research, 18, 668
Slatkin, M. 1972, Genetics, 72, 157
Smith, M. W., Patterson, N., Lautenberger, J. A., et al. 2004, The American Journal of Human Genetics, 74, 1001
Stam, P. 1980, Genetics Research, 35, 131
Stefanov, V. T. 2000, Genetics, 156, 1403
Sundquist, A., Fratkin, E., Do, C. B., & Batzoglou, S. 2008, Genome Research, 18, 676
Tang, H., Choudhry, S., Mei, R., et al. 2007, The American Journal of Human Genetics, 81, 626
Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. 2006, The American Journal of Human Genetics, 79, 1

Tang, H., Peng, J., Wang, P., & Risch, N. J. 2005, Genetic epidemiology, 28, 289
Verdu, P., & Rosenberg, N. A. 2011, Genetics, 189, 1413
Wakeley, J., King, L., Low, B. S., & Ramachandran, S. 2012, Genetics, 190, 1433
Walters, K., & Cannings, C. 2005, Theoretical Population Biology, 68, 55
Wiuf, C., & Hein, J. 1999, Theoretical Population Biology, 55, 248
Zhang, B., Li, M., Zhang, Z., et al. 2007, Molecular biology and evolution, 24, 1801