UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Inference of population history and mutation biology from human genetic variation

Permalink
https://escholarship.org/uc/item/5w42n2s6

Author
Harris, Kelley

Publication Date
2015

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Inference of population history and mutation biology from human genetic variation

By

Kelley Harris

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Applied Mathematics

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Rasmus Nielsen, Co-Chair
Steven N. Evans, Co-Chair
Yun S. Song
Lior Pachter

Spring 2015

Inference of population history and mutation biology from human genetic variation

Copyright 2015
by
Kelley Harris

Abstract

Inference of population history and mutation biology from human genetic variation

by

Kelley Harris

Doctor of Philosophy in Applied Mathematics

and the Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Rasmus Nielsen, Co-Chair
Steven N. Evans, Co-Chair

Human genetic diversity bears many imprints of our species' migration out of Africa. When our ancestral population expanded across the globe and settled into nearly every habitable environment, genetic diversity was reduced, distributed and redistributed by an intricate series of population bottlenecks and migrations. The coalescent process is a mathematical model that describes the probability distribution of modern DNA sequences that could have evolved from a common ancestor following a specific demographic scenario; in principle, this gives us the means to infer the history that is most consistent with large datasets of genomes sequenced from living people. In practice, however, the coalescent is such a complicated model that such calculations are intractable to perform exactly. Here, I introduce several new techniques for performing approximate demographic inference under the coalescent; these operate by condensing samples of genomes into more compact summary statistics and then mathematically approximating the probability distributions that these statistics should follow. I then use these techniques to infer joint demographic histories from European and African genomes, describing a complex out-of-Africa migration that involved multiple population size changes as well as a long period of migration between the diverging continental groups. The good match between predicted and observed genomic samples indicates that the coalescent is a useful framework for describing the evolution of humans; however, I also note systematic discrepancies between the model and the data. In the last two chapters of this thesis, I go on to show that some deviations of human data from coalescent predictions stem from the coalescent's oversimplication of the way mutations are generated. One standard assumption is that mutations occur independently; in contrast, at least 2% of human occur in linked clusters and are likely to have been generated by multinucleotide mutation events (MNMs). Examining the derived and ancestral alleles of these MNMs, I show that they are enriched for transversions and that many bear the specific signature of error-prone

Polymerase ζ. A second assumption I show to be violated is that the mutation rate has not changed over time and is constant across populations. I show this indirectly by demonstrating that C→T transitions, particularly in the context TCC→TTC, are more frequent in Europeans than in other populations. Although it is not clear whether this mutation rate change was functionally significant or driven by selection, it demonstrates that the process of genome evolution has not stayed constant during recent human history, but has been regionally differentiated by the forces that shape our primary genome sequences.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis could not have come to be without inspiration and guidance from Rasmus Nielsen and Yun Song, the research advisors who helped me develop into a scientist. I hope that everything presented here and all my work henceforth will live up to what they've taught me about pursuing important unknowns while maintaining rigor, clarity, and integrity in all facets of statistics and science. Rasmus contributed as senior author to the published versions of chapters 2 and 5, while Yun was senior author on chapters 3 and 4.

I am also indebted to the rich community of computational and evolutionary biologists who help make Berkeley one of the most intellectually exciting places in the world. Sara Sheehan deserves a special thank you for being my co-author as well as my friend; it was she and not I who was down in the trenches writing and deploying the software described in chapters 3 and 4, and we worked together through all stages of the theory and exposition of these studies. We shared co-first-authorship of the paper that became chapter 3. Jack Kamm also contributed as a co-author to chapter 4, helping translate our idea into a useful and somewhat clear end product. My M.Phil. advisor Richard Durbin also deserves credit for the theoretical ideas about identity by descent and identity by state that eventually became the backbone of Chapter 2. Richard was really the person who inspired me to pursue population genetics; my year under his mentorship and our ongoing interactions had a significant impact on my Ph.D. work.

Day to day and hour to hour, my science has been shaped by the many officemates and honorary officemates who've adopted me as a math-refugee into the Integrative Biology Department and stood by me over the years, including Melinda Yang, Josh Schraiber, Fernando Racimo, Mason Liang, Jacob Vestergaard, Emilia Huerta-Sanchez, Rori Rolphs, Jacob Crawford, Mehmet Somel, Malte Thodberg, Rebekah Rogers, Flora Jay, Mike Martin, Vincent Appel, Amy Ko, Eline Lorenzen, and Tyler Linderoth. Up the hill in Yun's sector of the Computer Science department, I've also treasured the sometimes-geekier companionship of Sara and Jack as well as Matthias Steinrücken, Josh Paul, Paul Jenkins, Ma'ayan Bresler, Anand Bhaskar, Andrew Chan, Shishi Luo, Geno Guerra, and Jeff Spence. I've learned innumerable things from conversations with these friends and colleagues, as well as other grad students and postdocs at Berkeley and elsewhere.

Through grad school and my entire life, my family have offered me constant support and love. My parents, Anne Katten and Glenn Harris, have always gone the extra mile to help me achieve my goals, embracing however eccentric my goals and I turned out to be.

This dissertation was typeset using the ucastrothesis LaTeX template.

Chapter 1

Introduction

In 2010, the completion of the 1000 Genomes Pilot Project ushered in new era of human genomics where high-quality human genome sequences are being generated from enclaves of human genetic diversity located all over the world (1000 Genomes Project 2010). Five years later, publicly available human genome sequences number in the thousands, and rare genetic variants have been ascertained from tens if not hundreds of distinct human populations (1000 Genomes Project 2012). High quality genomes have even been recovered from two extinct hominid species, the Neanderthals and the related but distinct Denisovans (Meyer et al. 2012; Prüfer et al. 2014).

This rapidly expanding storehouse of genomes has already provided a wealth of knowledge about human history and evolution. However, there are a great many questions about human evolution that genomic data has not yet been able to answer. Some of these questions may be answered by sequencing larger, more diverse datasets of living humans or by more opportune finds of hominid fossils containing ancient DNA. However, other questions can only be answered by developing more sophisticated theory and methodology for analyzing the genomic data we already have.

One powerful strategy for extracting information from genetic data is to mathematically model the process of evolution. A particularly useful mathematical framework is the coalescent, a stochastic process that describes a probability distribution on the ancestors of a DNA sample. In its simplest form, the coalescent is dual to the continuous-time limit of the neutral Wright-Fisher model, which describes genetic drift in an asexual population that maintains a constant size of N individuals (Kingman 1982). In addition, the coalescent model has been expanded to include sexual recombination, non-reciprocal gene conversion, natural selection, changes in effective population size, and migration among demes that maintain partial reproductive isolation. Recombination and gene conversion are modeled by specifying that each new individual is the offspring of two parents. The offspring's genome is generated as a mosaic of the two parent genomes, with randomly placed recombination events stitching together parentally-derived tracts into a new composite genome.

This additional complexity equips the coalescent to describe human evolution very accurately. Conversely, large genetic samples can provide the power necessary to test which types of complexity are most necessary for describing human diversity well. In principle, one can test whether a coalescent model including both population size changes and migration between demes has a higher likelihood of producing a given genomic dataset than a subtly different coalescent model including population size changes alone. This model testing approach allows many historical claims to be tested using only genetic data from humans alive today.

In practice, coalescent likelihoods are very complex to compute, even when drift and recombination are the only processes being modeled. For this reason, approximations are crucial to the feasibility of coalescent-based inference. Two approximations that have been widely used over the past few years are the sequentially Markov coalescent (SMC) (McVean & Cardin 2005) and the closely related SMC' (Marjoram & Wall 2006). These approximations are motivated by the earlier work of Wiuf & Hein (1999), who introduced a new way of looking at the un-approximated coalescent with recombination. Wiuf & Hein (1999) developed an algorithm for simulating coalescent histories spatially rather than temporally, parsing a set of homologous chromosomes into a sequence of adjacent blocks that have each been inherited from an ancestral individual with no intervening recombination. They presented an algorithm for calculating the joint distribution of the length and time to most recent common ancestry (TMRCA) of nth block as a function of the lengths and TMRCAs of the $n-1$ blocks that precede it moving spatially from 5' to 3' along the chromosome. This spatial algorithm enables much more efficient simulation of DNA samples than can be done using forward-time models of a population evolving over many generations from past to present.

Given a chromosome alignment containing n blocks with distinct TMRCAs, McVean & Cardin (2005) had the insight that the sequence of these n TMRCAs is nearly distributed like a Markov process. Although the exact distribution of the nth TMRCA depends jointly on all $n-1$ TMRCAs that precede it in the sequence, McVean & Cardin (2005) showed that little accuracy was lost by sampling this nth TMRCA from a distribution that depended on the $(n-1)$-st TMRCA but was conditionally independent of all preceding $n-2$ TMRCAs. This suggested that Hidden Markov Model (HMM) methods, which are theoretically well-developed and computationally efficient, could be used for the purpose of evolutionary inference.

Between 2007 and 2011, two pioneering HMM-based methods were developed for the purpose of evolutionary inference. The first, coalHMM, was designed to analyze an alignment of three reference genomes representing humans, chimpanzees, and gorillas (Hobolth et al. 2007). Using this method, Hobolth et al. (2007) inferred that these three species diverged recently and quickly from an ancestral population with a fairly large effective population size. A second method pioneered by Li & Durbin (2011) charted changes in human effective population size over a more recent period of time ranging from about 1 million years to 20,000 years before the present. These population sizes are inferred from the spacing of heterozygous sites in a single diploid genome, which is informative about the sequence of hidden TMRCAs shared by the two underlying haplotypes. This method, the Pairwise Sequentially Markov Coalescent (PSMC), infers similar ancient population sizes from African and non-African genomes but infers an out-of-Africa population size bottleneck in Europeans and Asians between 20,000 and 60,000 years ago.

The first three chapters of this dissertation describe demographic inference methodology that is rooted in the Markovian coalescent, as are coalHMM and the PSMC, but overcomes some limitations of these earlier methods. One of these limitations is that neither PSMC nor coalHMM can jointly model population subdivision and effective population size changes. In Chapter 2, I describe a more versatile method that allows the user to specify a fixed number of effective population size changes, population splits, and migration pulses and then numerically optimize the model parameters to fit an input dataset. This additional complexity is accommodated by summarizing the input data more extensively than is done by coalHMM or PSMC, collapsing an alignment of two DNA sequences into a spectrum of distances between consecutive polymorphic sites, or tracts of identity by states (IBS).

Although summarizing a sequence alignment as a spectrum of IBS tract lengths reduces the computational cost of inferring a complex demographic history, it also comes with disadvantages. For example, this approach makes it difficult to integrate over DNA sequence observations in genomic regions of low sequencing quality. To address these limitations, Chapters 3 and 4 describe improvements to the more traditional hidden Markov model decoding approach that underlies coalHMM and the PSMC. These chapters describe my contributions to the method diCal (Demographic Inference using Composite Approximate Likelihood), which I jointly developed together with Yun S. Song, Joshua Paul, Paul Jenkins, Sara Sheehan, Matthias Steinrücken, and John. A. Kamm.

Although Chapters 2, 3, and 4 focus on the problem of inferring demographic history, there are broader insights that can be gained from the ability to compute approximate likelihoods of genetic datasets given evolutionary models. In particular, searching for the evolutionary model that optimally describes a dataset can reveal ways in which broad categories of evolutionary models fail to describe genomic data well. One shortcoming of the coalescent with recombination, as revealed and described in Chapter 2 is that it does not accurately predict the abundance of polymorphisms that occur very close together in the genome (i.e. fewer than 100 base pairs apart). Chapter 5 explores this pattern more fully, providing evidence that this discrepancy between the data and the coalescent model is a consequence of error-prone polymerase activity in the human germline. When high-fidelity DNA polymerases stall because of irregularities in DNA structure, it has been shown that error-prone polymerases restart replication by taking over from the high-fidelity polymerase over a short stretch of DNA, sometimes introducing closely spaced mutations before the high-fidelity polymerase takes over again.

In Chapter 5, we suggest that error-prone polymerase activity should be mathematically modeled as follows: two sites that are close together (say $\ell < 10$ base pairs apart) will mutate simultaneously in a given generation with probability $\mu_{\text{MNM},\ell}$ that is about 1000-fold smaller than μ, the rate of single point mutations per site per generation. In contrast, the standard coalescent models mutation as a process that is homogeneous in space, where the probability of simultaneous mutations at nearby sites is μ^2 per generation. Given that μ is on the order of 10^{-8} in humans, the two models predict very different abundances of closely spaced SNPs and the data strongly support the error-prone polymerase model.

In addition to assuming that mutations accumulate homogeneously in space, the coalescent model assumes that they accumulate homogeneously in time, with a constant rate of μ mutations per generation throughout the recent and ancient past. This assumption predicts that many different methods should be able to estimate μ accurately, some utilizing mutations that occurred very recently and others quantifying variation that is much older. In contrast to this, μ values estimated from the frequency of fixed differences between humans and chimpanzees are empirically almost 2.5-fold higher than μ values estimated from *de novo* mutations detected by sequencing parent-child trios (Nachman & Crowell 2000; 1000 Genomes Project 2010; Ségurel et al. 2014). One explanation for this trend is that the mutation rate may have slowed down over time along the human lineage (Li & Tanimura 1987; Scally & Durbin 2012).

It is intrinsically harder to prove that the mutations happen inhomogeneously over time than to prove that they occur inhomogeneously in space. Older mutations have been affected by more selection and drift than young mutations have, while young mutations are harder to ascertain because they occur at low population frequencies and can be confused with sequencing errors. These complicating forces might confound different mutation rate estimators in different ways, leading to erroneous conclusions that μ has changed over time. In Chapter 6, I describe a strategy for overcoming some of these challenges and conclude that the human mutation rate has recently changed to some extent. I show that a particular mutation type TCC→TTC, occurs at a higher rate in Europeans than in Asians or Africans, a finding that is most satisfactorily explained by a recent mutation rate increase in Europeans. The magnitude of this mutation rate change is slight compared to the magnitude of the proposed “hominoid slowdown" since we diverged from chimpanzees, but it provides conclusive evidence that the mutation rate is capable of rapid evolution.

Chapter 2

Inferring demographic history from a spectrum of shared haplotype lengths

Originally published as: K. Harris and R. Nielsen. (2013) “Inferring demographic history from a spectrum of shared haplotype lengths." *PLoS Genetics* 9 (6) e1003521.

There has been much recent excitement about the use of genetics to elucidate ancestral history and demography. Whole genome data from humans and other species are revealing complex stories of divergence and admixture that were left undiscovered by previous smaller data sets. Much of the interest focuses on the timing of past admixture or divergence events, for example the time at which Neanderthals exchanged genetic material with humans or the time at which modern humans left Africa. Here, we present a method for using sequence data to jointly estimate the timing and magnitude of past admixture events, along with population divergence times and changes in effective population size. We infer demography from a collection of pairwise sequence alignments by summarizing their length distribution of tracts of identity by state (IBS) and maximizing an analytic composite likelihood derived from a Markovian approximation to the coalescent. Recent gene flow between populations leaves behind long tracts of identity by descent (IBD), and these tracts give our method its power by influencing the distribution of shared IBS tracts. In simulated data, we accurately infer the timing and strength of admixture events, population size changes, and divergence times over a variety of ancient and recent time scales. Using the same technique, we analyze deeply sequenced trio parents from the 1000 Genomes project. The data show strong evidence of extensive gene flow between Africa and Europe after the time of divergence as well as substructure and gene flow among ancestral hominids. In particular, we infer that recent Africa-Europe gene flow and ancient ghost admixture into Europe are both necessary to explain the spectrum of IBS sharing in the trios, rejecting simpler models that contain only one of these features.

2.1 Introduction

Over the past several decades, population genetics has made key contributions to our understanding of human demography, as well as the demographic history of other species. Early studies that inferred haplotype trees of mitochondria and the Y chromosome (Slatkin & Madison 1989; Templeton 2002) changed our view of human origins by prompting wide acceptance of the out of Africa replacement hypothesis. Equally important were early methods that modeled the distribution of pairwise differences (Tajima 1983; Slatkin & Hudson 1991) and polymorphic sites (Wakeley & Hey 1997) in genetic samples, using this information to estimate historical population sizes and detect recent population growth. These methods revealed that a population bottleneck accompanied the human migration out of Africa; they have also shed light on recent population growth brought on by agriculture.

Advances in computational statistics have gradually made it possible to test more detailed hypotheses about demography. One advancement has been computing the coalescent likelihood of one or a few markers sampled across many organisms (Griffiths & Tavaré 1994a,b; Kuhner et al. 1995; Nielsen 1998, 1997; Beerli & Felsenstein 2001). With the availability of likelihood methods, complex models including both gene flow and population divergence (Nielsen & Wakeley 2001), and/or involving multiple populations can be analyzed. Unfortunately, full likelihood methods are not applicable to genome-scale datasets because of two significant limitations: 1) they do not scale well in the number of loci being analyzed and 2) they are not well suited for handling recombination. Methods by Yang and Rannala, Gronau, *et al.*, and Nielsen and Wakeley, among others (Yang & Rannala 1997; Gronau et al. 2011; Nielsen & Wakeley 2001), integrate over explicitly represented coalescence trees to find the joint likelihoods of short loci sampled from far apart in the genome, assuming that recombination is absent within each locus and that different loci are unlinked. The second assumption is realistic if loci are sampled far apart, but the first is problematic given that mutation and recombination rates are the same order of magnitude in humans and many other species. Simulation studies have shown that neglecting intra-locus recombination can generate significant biases when inferring population sizes and divergence times by maximum likelihood (Schierup & Hein 2000; Strasburg & Rieseberg 2010).

A parallel advancement to likelihood methods has been the production and analysis of genome-scale datasets. These datasets provide enough signal to test demographic questions of significant interest that cannot be answered using data from a small number of loci. Genome-wide data were instrumental, for example, in unearthing the presence of Neanderthal ancestry in modern humans (Green et al. 2010) and the antiquity of the Aboriginal Australian population (Rasmussen et al. 2011).

Motivated by the limitations of full likelihood methods and the power of large datasets, there is great interest in developing scalable approximate methods for population genetic inference across many recombining loci. One popular strategy is approximate Bayesian computation (ABC) (Tavaré et al. 1997; Pritchard et al. 1999; Beaumont et al. 2002), where the basic idea is to simulate many datasets under parameters drawn from a prior and rejection-sample by accepting replicates that are similar to an observed dataset. Another popular strategy, which is especially useful for the analysis of large SNP sets and genome-wide sequence data, is to fit the site frequency spectrum (SFS) using a composite likelihood approach. The main approximation here is to regard every segregating site as an independent sample from an expected SFS that can be computed from coalescent simulations (Nielsen 2000) or by numerically solving the Wright-Fisher diffusion equation (Williamson et al. 2005; Gutenkunst et al. 2009).

It is computationally easier to model the SFS as if it came from a collection of unlinked sites than to work with distributions of sequentially linked coalescence times. This strategy is statistically consistent in the limit of large amounts of data (Nielsen & Wiuf 5–12 April 2005; Wiuf 2006), but entails the loss of useful linkage information. A different class of method that is able to harness linkage information for demographic inference is the coalescent HMM; examples include CoalHMM, the Pairwise Sequentially Markov Coalescent (PSMC), and the sequentially Markov conditional sampling distribution (SMCSD) (Hobolth et al. 2007; Li & Durbin 2011; Steinrücken et al. 2012). Unlike the SFS-based methods and full likelihood methods, which require data from tens to hundreds of individuals, coalescent HMMs can infer demography from one or a few individuals. These methods assume that the sequence of times to most recent common ancestry (TMRCAs) in a sample is distributed like the output of a Markov process, which is almost (though not quite) true under the classical coalescent with recombination (Wiuf & Hein 1999; McVean & Cardin 2005). They use more of the information from a DNA sample than SFS-based methods do, but at present have a more limited ability to model subdivision and size changes at the same time. The PSMC produces detailed profiles of past population size (Li & Durbin 2011), but has limited ability to infer migration and subdivision; CoalHMM was recently generalized to accommodate subdivision and migration, but only in the context of the 6-parameter isolation with migration (IM) model (Miller et al. 2012).

Linkage information can be particularly revealing about recent demographic history and recent common ancestry between individuals. Many HMM-based methods have been devised to identify long haplotype tracts inherited *identical by descent* (IBD) from a single common ancestor without recombination (Browning & Browning 2011; Purcell et al. 2007; Moltke et al. 2011; Gusev et al. 2009). Several recent papers model the effects of recent demography on the size distribution of IBD blocks shared within populations (Hayes et al. 2003; MacLeod et al. 2009; Palamara et al. 2012). Others use the distribution of *migrant tracts* that were inherited IBD between individuals from different populations as a result of recent migration between those populations (Pool & Nielsen 2009; Gravel 2012; Ralph & Coop 2013). In particular, Gravel used migrant tracts to show that at least two migration “pulses" are needed to account for tracts admixed from Europe into African Americans (Gravel 2012), and Palamara, *et al.* provide IBD-based evidence for a detailed hypothesis about the past 200 generations of demographic history in the Ashkenazi Jewish population (Palamara et al. 2012).

Until now, no methods have formally bridged the time gap between IBD-based accounts of recent demography and the various methods that interpret older demographic signals. In this paper, however, we present an analytic method that draws power from linked sites over all genomic scales, not just short blocks where linkage is strong or long stretches of homozygosity between sequences with recent common ancestors. To do this, we study the spacing between adjacent polymorphisms in a sample of two haplotypes. The distance between adjacent polymorphisms is inversely correlated with local TMRCA; an L-base locus in a pairwise alignment that coalesced t generations ago is expected to contain $2L\mu t$ polymorphisms, μ being the mutation rate per generation. This motivates us to summarize a pairwise alignment by cutting it up at its polymorphic sites and recording the length distribution of the resulting shared haplotype fragments. We then model the effects of demography on frequencies of tracts of *identity by state* (IBS), where an L-base IBS tract is defined to be a string of L non-polymorphic base pairs bracketed by polymorphisms on the left and right (see Figure 2.1).

AGGTCGAGCTTG
ACGTCGAGCTGG

Figure 2.1: **An eight base-pair tract of identity by state (IBS).**

In a non-recombining mitochondrial alignment with TMRCA t, coalescent theory predicts that IBS tract lengths should be Poisson-distributed with mean $1/(2\mu t)$. In recombining DNA, more work is required to derive an expected IBS tract length spectrum, but such work is rewarded by the fact that the observed spectrum is informative about a wide range of historical coalescence times. Working with McVean and Cardin's sequentially Markov coalescent (SMC) and the related SMC' model by Marjoram and Wall(McVean & Cardin 2005; Marjoram & Wall 2006), we derive an approximate closed-form formula for the expected IBS tract length distribution in a two-haplotype sample, incorporating an arbitrary number of population size changes, divergence events, and admixture pulses between diverged populations. The formula is numerically smooth and quick to compute, making it well suited to the inference of demographic parameters using a Poisson composite likelihood approach. Empirical and predicted spectra can be graphed and visually inspected in the same way that is done with the SFS, but they encode linkage information that the SFS is missing.

In simulated data, we can accurately infer the timing and extent of admixture events that occurred hundreds of generations ago, too old for migrant IBD tracts to be reliably identified and thus for the methods of Pool & Nielsen (2009); Gravel (2012), and Palamara et al. (2012) to be applicable. IBS tracts have the advantage that their length distribution is directly observable; by computing this distribution under a model that incorporates intra-tract recombination, we can use the entire length spectrum for inference instead of only those short enough or long (and thus recently inherited) enough for internal recombination to be negligible. Although our derivation is for a sample size of only two haplotypes, we can parse larger samples by subsampling all haplotype pairs and regarding them as independent. Given sufficient data, this subsampling should not bias our results, though it may reduce our power to describe the very recent past.

To illustrate the power of our method, we use it to infer a joint history of Europeans and Africans from the high coverage 1000 Genomes trio parents. Previous analyses agree that Europeans experienced an out-of-Africa bottleneck and recent population growth, but other aspects of the divergence are contested (Pritchard 2011). In one analysis, Li and Durbin separately estimate population histories of Europeans, Asians, and Africans and observe that the African and non-African histories begin to look different from each other about 100,000-120,000 years ago; at the same time, they argue that substantial migration between Africa and Eurasia occurred as recently as 20,000 years ago and that the out-of-Africa bottleneck occurred near the end of the migration period, about 20,000-40,000 years ago. In contrast, Gronau et al. (2011) use their Bayesian analysis of 1 kb blocks to infer a Eurasian-African split that is recent enough (50 kya) to coincide with the start of the out of Africa bottleneck, detecting no evidence of recent gene flow between Africans and non-Africans. An earlier model by Schaffner et al. (2005) also explains a collection of summary statistics of human data with a model that includes recombination rate variation, but no recent gene flow. Gutenkunst et al. (2009) and Gravel et al. (2011) use SFS data to infer divergence times and gene flow levels that are intermediate between these two extremes. We aim to contribute to this dialogue by studying the same class of complex demographic models employed by Gutenkunst et al. (2009) and Gravel et al. (2011), models that have only been previously used to study the frequencies of alleles and of short haplotypes that are assumed not to recombine. Our method is the first to fit these models to haplotype sharing data similar to what is used by the PSMC and other coalescent HMMs. Through this effort, we hope to begin teasing apart which models and which data sources yield accurate information about our history.

2.2 Results

2.2.1 An accurate analytic IBS tract length distrubution

In the methods section, we derive a formula for the expected IBS tract length distribution shared between two DNA sequences from the same population or diverging populations. Our formula approximates the distribution expected under the SMC' model of Marjoram & Wall (2006), which in turn approximates the coalescent with recombination. We evaluate the accuracy of the approximation using simulations under a full coalescence model, by comparing our analytical formulae to the predictions obtained from the simulations. In general, we find that the approximations are very accurate as illustrated in the examples shown in Figures 2.2 and 2.3. To create each plot in Figure 2.2, we simulated several gigabases of pairwise alignment between populations that split apart 2,000 generations ago and experienced a 5% strength pulse of recent admixture, plotting the IBS tract spectrum of the alignment (for more details, see section A.2 of the supporting information). Figure 2.3 was generated under models of a single population with bottlenecks of varying intensity. In both of these scenarios the analytical approximations closely follow the distribution expected from full coalescent simulations.

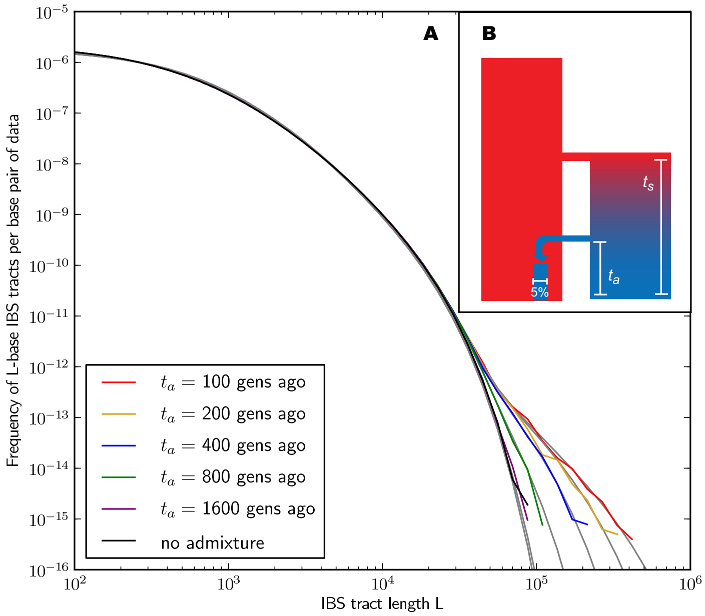


Figure 2.2: **Spectra of IBS sharing between simulated populations that differ only in admixture time.** Each of the colored tract spectra in Figure 2.2A was generated from 4.8×10^{10} base pairs of sequence alignment simulated with Hudson's MS (Hudson 2002). The IBS tracts are shared between two populations of constant size 10,000 that diverged 2,000 generations ago, with one haplotype sampled from each population. 5% of the genetic material from one population is the product of a recent admixture pulse from the other population. Figure 2.2B illustrates the history being simulated. When the admixture occurred less than 1,000 generations ago, it noticeably increases the abundance of long IBS tracts. The gray lines in 2.2A are theoretical tract abundance predictions, and fit the simulated data extremely well. To smooth out noise in the simulated data, abundances are averaged over intervals with exponentially spaced endpoints $\{\lfloor 1.25^n \rfloor\}_{n\geq 1}$.

If we wish to infer demography from IBS tract lengths, the following must be true: 1) IBS tract distribution must differ significantly between data sets simulated under coalescent histories we hope to distinguish, and 2) these differences must be predictable within our theoretical framework. Figures 2.2 and 2.3 provide evidence for both of these claims. For populations that diverged 2,000 generations ago, 5% admixture is detectible if it occurred less than 1,000 generations ago, late enough for the admixed material to significantly diverge from the recipient population. Likewise, population bottlenecks with the same strength-to-duration ratio are distinguishable if the population sizes differ by at least a factor of two during the bottleneck. As expected, longer IBS tracts are shared between populations that

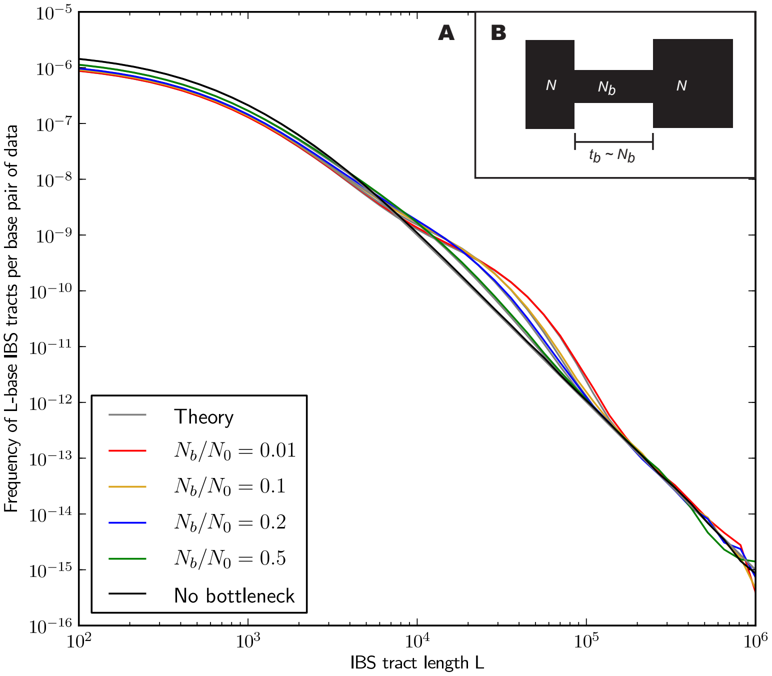


Figure 2.3: **Shared IBS tracts within bottlenecked populations.** As in Figure 2.2, each colored spectrum was generated by using MS to simulate 4.8×10^{10} base pairs of pairwise alignment. Both sequences are derived from the population depicted in panel B that underwent a bottleneck from size $N_0 = 10,000$ to size N_b, the duration of the bottleneck being $N_b/2$ generations. 1,000 generations ago, the population recovered to size 10,000. These bottlenecks leave similar frequencies of very long and very short IBS tracts because they have identical ratios of strength to duration, but they leave different signature increases compared to the no-bottleneck history in the abundance of 10^4–10^5-base IBS tracts. In grey are the expected IBS tract spectra that we derive analytically under each simulated history.

exchanged DNA more recently, suggesting that IBS tracts are highly informative about past admixture times, motivating the development of a statistical method for inferring demography based on IBS tract length distributions.

2.2.2 Estimates from simulated data

Inferring simulated population histories: Figures 2.2 and 2.3 suggest that by numerically minimizing the distance between observed and expected IBS tract spectra, we should be able to infer demographic parameters. We accomplish this by maximizing a Poisson composite likelihood function formed by multiplying the likelihoods of individual IBS tracts. Maximization is done numerically using the BFGS algorithm (Press et al. 2007).

To assess the power and accuracy of the method, we simulated 100 replicates of each of two histories with different admixture times and simultaneously inferred four parameters: admixture time, split time, admixture fraction, and effective population size. We obtain estimates that are extremely accurate and low-variance (see Table 2.1).

Table 2.1: **Inferring the parameters of a simple admixture scenario**

	t_a (gens)	t_s (gens)	f	N
True value:	400	2,000	0.05	10,000
Mean:	431	1,990	0.0505	9,806
Std dev:	51	41	0.00652	27
Bias:	31	-10	0.0005	-194
Mean squared error:	3280	1781	4.27×10^{-5}	3.84×10^{4}
True value:	200	2,000	0.05	10,000
Mean:	220	1,983	0.0499	10,003
Std dev:	28	39	0.00328	287
Bias:	20	-17	-0.0001	-3
Mean squared error:	1184	1810	1.08×10^{-5}	8.23×10^{4}

Using MS, we simulated 200 replicates of the admixture scenario depicted in Figure 2.2B. In 100 replicates, the gene flow occurred 400 generations ago, while in the other 100 replicates it occurred 200 generations ago. Our estimates of the four parameters t_a, t_s, f, N are consistently close to the true values, showing that we are able distinguish the two histories by numerically optimizing the likelihood function.

Comparison to ∂a∂i : We compared the new method to the method implemented in ∂a∂i, which can evaluate demographic scenarios with the same parameterization as ours. We focused on the simple admixture history summarized in Table 2.1. After simulating equal amounts of IBS tract and SFS data, we performed 20 numerical optimizations with each method starting from random points in the parameter space. Optimizations for the IBS method were almost always successful, converging to the global optimum, but optimizations in ∂a∂i using the default settings in the program often terminated near random initial starting points (see Supplementary Tables A.1 and A.2). This suggests that the implementation of the IBS based method has greater numerical stability than the implementation of ∂a∂i evaluated here, at least for scenarios involving discrete admixture pulses. This is not surprising as evaluation of the likelihood function in ∂a∂i involves numerical solution of partial differential equations.

For a simple four-parameter history, it is feasible to identify the maximum through a grid search that will be less sensitive to minor numerical instabilities. Using this type of optimization strategy both methods provide similar results (see Supplementary Figure A.9). Inspection of the likelihood surface also reveals that the two composite likelihood surfaces have different shapes–the IBS tract likelihood surface has a steeper gradient in the direction of admixture time, while the SFS likelihood surface is steeper along the split time axis (Supplementary Figure A.9).

2.2.3 IBS tracts in human data

Our analyses of simulated data indicate that real genomic IBS tracts should contain high-resolution demographic information. A potential obstacle, especially concerning recent demography, is that random sequencing and phasing errors will tend to break up long IBS tracts. To avoid this obstacle as much as possible, we chose to study IBS sharing within the 1000 Genomes trios: one mother-father-child family recruited from Utah residents of central European descent (the CEU) and another family recruited from the Yorubans of Ibadan, Nigeria (YRI).

We recorded the spectrum of IBS tracts shared between each pair of the eight parental haplotypes, which were sequenced at 20–60x coverage and phased with the help of the children by the 1000 Genomes consortium (1000 Genomes Project 2010). As expected, we observe much longer tracts shared within each population than between Europeans and Africans. The distribution of tracts shared between the populations, as well as within each population, were extremely robust to block bootstrap resampling (see Figure 2.4).

Sequencing and phasing errors: To gauge the effects of sequencing and phasing errors on IBS tract frequencies in real data, we also generated IBS tract spectra from samples that were sequenced at 2–4x coverage from the CEU and YRI populations, also as part of the 1000 Genomes pilot project (1000 Genomes Project 2010). Within each population, we found that samples sequenced at low coverage shared a higher frequency of short tracts and a lower frequency of long tracts than the high coverage trio parents did. (see Figure 2.5). In section A.3.2 of the supporting information, we mathematically describe how uniformly distributed errors can account for much of the difference between the high and low coverage data sets. It is encouraging that the frequencies of IBS tracts between 1 and 100 kB in length are almost the same between the two data sets, as are the frequencies of tracts shared between European and African sequences; this suggests that IBS sharing between low coverage sequences can yield reliable information about divergence times and the not-too-recent past. If we inferred demographic parameters from low coverage data without correcting for errors, however, the errors would lead to an upward bias in our estimates of recent population sizes.

Mutation and recombination rate variation: Regardless of data quality, all empirical IBS tract spectra are potentially affected by mutation and recombination rate variation (Hodgkinson et al. 2009; Kong et al. 2002). Our theoretical framework would make it possible

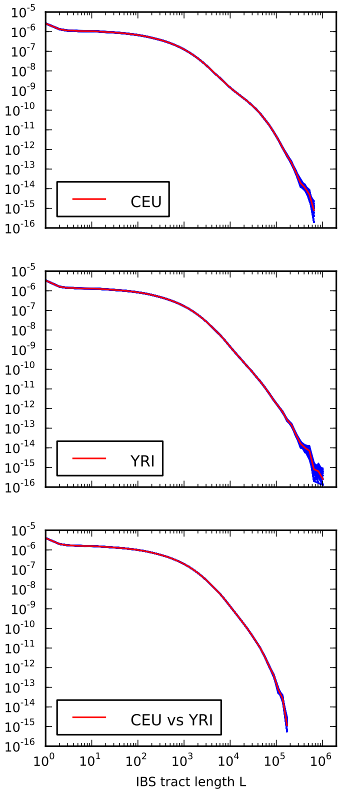


Figure 2.4: **Frequencies of IBS tracts shared between the 1000 Genomes trio parental haplotypes.** Each plot records the number of L-base IBS tracts observed per base pair of sequence alignment. The red spectrum records tract frequencies compiled from the entire alignment, while the blue spectra result from 100 repetitions of block bootstrap resampling. A slight upward concavity around 10^4 base pairs is the signature of the out of Africa bottleneck in the CEU.

to incorporate hotspots of mutation and recombination, but doing so would incur substantial computational costs when analyzing data sampled across the entire genome. We therefore made an effort to look for signatures of rate-variation bias in the real IBS tract data and to correct for such bias in the most efficient way possible.

To gauge the effects of recombination rate variation, we used the DECODE genetic map (Kong et al. 2002) to calculate the average recombination rate across all sites that are part of L-base IBS tracts. The results, plotted in Figure 2.6A, show no significant difference between the average recombination rate within long IBS tracts versus short ones. If recombination hotspots significantly reduced the frequency of long IBS tracts compared to what we would expect under the assumption of constant recombination rate, then the longest

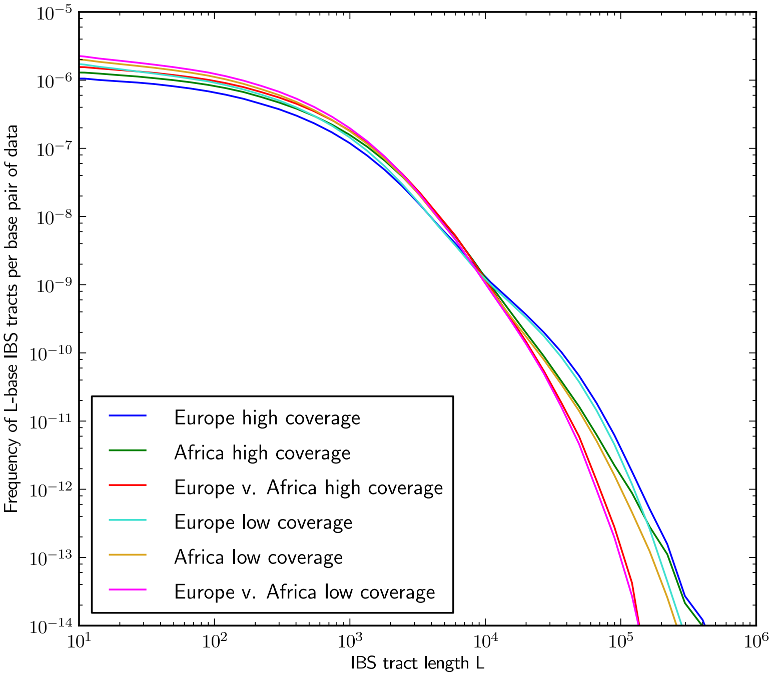


Figure 2.5: **IBS tract lengths in the 1000 Genomes pilot data: trios v. low coverage.** These IBS tract spectra were generated from pairwise alignments of the 1000 Genomes high coverage trio parental haplotypes and the CEU (European) and YRI (African) low coverage haplotypes, aligning samples within each population and between the two populations. Due to excess sequencing and phasing errors, the low coverage alignments have excess closely spaced SNPs and too few long shared IBS tracts. Despite this, frequencies of tracts between 1 and 100 kB are very similar between the two datasets and diagnostic of population identity.

observed IBS tracts should span regions of lower-than-average recombination rate; conversely, if recombination hotspots significantly increased the frequency of short IBS tracts, we would expect to see short tracts concentrated in regions of higher-than-average recombination rate. We observed neither of these patterns and therefore made no special effort to correct for recombination rate variation. Li and Durbin made a similar decision with regard to the PSMC, which can use IBS tract lengths to accurately infer past population sizes from data that was simulated using msHOT to reflect the true human recombination map.

To judge whether non-uniformity of the mutation rate was biasing the IBS tract spectrum, we computed the frequency of human/chimp fixed differences within IBS tracts of length L. We observed that short IBS tracts of < 100 bp are concentrated in regions with elevated rates of human-chimp substitution, suggesting that mutation rate variation has a significant impact on this part of the IBS tract spectrum. IBS tracts shorter than 5 base pairs long

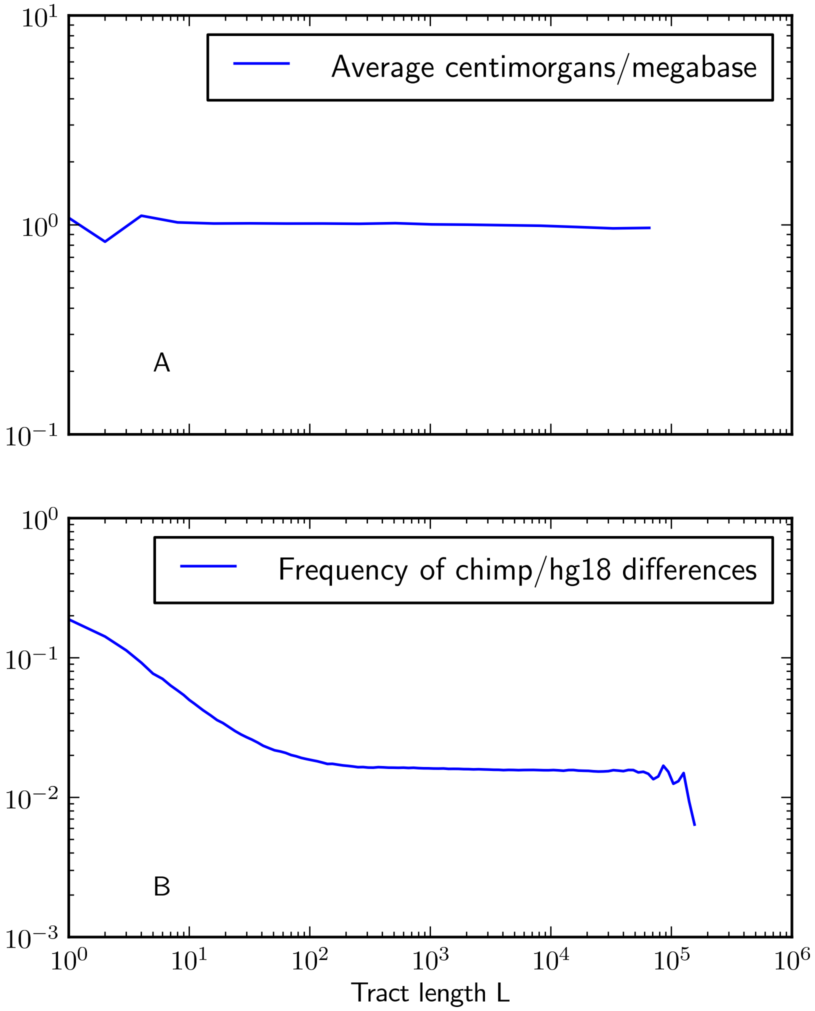


Figure 2.6: **Mutation and recombination rates within L-base IBS tracts.** Figure 2.6A shows that there is no length class of IBS tracts with a significantly higher or lower mutation rate than the genome-wide average (recombination rates are taken from the deCODE genetic map (Kong et al. 2002)). In contrast, Figure 2.6B shows that IBS tracts shorter than 100 base pairs occur in regions with higher rates of human-chimp differences than the genomewide average. These plots were made using IBS tracts shared between the CEU and YRI, but the results are similar for IBS sharing within each of the populations.

are dispersed fairly evenly throughout the genome, but human-chimp fixed differences cover more than 10% of the sites they span (see Figure 2.6B) as opposed to 1% of the genome overall.

In Hodgkinson, *et al.*'s study of cryptic human mutation rate variation, they estimated that the rate of coincidence between human and chimp polymorphisms could be explained by 0.1% of sites having a mutation rate that was 33 times the mutation rate at other sites (Hodgkinson et al. 2009). We modified our method to reflect this correction when analyzing real human data, assuming that a uniformly distributed 0.1% of sites have a scaled mutation rate of $\theta' = 0.033$, elevated above a baseline value of $\theta = 0.001$. We also excluded IBS tracts shorter than 100 base pairs from all computed likelihood functions (see Methods for more detail).

2.2.4 Human demography and the migration out of Africa

Previously published models of human demography: After generating spectra of empirical IBS sharing in the 1000 Genomes trios, we simulated IBS tract data under several conflicting models of human evolution that have been proposed in recent years. Two of these models were obtained from SFS data using the method ∂a∂i of Gutenkunst et al. (2009); these models are identically parameterized but differ in specific parameter estimates, which were inferred from different datasets. One model was fit to the SFS of the National Institute of Environmental and Health Sciences (NIEHS) Environmental Genome Project data, a collection of 219 noncoding genic regions (Gutenkunst et al. 2009); the other was fit by Gravel, *et al.* to a SFS of the 1000 Genomes low coverage data that was corrected for low coverage sampling bias (Gravel et al. 2011). The IBS tract length distributions corresponding to these models are qualitatively similar to each other but different from the tract length distribution of the 1000 Genomes trio data (see Supplementary Figure A.10). They also differ from the tract length distribution of the 1000 Genomes low coverage data, which is much more similar to the tract length distribution of the trio data as discussed under the heading “sequencing and phasing errors."

The models inferred from SFS data predict too few long IBS tracts shared between simulated Europeans and Africans, indicating too ancient a divergence time, too little subsequent migration, or both. There is also a dearth of long tracts shared within each population, a discrepancy that could be caused by too mild a European bottleneck and the lack of any historical size reduction in the African population.

A mild African bottleneck is a feature of the history that Li and Durbin infer using the PSMC, which also includes a more extreme European bottleneck than the ones inferred using ∂a∂i. Compared to the ∂a∂i histories, the PSMC predicts IBS tract sharing within Europe and Africa that is more similar to the pattern observed in the data (see Supplementary Figure A.10), which is not surprising given that the PSMC implicitly uses IBS tract sharing for inference.

A new demographic model: We were not able to match empirical IBS tract sharing in the trios by re-optimizing the parameters of a previously published history, but we were able to devise a new demographic model that is consistent with the distribution of IBS tract sharing in the trios. This model is illustrated in Figure 2.7. It bears many similarities to the model used by Gutenkunst, *et al.* and Gravel, *et al.*, including an ancestral population expansion, gene flow after the European-African divergence, a European bottleneck, and a recent European expansion. Unlike Gutenkunst, *et al.*, we also include a pulse of ghost admixture from an ancient hominid population into Europe, as well as a modest African population size reduction. All size changes are approximated by instantaneous events instead of gradual exponential growth.

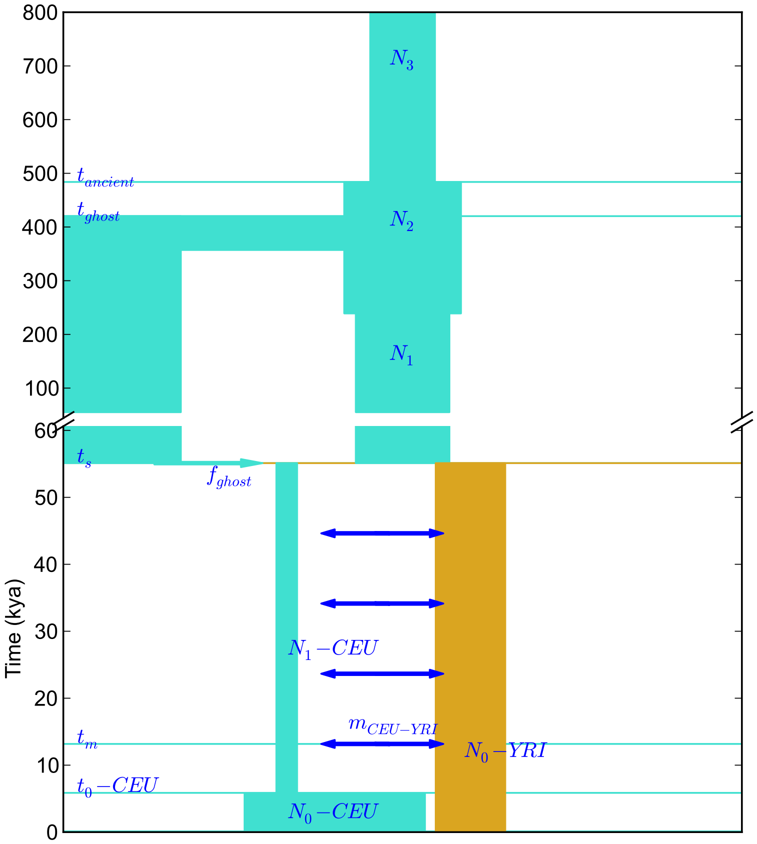


Figure 2.7: **A history inferred from IBS sharing in the CEU and YRI.** This history has many features in common with the one inferred by Li and Durbin, with much more detail about the European and African divergence. It is the simplest history we found that satisfactorily explained IBS tract sharing in the trio data.

We fit our model to the data using a Poisson composite likelihood approach; maximum likelihood parameters are listed in Table 2.2. We estimate that the European-African divergence occurred 55 kya and that gene flow continued until 13 kya. About 5.8% of European genetic material is derived from a ghost population that diverged 420 kya from the ancestors of modern humans. The out-of-Africa bottleneck period, where the European effective population size is only 1,530, lasts until 5.9 kya. Given this history and parameter estimates, we simulated 12 gigabases each of European and African sequence data under the full coalescent with recombination, obtaining an IBS tract length distribution that is very close to the one observed in the trios (see Figure 2.8).

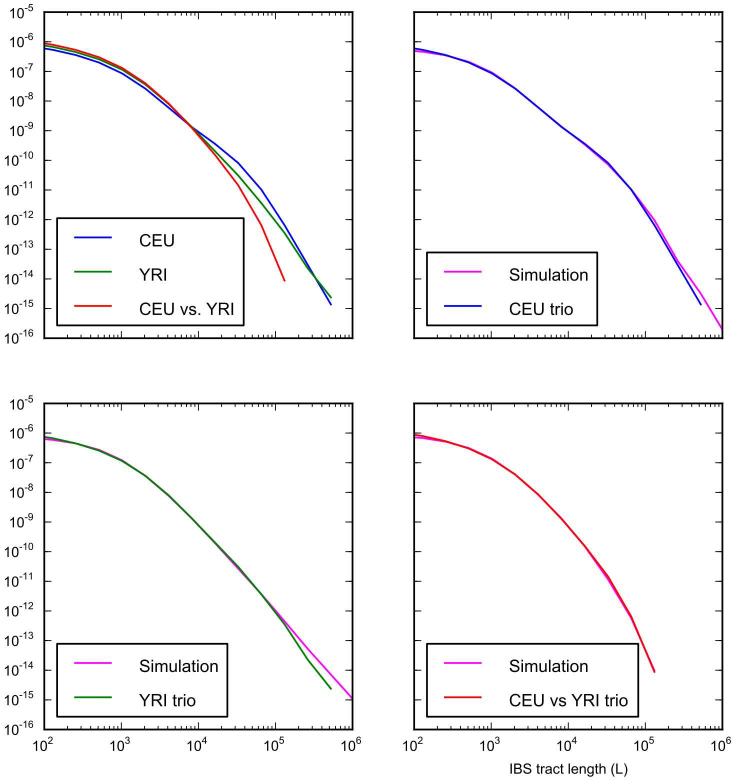


Figure 2.8: **Accurate prediction of IBS sharing in the trio data.** The upper left hand panel summarizes IBS tracts shared within the CEU and YRI 1000 Genomes trio parents, as well as IBS tract sharing between the two groups. The remaining three panels compare these real data to data simulated according to the history from Figure 2.7 with the maximum likelihood parameters from Table 2.2.

Assessing uncertainty: block bootstrap and replicate simulations To gauge the effects of local variation in the trio data, we re-optimized the parameters of our inferred history for each of 14 IBS tract spectra generated by block bootstrap resampling (see Fig-

Table 2.2: **Demographic parameters estimated from trio data**

Parameter	Estimate (kya)	Mean est. from simul.
t_0-CEU	5.86	5.0
t_m	13.17	15.11
t_0-YRI	55.11	48.42
t_{med}	239.06	146.44
t_s	55.11	57.32
t_{ghost}	365.12	278.73
t_{ancient}	483.89	426.04
f_{ghost}	0.0589	0.0404
N_0-CEU	13,298	117,149
N_1-CEU	1,531	1,694
N_0-YRI	5,125	5,246
N_1	6,900	6,325
N_2	8,606	7,901
N_3	4,772	5,608
$m_{\text{CEU-YRI}}$	1.52×10^{-4} gen^{-1}	1.80×10^{-4}

These times, population sizes and migration rates parameterize the history depicted in Figure 2.7. The migration rate $m_{\text{CEU-YRI}}$ is the fraction of the CEU population made up of new migrants from the YRI population each generation between t_m and t_s; it is also the fraction of the YRI population made up of new CEU immigrants each generation during the same time period.

ure 2.4). These inference results were consistent and low-variance. In addition, we used MS to simulate 27 datasets of the same size as the 1000 genomes trios, then inferred demographic parameters from each simulated dataset. This parametric bootstrapping revealed that some parameter estimates were biased, though there were no qualitative differences between the histories inferred from replicate simulations and the histories inferred from real data. Supplementary Figure A.5 compares the history inferred from real data to the mean parameters inferred from simulations.

To obtain further evidence for both ghost admixture and recent migration, we inferred parameters from the trio data under two models nested within our best-fit model. For one nested model, we set the recent migration rate to zero, obtaining parameters with a significantly worse fit to the data (composite log likelihood ratio 12891 compared to the best fit model). We then simulated data under the model with no recent migration and estimated the parameters of the full model. We inferred a migration period lasting only 5 ky, the minimum length permitted by the optimization bounds.

We also considered a nested model with the ghost admixture fraction set to zero. The best model with no ghost admixture also fit significantly worse than the maximum likelihood model, with a composite log likelihood ratio of 11796. When we simulated data under the restricted model and inferred a full set of 14 parameters from the simulated data, these included a ghost admixture fraction of 0.01, the smallest fraction permitted by the optimization bounds.

Given that models inferred from site frequency spectra do not fit the IBS tracts in human data, we simulated site frequency data under our inferred demographic model to see whether the reverse was true. The resulting spectrum had more population-private alleles than the NIEHS frequency spectrum previously analyzed by Gutenkunst, *et al* (see Section A.4.3 of the supporting information), a discrepancy that might result from biased population size estimates or from differences in error between IBS tract and SFS data.

2.3 Discussion

IBS tracts shared between diverging populations contain a lot of information about split times and subsequent gene flow; we can distinguish not only between instantaneous isolation and isolation with subsequent migration, but between recent admixture events that occur at modestly different times. We can accurately estimate the times of simulated admixture events that occurred hundreds of generations ago, too old for migrant tracts to be identified as IBD with tracts from a foreign population. In addition, we can distinguish short, extreme population bottlenecks from longer, less extreme ones that produce similar losses in the number of pairwise differences.

Our method harnesses most of the linkage information that is utilized by Li and Durbin's PSMC and the related coalescent HMMs of Hobolth, *et al.* and Paul, Steinrücken, and Song (Li & Durbin 2011; Hobolth et al. 2007; Paul et al. 2011), losing only the information about which IBS tracts are adjacent to each other in the data. In exchange, the method enjoys several advantages in computational efficiency over HMMs. The runtime of an HMM is linear in the number of base pairs being analyzed, whereas we incur only a small fixed computational cost when increasing the input sequence length and/or sample size. It takes $O(n^2\ell)$ time to compute the pairwise IBS tract spectrum of n sequences that are ℓ bases long, but this length distribution need only be computed once. After this is done, the time needed to find the composite likelihood of a demographic history does not depend on either n or ℓ. In addition, our runtime only grows linearly in the number of parameters d needed to describe a demographic history, whereas HMM decoding is $O(d^2)$. This scalability allows our program to handle all the demographic complexity that Gutenkunst et al. (2009) can, whereas Li and Durbin are limited to a *post hoc* argument linking large or infinite population size to periods of divergence.

Although the parameter estimates, including admixture times, in the simple four parameter model were fund to be approximately unbiased, we observed a weak estimation bias for some parameters when estimating a complex history with 14 parameters and very ancient demographic events. To our knowledge, no other methods have estimated such complex histories directly from the data, and we are hopeful that future improvements will help us infer complex histories more accurately. While perhaps it is disappointing that there is some bias, we emphasize that the bias is so small that it does not affect any qualitative conclusions. Two estimates that seem to be unbiased under parametric bootstrapping based on the full coalescence process, are the divergence time of 55 kya and the date of last gene flow of 13 kya; across simulated data, we estimate a mean divergence time of 57 kya and a mean date of last gene flow of 15 kya for simulated data. We have reduced the bias of the estimators over that of the SMC, by using Marjoram and Wall's SMC', which provides a more accurate approximation to the correlation structure along the sequence. It is possible that our method could be further improved by allowing IBS tracts to contain more than two internal recombinations; it could also be improved by allowing different parts of single tract to coalesce in epochs with different population sizes.

Our inferred human history mirrors several controversial features of the history inferred by Li and Durbin from whole genome sequence data: a post-divergence African population size reduction, a sustained period of gene flow between Europeans and Yorubans, and a “bump" period when the ancestral human population size increased and then decreased again. Unlike Li and Durbin, we do not infer that either population increased in size between 30 and 100 kya in either population. Li and Durbin postulate that this size increase might reflect admixture between the two populations rather than a true increase in effective population size; since our method is able to model this gene flow directly, it makes sense that no size increase is necessary to fit the data. In contrast, it is possible that the size increase we infer between 240 kya and 480 kya is a signature of gene flow among ancestral hominids.

Our estimated divergence time of 55 kya is very close to estimates published by Gravel et al. (2011) and Gronau et al. (2011), who use very different methods but similar estimated mutation rates to the $\mu = 2.5 \times 10^{-8}$ per site per generation that we use in this paper. However, recent studies of *de novo* mutation in trios have shown that the mutation rate may be closer to 1.0×10^{-8} per site per generation (1000 Genomes Project 2010; Scally & Durbin 2012; Kong et al. 2012). We would estimate older divergence and gene flow times (perhaps $1.75 = (2.5 \times 10^{-8} + 1.0 \times 10^{-8})/(1.0 \times 10^{-8} + 1.0 \times 10^{-8})$ times older) if we used the lower, more recently estimated mutation rate. This is because the lengths of the longest IBS tracts shared between populations should be approximately exponentially distributed with decay rate $T_{\text{split}}(\theta + \rho)$.

Sustained gene flow is essential to predict the true abundance of long IBS tracts shared between the African and European populations. The inferred rate of gene flow, $m = 1.78 \times 10^{-4}$ per generation, is the same order of magnitude as gene flow rates inferred from site frequency spectra using the method of Gutenkunst, *et al.* (Gutenkunst et al. 2009; Gravel et al. 2011) and by a analysis of human X chromosome diversity that employed the IM method of Hey and Nielsen (Cox et al. 2008). The two SFS-based analyses differ from ours, however, in that global gene flow drops off at the time of the European-Asian split about 23 kya. We find that high levels of gene flow must endure past this point to explain the abundance of long IBS tracts shared between the populations in these data.

Recent gene flow is not the only form of complex population structure that has left a signature in the IBS tracts shared between Africans and Europeans–we find strong log likelihood support for a pulse of ghost admixture from an ancient hominid species into non-

Africans. The admixture fraction and ghost population age are subject to some uncertainty, but our estimates of 6 % and 365 kya fit the profile of admixture between non-Africans and Neanderthals that was discovered through direct comparison of ancient and modern DNA (Noonan et al. 2006; Green et al. 2010). Without an ancient DNA sample, we lack power to date the ghost gene flow event and assume that it occurs immediately after the European-African divergence. Sankararaman et al. (2012) recently estimated that the Neanderthal gene flow event happened at least 47,000 years ago, much closer to estimates of the divergence date than to the present day.

To establish a less circumstantial link between Neanderthals and our inference of ghost admixture, it would be necessary to examine ancient DNA within our framework. This would be complicated by the higher error rates associated with ancient DNA sequencing and the lack of a reliable way to phase ancient samples. In general, it remains an open challenge to analyze IBS tracts shared between less pristine sequences than the ones we study here. Computational phasing programs like BEAGLE and MaCH effectively try to maximize the abundance of long IBS tracts among inferred haplotypes (Browning & Browning 2009; Li et al. 2010a), a fact that could seriously confound efforts to use IBS tracts for inference.

An opposite bias should result from excess sequencing errors, which have the potential to break up long shared haplotypes and degrade signals of gene flow and reduced population size. We see evidence of this degradation effect in low-coverage European and African sequences, but in the 1000 Genomes low coverage data this effect is very modest and does not noticeably influence IBS tract sharing between haplotypes from different populations. This suggests that IBS tracts in low coverage, computationally phased datasets can be used to make inferences about an intermediate-aged window of demographic history, inferences that would contribute valuable information about species where high quality data is not available and little to nothing is presently known about demography.

Even in high quality data, inference is complicated by departures of real evolutionary processes from the coalescent with uniform mutation and recombination. It is remarkable that real IBS tracts longer than 10 base pairs are distributed in a way that can be so closely approximated by our analytic predictions and by IBS tracts in simulated data; at the same time, real sequence alignments consistently harbor an excess of very short IBS tracts compared to simulated alignments, an excess we attribute to the non-uniformity of mutation rate in the genome. In this paper it was straightforward to neglect the frequencies of short tracts and correct the distribution of the remaining tracts for non-uniform mutation. In the future, however, it would be valuable to model the distribution of short tract frequencies and use them to learn more about the mutational process. At the moment, mutation rate variation is poorly understood compared to recombination rate variation, which does not appear to bias IBS tract frequencies (as seen in Figure 2.6). Because mutation rate variation does appear to affect IBS tract frequencies, we hope that IBS tracts can be used to obtain a more detailed picture of the mutational process just as we have used them to perform detailed inferences about demography.

Natural selection is beyond the scope of the models in this paper, but will be important for us to address in future work. One impetus for studying demography is to characterize long shared haplotypes caused by neutral events like bottlenecks so that they can be differentiated from the long shared haplotypes that hitchhike to high frequency around selected alleles (Sabeti et al. 2002; Pickerell et al. 2009). Histories with high SFS-based likelihoods can be quite inconsistent with genomic LD (Gutenkunst et al. 2009); to accurately describe neutral linkage in the genome, it is essential to harness linkage information as we have done here. Schaffner et al. (2005) addressed this need with their 2005 demographic model that reproduces r^2 correlations between pairs of common SNPs, but our model explains genome-wide LD on a much finer scale.

The empirical IBS tract distributions studied here are highly similar among bootstrap samples, making it unlikely that they are influenced by isolated loci under strong selection or other regional peculiarities. However, the data and model are surely influenced by the background selection (Charlesworth et al. 1995; McVicker et al. 2009). Background selection reduces diversity in a way that has been compared to a simple reduction in effective population size (Charlesworth et al. 1995; Lohmueller et al. 2011), and if selection is not being modeled explicitly, it is arguably better to report sizes that have been downwardly biased by background selection than sizes that do not accurately predict nucleotide diversity and LD.

In the future, it will be important to explain the discrepancy between the European-African site frequency spectrum studied by Gutenkunst, *et al.* and the SFS predicted by our model. The discrepancy has several potential causes, one being that the data were taken from different individuals. This could be especially important if Northern Europeans or Yorubans have significant population substructure. Another potential cause could be background selection–as previously mentioned, background selection makes coding regions look like they were generated under lower effective population size than neutral regions. We did not exclude coding regions here, opting to use as much data as possible, whereas the NIEHS frequency spectrum was recorded from intergenic loci. Bioinformatical issues may also play a role; the datasets were generated using different sequencing and filtering protocols, and even consistent bioinformatical protocols can have different effects on IBS tracts and site frequency data. A final culprit could be model specification–it is possible that a history with more structure than the one considered here could better fit the IBS tract length spectrum and the SFS simultaneously.

These caveats aside, we have here provided analytical results for the expected IBS tract length distribution within and between individuals from the same or different populations, and have shown that these results can be used to efficiently estimate demographic parameters. In the absence of likelihood-based methods for analyzing genome-wide population genetic data, methods such as the one presented here provide a computationally efficient solution to the demographic inference problem in population genetics.

2.4 Methods

2.4.1 Derivation of a frequency spectrum of shared haplotype lengths

A formula that is exact under the SMC: To derive an efficiently computable spectrum of shared haplotype lengths, we work within the setup of McVean and Cardin's sequentially Markov coalescent (SMC) (McVean & Cardin 2005) and introduce additional approximations as needed. We do not address the subject of IBS tracts in multiple sequence alignments; all alignments we refer to are pairwise.

The coalescent with recombination specifies a probability distribution on the coalescent histories that could have produced a sequence of base pairs $b_1 \cdots b_n$. Such a history assigns a TMRCA t_i to each base pair b_i, and in general the times $t_1, \ldots, t_n$ are related in a complex non-Markov way (Wiuf & Hein 1999). Because inference and computation under this model are so challenging, McVean & Cardin (2005) introduced a simpler coalescent process (the SMC) for which

$$p(t_n \mid t_1, \ldots, t_{n-1}) = p(t_n \mid t_{n-1}) \tag{2.1}$$

and coalescences are disallowed between sequences with no overlapping ancestral material. In a population with stationary coalescence time density $\zeta(t)$ and recombination probability ρ per base pair per generation, the SMC stipulates the following: If the nth base pair in a sequence coalesces at time t, then with probability $e^{-\rho t}$ there is no recombination in the joint history of base pairs n and $n+1$ before the find a common ancestor, meaning that base pair $n+1$ coalesces at time t as well. With infinitesimal probability $\rho e^{-\rho t_r} \mathbf{1}(t_r < t) dt_r$, however, the joint history of the two base pairs contains a recombination at time $t_r < t$. Given such a recombination, base pair $n+1$ is constrained to coalesce more anciently than t_r. Because of the assumption of no coalescence between sequences with nonoverlapping ancestral material, the distribution of t_{n+1} is independent of t_n given t_r. It is a renormalized tail of $\zeta(t)$:

$$p(t_n \mid t_{n-1}) = \exp(-\rho t_{n-1}) \delta_{t_{n-1}, t_n} + \int_{t_{(r)}=0}^{\min(t_{n-1}, t_n)} \rho \exp(-\rho t_{(r)}) \left(\int_{\tau=0}^{t_{(r)}} \zeta(\tau) d\tau \right)^{-1} dt_{(r)} \cdot \zeta(t_n)$$

For an alignment between sequences from constant-size populations that diverged at time τ_s, we can derive a formula for the expected IBS tract spectrum that is exact under the SMC. Specifically, we compute the expected value of $n_{L_{\text{tot}}}(L)$, the number of L-base IBS tracts in an L_{tot}-base sequence alignment. By setting $\tau_s = 0$, we can also compute this value for two sequences sampled within the same population.

In an alignment of length L_{tot}, any of the leftmost $L_{\text{tot}} - L - 1$ base pairs could be the leftmost polymorphic endpoint of an L-base IBS tract. Moreover, each of these $L_{\text{tot}} - L - 1$ base pairs has the same *a priori* probability of being such a leftmost endpoint. This motivates us to define $H_{\tau_s}(L)$ as the probability that a randomly chosen locus will be a) polymorphic and b) followed on the left by L homozygous base pairs, assuming that b) is not made impossible by edge effects. Assuming uniform mutation and recombination rates $\theta = 2N\mu$

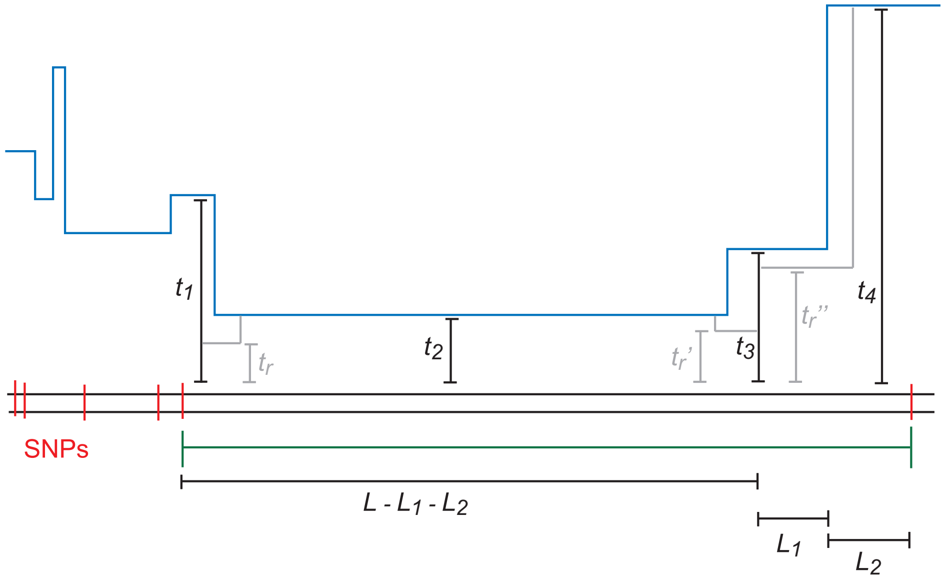


Figure 2.9: **An L-base IBS tract with three recombination events in its history.** A blue skyline profile represents the hidden coalescence history of this idealized IBS tract. In order to predict the frequency of these tracts in a sequence alignment, we must integrate over the coalesence times t_1, t_2, t_3 as well as the times $t_r < \min(t_1, t_2)$, $t'_r < \min(t_2, t_3)$, and $t''_r < \min(t_3, t_4)$ when recombinations occurred.

and $\rho = 2Nr$, it follows that

$$E[n_{L_{\text{tot}}}(L)] = (L_{\text{tot}} - L - 1)(H_{\tau_s}(L) - H_{\tau_s}(L+1)).$$

It is straightforward but computationally costly to relax the assumption of uniform mutation and recombination rates. We will wait to revisit this issue in the context of data analysis.

For now, let $H_{\tau_s}(L,t)$ be the joint infinitesimal probability that a) a randomly selected locus b_0 is polymorphic, b) the next L base pairs $b_1,\ldots,b_L$ sampled from left to right are non-polymorphic, and c) the rightmost base pair b_L has TMRCA t. We can use the sequential Markov property of the SMC to write down a recursion for $H_{\tau_s}(L,t)$ in L: if $\mathbf{1}_{\mathrm{hom}}(b_L)$ denotes an indicator function for the event that base pair b_L is homozygous and t_L denotes the coalescence time of base pair L, then

$$\begin{aligned} H_{\tau_s}(L,t) &= \int_{t_{L-1}=\tau_s}^{\infty} H_{\tau_s}(L-1,t_{L-1})\cdot P(t_L=t \mid t_{L-1})\cdot P(\mathbf{1}_{\mathrm{hom}}(b_L)\mid t_L=t)dt_{L-1} \\ &= e^{-t\theta}\int_{t_{L-1}=\tau_s}^{\infty} H_{\tau_s}(L-1,t_{L-1})\cdot P(t_L=t\mid t_{L-1})dt_{L-1}. \end{aligned}$$

When $t=t_{L-1}$, the quantity $P(t_L=t\mid t_{L-1})$ is simply $e^{-t\rho}$, the probability that neither lineage undergoes recombination. Conversely, a recombination is required whenever $t\neq t_{L-1}$; to compute $P(t_L=t\mid t_{L-1})$ when $t_{L-1}\neq t$, we must marginalize over the time $t_{(r)}$ of the recombination that caused the change in TMRCA (see Figure 2.10). Paul, Steinrücken, and Song used a similar computation to motivate the transition probabilities of their sequentially Markov conditional sampling HMM (Paul et al. 2011):

$$\begin{aligned} p(t_L\mid t_{L-1}) &= \int_{t_{(r)}=0}^{\min(t_{L-1},t)} \mathbb{P}(\text{an ancestor of } b_{L-1} \text{ recombined at time } t_{(r)}) \\ &\quad \cdot\mathbb{P}(t_L=t\mid t_L>t_{(r)})dt_{(r)}+e^{-\rho t}\delta_{t_{L-1},t} \\ &= \int_{t_{(r)}=t_{L-1}}^{\min(t_{L-1},t)} \rho e^{-\rho t_{(r)}}\cdot\frac{e^{-(t-\tau_s)}}{e^{\min(0,-t_{(r)}+\tau_s)}}dt_{(r)}+e^{-\rho t}\delta_{t_{L-1},t} \\ &= \frac{\rho e^{-(t-\tau_s)-\rho\tau_s}}{1-\rho}(e^{\min(t-\tau_s,t_0-\tau_s)(1-\rho)}-1)+e^{-(t-\tau_s)}(1-e^{-\rho\tau_s}) \\ &\quad + e^{-\rho t}\delta_t(t_{L-1}) \end{aligned}$$

This yields that

$$\begin{aligned} H_{\tau_s}(L,t) &= \int_{t_{L-1}=\tau_s}^{\infty} H_{\tau_s}(L-1,t_{L-1})\cdot e^{-t\theta}\Bigg(\frac{\rho e^{-(t-\tau_s)-\rho\tau_s}}{1-\rho}(e^{\min(t-\tau_s,t_{L-1}-\tau_s)(1-\rho)}-1) \\ &\quad + e^{-(t-\tau_s)}(1-e^{-\rho\tau_s})\Bigg) \\ &\quad + H_{t_s}(L-1,t)e^{-t(\rho+\theta)}. \end{aligned}$$

To find $H_{\tau_s}(L,t)$, all we need to do is apply the integral operator (2.8) $L-1$ times to the base case

$$H_{\tau_s}(0,t)=e^{-(t-\tau_s)}-e^{-(t-\tau_s)-t\theta}.$$

Moreover, it turns out that this integral recursion can be transformed into an algebraic recursion that is more efficient to compute:

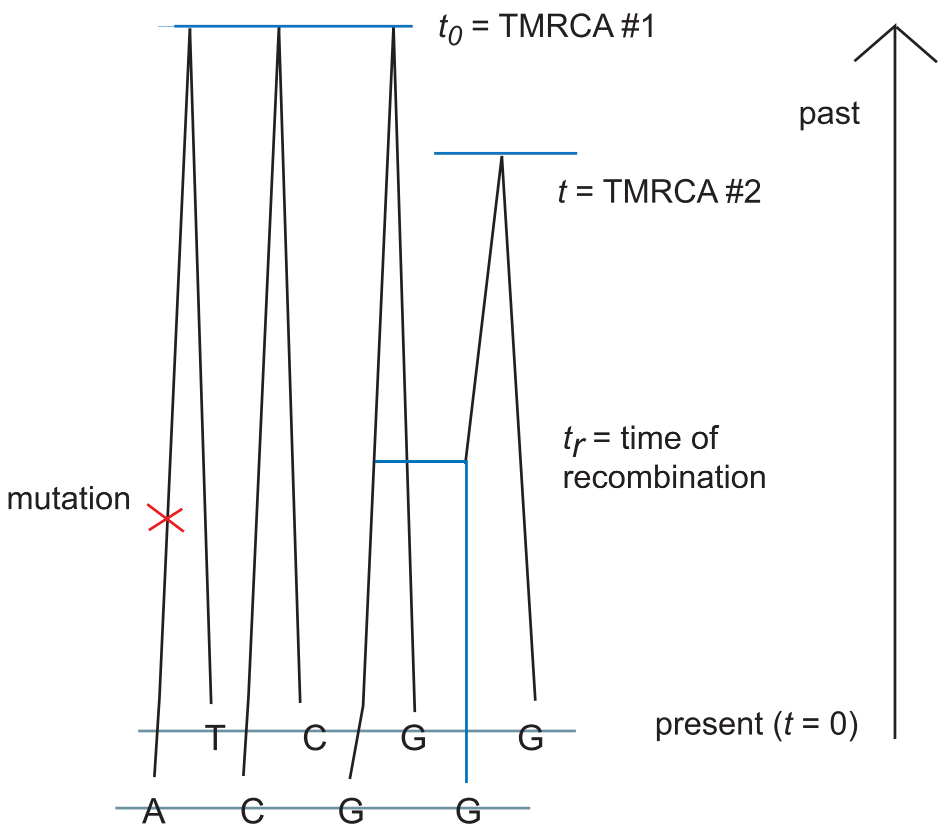


Figure 2.10: **The coalescent with recombination and the sequentially Markov coalescent associate an observed pair of DNA sequences with a history that specifies a time to most recent common ancestry for each base pair.** Polymorphisms are caused by mutation events, while changes in TMRCA are caused by recombination events.

Claim 1 *The sampling probability $H_{\tau_s}(L,t)$ can be written in the form*

$$H_{\tau_s}(L, t + \tau_s) = \sum_{i=0}^{L} A_i(L) e^{-t(1+i(\rho+\theta))} + B_i(L) e^{-t(1+i(\rho+\theta)+\theta)}, \tag{2.9}$$

with coefficients that satisfy the following recursions and base cases:

$$\begin{aligned}
A_{i+1}(L+1) &= A_i(L)e^{-\tau_s(\rho+\theta)}\left(1-\frac{\rho}{(i(\rho+\theta)+\rho)(1+i(\rho+\theta))}\right)\\
B_{i+1}(L+1) &= B_i(L)e^{-\tau_s(\rho+\theta)}\left(1-\frac{\rho}{(i+1)(\rho+\theta)(1+i(\rho+\theta)+\theta)}\right)\\
B_0(L+1) &= \sum_{i=0}^{L}\frac{\rho A_i(L)e^{-\tau_s(\rho+\theta)}}{(i(\rho+\theta)+\rho)(1+i(\rho+\theta))}+\frac{\rho B_i(L)e^{-\tau_s(\rho+\theta)}}{(i+1)(\rho+\theta)(1+i(\rho+\theta)+\theta)}\\
&\quad+\left(e^{-\tau_s\theta}-e^{-\tau_s(\theta+\rho)}\right)\sum_{i=0}^{L}\left(\frac{A_i(L)}{1+i(\rho+\theta)}+\frac{B_i(L)}{1+\theta+i(\rho+\theta)}\right)\\
A_0(L+1) &= 0\\
A_0(0) &= 1,\\
B_0(0) &= -e^{-\theta\tau_s}
\end{aligned}$$

It is straightforward to prove Claim 1 by applying the integral operator (2.8) to expression (2.9). The upshot is that

$$H_{\tau_s}(L)=\int_{t=\tau_s}^{\infty}H_{\tau_s}(L,t)dt$$

can be computed in $O(L^2)$ time using elementary algebra.

While Claim 1 enables an exact computation that is orders of magnitude faster than using numerical integration to solve recursion (2.8), it is still too slow for our purposes. It will prove more useful to derive an approximate formula for $H_{\tau_s}(L)$ that is not exact with respect to the SMC but whose computation time does not depend on L; this is accomplished by limiting the total number of recombinations that can occur within the history of an IBS tract.

Restricting the number of ancestral recombination events: In principle, each base pair of an L-base IBS tract could coalesce at a different time, with each TMRCA partially decoupled from its neighbors by an ancestral recombination event. In practice, however, most L-base IBS tracts will contain many fewer than L distinct TMRCAs. As we move left along the history of a sequence, the probability of seeing k ancestral recombinations before we see a single ancestral mutation declines geometrically as $(\rho/(\rho+\theta))^k$. Moreover, each ancestral recombination represents a chance for the TMRCA to become ancient and force mutation to end the IBS tract soon. Lohse and Barton were able to show under the full coalescent with recombination (not the SMC) that if $t_L \neq t_{L-1}$, then $\mathbb{E}[t_L] \gg \mathbb{E}[t_{L-1}]$ (Barton June 28, 2012).

To speed the computation, we assume that an L-base IBS tract contains at most two internal recombinations. To make this precise, we let $H_{\tau_s}(L) \approx H_{\tau_s}^{(0)}(L)+H_{\tau_s}^{(1)}(L)+H_{\tau_s}^{(2)}(L)$, where $H_{\tau_s}^{(i)}$ is the joint probability that a) a randomly selected base pair is polymorphic, b) the next L base pairs to the left are IBS, and c) the coalescent history of these $L+1$ base pairs contains exactly i ancestral recombinations.

Computing $H^{(0)}_{\tau_s}(L)$ is easy because it involves integrating over only one coalescence time:

$$\begin{aligned} H^{(0)}_{\tau_s}(L) &= \int_{t=\tau_s}^{\infty} e^{-(t-\tau_s)} \cdot (1-e^{-t\theta}) \cdot e^{-tL(\rho+\theta)} dt \\ &= \frac{e^{-\tau_s L(\rho+\theta)}}{1+L(\rho+\theta)} - \frac{e^{-\tau_s(L(\rho+\theta)+\theta)}}{1+L(\rho+\theta)+\theta} \end{aligned}$$

When $i>0$, however, the complexity of the integral grows quickly. We must marginalize over $i+1$ different coalescence times $t_0, \dots, t_i$, i different times of recombination $t_{(r,1)}, \dots, t_{(r,i)}$, and i recombination breakpoint locations $L_1 < \cdots < L_i$. For example,

$$\begin{aligned} H^{(1)}_{\tau_s}(L) &= \sum_{L_1=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=0}^{\min(t_0,t)} e^{-t_0}(1-e^{-t_0\theta}) \cdot e^{-t_0 L_1(\rho+\theta)} \cdot \rho e^{-\rho t_{(r)}} \\ &\cdot e^{-(t-t_{(r)})} \cdot e^{-t(L-L_1)(\rho+\theta)} dt_{(r)} dt dt_0 \end{aligned}$$

In the supplementary information, we evaluate this expression in closed form after approximating the sum by an integral. In the same way, we compute $H^{(2)}_{\tau_s}(L)$.

Adding recombination and population size changes: As demonstrated in the results section, IBS tract lengths are very informative about the timing of admixture pulses. This makes it interesting to look at IBS tracts shared between two populations A and B that diverged at time τ_s but exchanged genetic material at a more recent time τ_a. To this end, we let $H_{\tau_a,\tau_s,f}(L)$ be the frequency of L-base IBS tracts shared between A and B assuming that a fraction f of A's genetic material was transferred over from B in a single pulse at time τ_a, with the remaining fraction constrained to coalesce with B before τ_s. If we define $H^{(0)}_{\tau_a,\tau_s,f}(L), H^{(1)}_{\tau_a,\tau_s,f}(L), \dots$ the same way as before, then $H^{(0)}_{\tau_a,\tau_s,f}(L)$ is simply a linear combination of $H^{(0)}_{\tau_s}(L)$ and $H^{(0)}_{\tau_a}(L)$:

$$\begin{aligned} H^{(0)}_{\tau_a,\tau_s,f}(L) &= \frac{fe^{-\tau_a L(\rho+\theta)} + (1-f)e^{-\tau_s L(\rho+\theta)}}{1+L(\rho+\theta)} - \frac{fe^{-\tau_a(L(\rho+\theta)+\theta} + (1-f)e^{-\tau_s(L(\rho+\theta)+\theta)}}{1+L(\rho+\theta)+\theta} \\ &= fH^{(0)}_{\tau_a} + (1-f)H^{(0)}_{\tau_s}. \end{aligned}$$

The next term $H^{(1)}_{\tau_a,\tau_s,f}(L)$ is much more challenging to compute exactly; this is done in the supplementary information. The challenge stems from the fact that the recombination site might partition the tract into two components that have different “admixture statuses"– one side might be constrained to coalesce before the ancestral split time, and the other side might not. As a result $H^{(1)}_{\tau_a,\tau_s,f}(L)$ is not an exact linear combination of $H^{(1)}_{\tau_a}(L)$ and $H^{(1)}_{\tau_s}(L)$.

A similar challenge arises when we consider histories where the effective population size varies with time. For a simple example, consider the vector of times $\nu = (\nu_0 = 0, \nu_1, \nu_2, \nu_3 = \infty)$ with $\nu_1 < \nu_2$ and the vector of sizes $\mathbf{N} = (N_1, N_2, N_3)$. It will be useful to let $H_{\tau_s}(L)$ denote $H_{\tau_s}(L)$ in a population where the constant effective population size is N. Let $H_{\nu,\mathbf{N}}(L)$

denote the frequency of L-base IBS tracts in a population that underwent a bottleneck, such that the population size function $N(t)$ is piecewise constant with $N(t) = \sum_i N_i \mathbf{1}(\nu_i \le t < \nu_{i+1})$. This population has a coalescence density function that is a linear combination of exponentials, which implies that $H^{(0)}_{\nu,\mathbf{N}}(L)$ is a linear combination of the quantities $H^{(0)}_{\nu_i,N_i}$:

$$\begin{aligned} H^{(0)}_{\nu,\mathbf{N}}(L) &= \int_{t=0}^{\infty} \left(\mathbf{1}(t < \nu_0) \frac{1}{N_0} e^{-t/N_0} + \mathbf{1}(\nu_1 \le t < \nu_2) \frac{1}{N_1} e^{-\nu_1/N_0 - (t-\nu_1)/N_1} \right. \\ &\quad \left. + \mathbf{1}(t \ge \nu_2) \frac{1}{N_2} e^{-\nu_1/N_0 - (\nu_2-\nu_1)/N_1 - (t-\nu_2)/N_2} \right) e^{-tL(\rho+\theta)} dt \\ &= H^{(0)}_{0,N_0}(L) + e^{-\nu_1/N_0} \left(H^{(0)}_{\nu_1,N_1}(L) - H^{(0)}_{\nu_1,N_0}(L) \right) \\ &\quad + e^{-\nu_1/N_0 - (\nu_2-\nu_1)/N_1} \left(H^{(0)}_{\nu_2,N_2}(L) - H^{(0)}_{\nu_2,N_1}(L) \right). \end{aligned}$$

As in the case of an admixed population, the next term $H^{(1)}_{\nu,\mathbf{N}}(L)$ is harder to compute because it is difficult to write down the frequencies of IBS tracts that span multiple epochs (i.e. when the left hand part of a tract coalesces earlier than ν_1 and the right hand part coalesces later than ν_1 during a time period of smaller effective population size). The higher terms ($H^{(2)}_{\nu,\mathbf{N}}$, etc.) are more complicated still. Rather than attempt to compute these terms for a simple bottleneck history, we have developed an approximation for $H_{\nu,\mathbf{N}}(L)$ that involves little extra computation and generalizes easily to more complicated histories. The approximation can be described as the following modification to the SMC: if the left hand side of an IBS tract coalesces between ν_i and ν_{i+1} and the tract then recombines at time $t_{(r)}$, the probability distribution of the new coalescence time is $\frac{1}{N_i} e^{-(t-t_{(r)})/N_i}$ instead of $\sum_j \frac{1}{N_j} e^{-(t-t_{(r)})/N_j} \mathbf{1}(\nu_j < t \le \nu_{j+1})$. If we let $\hat{H}_{\nu,\mathbf{N}}(L)$ be the IBS tract spectrum under this assumption, we have that

$$\begin{aligned} \hat{H}_{\nu,\mathbf{N}}(L) &= H_{0,N_0}(L) + e^{-\nu_1/N_0} \left(H_{\nu_1,N_1}(L) - H_{\nu_1,N_0}(L) \right) \\ &\quad + e^{-\nu_1/N_0 - (\nu_2-\nu_1)/N_1} \left(H_{\nu_2,N_2}(L) - H_{\nu_2,N_1}(L) \right) \end{aligned}$$

This linear approximation strategy generalizes to any history that is described by size changes, splits, and admixture pulses, since every such history has a coalescence density function that is a linear combination of exponentials. Figure 2.3 shows a close agreement between $\hat{H}_{\nu,\mathbf{N}}(L)$ and the IBS tracts in data simulated under bottleneck histories with MS.

Increasing accuracy via the SMC': If we approximate the frequency of L-base IBS tracts by calculating $H^{(0)}(L) + H^{(1)}(L) + H^{(2)}(L)$ as described above, we slightly underestimate the frequency of intermediate-length tracts between 10^3 and 10^5 base pairs long. This bias can be reduced by replacing $H^{(0)}(L)$, the largest summand, with a term $H^{(0')}(L)$ that is derived from Marjoram and Wall's SMC'.

The SMC' is a coalescent approximation that is slightly more complex and more accurate than the SMC (Marjoram & Wall 2006). Both the SMC and the SMC' are efficient for simulating long DNA samples with many ancestral recombinations, and both satisfy the Markov property from equation (2.1).

Under McVean and Cardin's SMC, t_{n-1} and t_n are distinct whenever a recombination occurs between b_{n-1} and b_n. As a result, $t_{n-1} = t_n$ with probability $\mathbb{P}_{\text{SMC}}(t_n = t \mid t_{n-1} = t) = e^{-\rho t}$. Under the SMC', the situation is more complex: in the event of an ancestral recombination between base pairs b_{n-1} and b_n, it is possible for the times t_{n-1} and t_n to be equal because of a "back-coalescence" involving part of the ancestral recombination graph that the SMC does not retain. In particular,

$$\begin{aligned}
\mathbb{P}_{\text{SMC}',\tau_s}(t_n = t \mid t_{n-1} = t) &= e^{-\rho t} + \int_{t_r=0}^{\tau_s} \rho e^{-\rho t_r} \left(\int_{t'=t_r}^{\tau_s} e^{-(t'-t_r)} dt' \right. \\
&\quad \left. + \int_{t'=\tau_s}^{t} \frac{1}{2} \cdot 2e^{-(\tau_s - t_r) - 2(t'-\tau_s)} dt' \right) dt_r \\
&\quad + \int_{t_r=\tau_s}^{t} \rho e^{-\rho t_r} \int_{t'=t_r}^{t} \frac{1}{2} \cdot 2e^{-2(t'-t_r)} dt' dt_r \\
&= 1 - \frac{\rho}{4} \left(3 - 2e^{-\tau_s} - 2e^{-t-(t-\tau_s)} + e^{-2(t-\tau_s)} + 2t - 2\tau_s\right) + O(\rho^2)
\end{aligned}$$

Motivated by Equation (2.28), we can replace $e^{-\rho t}$ with $\exp(-\rho(3 - 2e^{-\tau_s} - 2e^{-t-(t-\tau_s)} + e^{-2(t-\tau_s)} + 2t - 2\tau_s)/4)$ in Equation (2.16) to compute the probability of observing L base pairs that are IBS with no internal recombinations that change the coalescence time. We obtain that

$$\begin{aligned}
H_{\tau_s}^{(0')}(L) &= \int_{t=\tau_s}^{\infty} e^{-(t-\tau_s)} \cdot (1 - e^{-t\theta}) \cdot e^{-tL\theta} \\
&\quad \cdot \exp(-\rho L(3 - 2e^{-\tau_s} - 2e^{-t-(t-\tau_s)} + e^{-2(t-\tau_s)} + 2t - 2\tau_s)/4) dt \\
&= 2e^{-L((1-e^{-\tau_s} - e^{\tau_s} + e^{2\tau_s})\rho/2 + \tau_s\theta)} \cdot \\
&\quad \left(\frac{1}{2 + L\rho + 2L\theta} \cdot {}_1F_1\left(1, \frac{6 + L\rho + 2L\theta}{4}, -\frac{L\rho}{4}\left(1 + 2e^{\tau_s} - 2e^{2\tau_s}\right)\right) \right. \\
&\quad \left. - \frac{e^{-\tau_s\theta}}{2 + L\rho + 2(L+1)\theta} \cdot {}_1F_1\left(1, \frac{6 + L\rho + 2(L+1)\theta}{4}, -\frac{L\rho}{4}\left(1 + 2e^{\tau_s} - 2e^{2\tau_s}\right)\right) \right).
\end{aligned}$$

In this formula,

$$ {}_1F_1(a, b, z) = \sum_{k=0}^{\infty} \frac{a(a+1)\cdots(a+k-1)}{b(b+1)\cdots(b+k-1)} \frac{z^k}{k!} $$

is a confluent hypergeometric function of the first kind, which we compute via the Python `mpmath` library.

2.4.2 Inference strategy

The previous section described how we compute $\mathbb{E}[n_{L_{\text{tot}}}(L)]$, the expected number of L-base IBS tracts present in L_{tot} base pairs of sequence alignment. As L_{tot} approaches infinity, the law of small numbers predicts that $n_{L_{\mathrm{tot}}}(L)$ should become Poisson-distributed about its mean. This motivates us to compare models Θ, Θ' by evaluating the Poisson composite log likelihood of the IBS tract spectrum under each model:

$$\log \mathcal{L}(\Theta) = \sum_{L=1}^{\infty} -\mathbb{E}_{\Theta}[n_{L_{\mathrm{tot}}}(L)] + n_{L_{\mathrm{tot}}}(L) \log \mathbb{E}_{\Theta}[n_{L_{\mathrm{tot}}}(L)] - \log n_{L_{\mathrm{tot}}}(L)!$$

We emphasize that this is a composite likelihood function formed by multiplying likelihoods together that are not necessarily independent of each other. Nonetheless, the resulting function may provide estimators with desirable statistical properties, as illustrated in the Results section. Throughout this paper, when discussing composite likelihood functions we will use the shorter term 'likelihood function'. However, we emphasize that we never apply general asymptotic theory for likelihood function to the composite likelihood functions derived and applied in this paper.

This formula above has a tendency to destabilize numerically; its many alternating terms must be computed by multiplying small $P_{t_s,N}(L)$ numbers by the very large number L_{tot}, leading to a rapid loss of machine precision. This loss of precision can be avoided, however, by grouping IBS tracts into bins with endpoints $0 < b_1 < b_2 < \cdots < b_n$ and evaluating a log likelihood function with one term per bin. In addition to improving numerical stability, binning reduces the time required to compute and optimize the likelihood function. Letting $n_{L_{\mathrm{tot}}}(b_i, b_{i+1}) = \sum_{L=b_i}^{b_{i+1}-1} n_{L_{\mathrm{tot}}}(L)$, we define

$$\log \mathcal{L}_{\mathbf{b}}(\Theta) = \sum_{i} -\mathbb{E}_{\Theta}[n_{L_{\mathrm{tot}}}(b_i, b_{i+1})] + n_{L_{\mathrm{tot}}}(b_i, b_{i+1}) \log \mathbb{E}_{\Theta}[n_{L_{\mathrm{tot}}}(b_i, b_{i+1})] - \log n_{L_{\mathrm{tot}}}(b_i, b_{i+1})!$$

The ideal choice of bins depends on the nature of the demography being inferred. We found that exponentially spaced bins ($b_i = 2^i$) performed well for most inference purposes, and these are the bins we used to infer human demography from the 1000 Genomes trios. The optimization results were not sensitive to the fine-scale choice of binning scheme. For inferring admixture times from data simulated without population size changes, a different binning scheme was more efficient because only the longest tracts were truly informative (this is clear from looking at Figure 2.2). We took $b_0 = 20,000$ and $b_{i+1} = 1.3 \cdot b_i$.

To infer the joint history of two populations A and B, we use the quasi-Newton BFGS algorithm to simultaneously maximize the likelihood of three different IBS tract spectra: the first summarizes an alignment of two sequences from population A, the second summarizes an alignment of two sequences from population B, and the third summarizes an alignment between population A and B. The three likelihoods are computed with respect to the same set of parameters Θ and multiplied together. Computing the joint likelihood of an n-population history requires $O(n^2)$ computational time compared to the likelihood of a one-population history with the same number of size change and admixture parameters.

2.4.3 Mutation rate variation

The human genome is known to contain complicated patterns of mutation rate variation, as well as a better-understood map of recombination rate variation (Hodgkinson et al. 2009; Kong et al. 2002). As discussed in the results, only mutation rate variation appears to bias the distribution of IBS tracts and is therefore taken into account by our method. Long regions of elevated mutation rate should elevate the abundance of short IBS tracts but have little effect on the abundance of longer IBS tracts. Because the distribution of such regions is not well understood and is outside the scope of this paper, we simply restrict our inference to the spectrum of tracts longer than 100 base pairs.

Hodgkinson et al. (2009), among others, have shown that sites of elevated mutation rate are not always grouped together in the human genome. They propose several models of cryptic, dispersed variation that could explain observations of correlation between human and chimp polymorphism. Of the models that they deem consistent with the data, the one that we incorporate into our method is a bimodal distribution of mutation rate where 99.9 % of all sites have the baseline rate $\mu = 2.5 \times 10^{-8}$ mutations per base per generation and the remaining 0.1% have an elevated rate $\mu' = 38\mu$. It is straightforward to compute the probability $P'_{\text{IBS}}(L)$ that a site of elevated mutation rate followed by $L + 1$ bases of normal mutation rate is the left endpoint of an L-base IBS tract. If we were to randomly assign a higher mutation rate to 0.1% of the L IBS bases and compute the resulting probability $P''_{\text{IBS}}(L)$, the difference between $P'_{\text{IBS}}(L)$ and $P''_{\text{IBS}}(L)$ would be on the order of the miniscule difference between $P_{\text{IBS}}(L)$ and $P_{\text{IBS}}(1.038 \times L)$. Neglecting this second effect, we replace $P_{\text{IBS}}(L)$ with $0.999 \times P_{\text{IBS}}(L) + 0.001 \times P'_{\text{IBS}}(L)$ for the purpose of inferring demography from human data.

2.4.4 Data analysis

For human demographic inference, we used the European and Yoruban parents who were sequenced at high coverage and phased with the help of their children by the 1000 Genomes pilot project (1000 Genomes Project 2010). We generated a set of IBS tract lengths from each of the six pairwise alignments between distinct CEU haplotypes, excising centromeres, telomeres, and other gaps annotated in the UCSC Genome Browser. To enable comparison of this spectrum with the spectrum of shared IBS tracts in the low coverage pilot data, we also excised regions that were inaccessible to the low coverage mapping or contained conspicuously few SNP calls in the low coverage data (see supporting information). The IBS tracts shared in the remaining parts of the genome were pooled to generate a spectrum of IBS sharing within the CEU population. The same regions were used to find the IBS tract shared within the six pairwise alignments of YRI haplotypes, as well as within the 16 pairwise alignments between a CEU haplotype and YRI haplotype.

Because of our interest in comparing our method to the closely related method of Li & Durbin (2011), we used the same mutation and recombination rates used in that paper ($\mu = 2.5 \times 10^{-8}$ mutations per base per generation; $\rho = 1.0 \times 10^{-8}$ recombinations per base per generation), as well as the same generation time (25 years).

2.4.5 Block bootstrapping

We performed block bootstrapping on IBS tract sharing within the CEU population by resampling large blocks, with replacement, from the 1.53×10^{10} base pairs of pairwise alignment data that were obtained by matching CEU parental haplotypes with each other. We did this by partitioning the total pool of CEU-CEU sequence alignment into 100 nonoverlapping regions that were each approximately 1.53×10^8 base pairs long. These regions were drawn with their boundaries at polymorphic sites so that no IBS tracts were broken up and divided between two blocks. By necessity, most blocks contain pieces of more than one continuous chromosomal region, but each is taken from a single pair of individuals. Each of the blue IBS tract length spectra from Figure 2.4 was created by sampling 100 blocks uniformly at random with replacement and recording the IBS tract lengths found within these blocks. The same procedure was used to sample from the distributions of tract lengths within the YRI population and between the CEU-YRI populations. Because the amount of pairwise CEU-YRI alignment totaled 4.06×10^{10} base pairs, the blocks of sequence alignment sampled from between populations were each approximately 4.06×10^8 base pairs long.

Chapter 3

Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach

Originally published as: S. Sheehan*, K. Harris* and Y.S. Song. (2013) "Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach." *Genetics* 194 (3) 647–662.

* These authors contributed equally to this work

Recently, Li and Durbin developed a coalescent-based hidden Markov model, called PSMC, to estimate changes in population size over time from a single pair of chromosomes (or one diploid individual). This is an efficient, useful approach, but its accuracy in the very recent past is hampered by the fact that, because of the small sample size, only few coalescence events occur in that period. Multiple genomes from the same population contain more information about the recent past, but are also more computationally challenging to study jointly in a coalescent framework. Here, we present a new coalescent-based method that can efficiently infer population size changes from multiple genomes, providing access to a new store of information about the recent past. Our work generalizes the recently developed sequentially Markov conditional sampling distribution framework, which provides an accurate approximation of the probability of observing a newly sampled haplotype given a set of previously sampled haplotypes. Simulation results demonstrate that we can accurately reconstruct the true population histories, with a significant improvement over the PSMC in the recent past. We apply our method, called *diCal* (Demographic Inference using Composite Approximate Likelihood), to the genomes of multiple human individuals of European and African ancestry to obtain a detailed population size change history during recent times.

3.1 Introduction

In the previous chapter, we saw that inferring demography from whole genomes is a difficult problem in general, and that it is particularly challenging to infer population size changes that occurred very recently. Information about recent demographic history comes from rare alleles and long tracts of shared IBS, which are both less abundant and more vulnerable to sequencing error than the common alleles and shorter IBS tracts that are informative about the ancient past.

Full-likelihood methods are better suited to coping with sequencing errors than methods that rely on summary statistics such as the SFS or the IBS tract length spectrum. This is because complex patterns of missing data can be incorporated directly into a likelihood function and do not have to be over-simplified in a probabilistic manner. Despite this theoretical advantage, existing likelihood methods have not resolved many questions surrounding either recent human population expansions or the accompanying migrations of *Drosophila* fruit flies out of Africa (Haddrill et al. 2005; Thornton & Andolfatto 2006; Wang & Hey 2010). The method PSMC of Li & Durbin (2011) produces reasonably accurate population size estimates overall, but its accuracy in the very recent past is hampered by the fact that, because of the small sample size, few coalescence events occur in that period. As a consequence, the information in the pattern of genetic variation for a pair of sequences is insufficient to resolve very recent demographic history.

The major obstacle to generalizing the PSMC to multiple sequences is the explosion in the state space with the number of sequences; the number of distinct coalescent tree topologies grows super-exponentially with the number of leaves, and we furthermore need to consider edge-weighted trees (i.e., include time information). In a related line of research, interesting progress has been made (Hobolth et al. 2007; Dutheil et al. 2009; Mailund et al. 2011) on performing "ancestral population genomic" inference under a coalescent HMM, but its applicability is limited to only a modest number of sequences, again due to the explosion in the state space.

In this paper, we describe an alternative method that is efficient in the number of sequences, while retaining the key generality of the PSMC in incorporating an arbitrary piecewise constant population size history. More precisely, the computational complexity of our method depends quadratically in the number of sequences, and the computation involved can be easily parallelized. As more sequences are considered, we expect to see a larger number of coalescence events during the recent past and should be able to estimate recent population sizes at a higher resolution. With only two sequences, the distribution of coalescence events is shifted toward the ancient past, relative to the distribution of the time a new lineage joins a coalescent tree for multiple sequences. Thus, even if all sequences are considered pairwise, the resolution in the recent past may not be as clear as that achieved by jointly modeling multiple sequences.

The input to our method, which is also based on an HMM, is a collection of haplotype sequences. At present, our method assumes that mutation and recombination rates are given, and it employs the expectation-maximization (EM) algorithm to infer a piecewise constant history of population sizes, with an arbitrary number of change points.

Our work generalizes the recently developed sequentially Markov conditional sampling distribution (SMCSD) framework (Paul et al. 2011) to incorporate variable population size. This approach provides an accurate approximation of the probability of observing a newly sampled haplotype given a set of previously sampled haplotypes, and it allows one to approximate the joint probability of an arbitrary number of haplotypes. Through a simulation study, we demonstrate that we can accurately reconstruct the true population histories, with a significant improvement over the PSMC in the recent past. Moreover, we apply our method to the genomes of multiple human individuals of European and African ancestry to obtain a detailed population size change history during recent times. Our software, called *diCal* (Demographic Inference using Composite Approximate Likelihood), is publicly available at http://sourceforge.net/projects/dical/.

3.2 Notation and a review of the SMCSD framework

Our work stems from the SMCSD framework (Paul et al. 2011), which describes the conditional genealogical process of a newly sampled haplotype given a set of previously sampled haplotypes. In what follows, we briefly review the key concepts underlying the SMCSD model.

We consider haplotypes each of length L from the same genomic region. Suppose we have already observed n haplotypes $\mathcal{O}_n = \{h_1, \dots, h_n\}$ sampled at random from a well-mixed population; note that some of the observed haplotypes may be identical. In this paper, we use the terms "sites" and "loci" interchangeably. Recombination may occur between any pair of consecutive loci, and we denote the set of potential recombination breakpoints by $B = \{(1,2), \dots, (L-1, L)\}$. Given a haplotype h, we denote by $h[\ell]$ the allele at locus ℓ, and by $h[\ell : \ell']$ (for $\ell \leq \ell'$) the subsequence $(h[\ell], \dots, h[\ell'])$.

As described in Paul & Song (2010), given the genealogy $\mathcal{A}_{\mathcal{O}_n}$ for the already observed sample $\mathcal{O}_n$, it is possible to sample a *conditional genealogy* $\mathcal{C}$ for the additional haplotype according to the following description: An ancestral lineage in $\mathcal{C}$ undergoes mutation at locus ℓ at rate $\theta_\ell/2$ according to the stochastic mutation transition matrix $\mathbf{P}^{(\ell)}$. Further, as in the ordinary coalescent with recombination, an ancestral lineage in $\mathcal{C}$ undergoes recombination at breakpoint $b \in B$ at rate $\rho_b/2$, giving rise to two lineages. Each pair of lineages within $\mathcal{C}$ coalesce with rate 1, and lineages in $\mathcal{C}$ get *absorbed* into the known genealogy $\mathcal{A}_{\mathcal{O}_n}$ at rate 1 for each pair of lineages. See Figure 3.2.1 for an illustration.

Unfortunately, we do not generally have access to the true genealogy $\mathcal{A}_{\mathcal{O}_n}$, and marginalizing over all possibilities is a challenging problem. However, Paul & Song (2010) showed that the diffusion-generator approximation described in De Iorio & Griffiths (2004a,b); Griffiths et al. (2008) implies the following approximation to $\mathcal{A}_{\mathcal{O}_n}$ which simplifies the problem considerably:

3.2.1 Approximation 1 (The trunk genealogy):

Approximate $\mathcal{A}_{\mathcal{O}_n}$ by the so-called *trunk* genealogy $\mathcal{A}^*_{\mathcal{O}_n}$ in which lineages do not mutate, recombine, or coalesce with one another, but instead form a non-random "trunk" extending infinitely into the past, as illustrated in Figure 3.2.1. Although $\mathcal{A}^*_{\mathcal{O}_n}$ is not a proper genealogy, it is still possible to sample a well-defined conditional genealogy $\mathcal{C}$ for the additional haplotype given $\mathcal{A}^*_{\mathcal{O}_n}$ in much the same way as described above, except that rates need to be modified. Specifically, lineages within $\mathcal{C}$ evolve backwards in time subject to the following events:

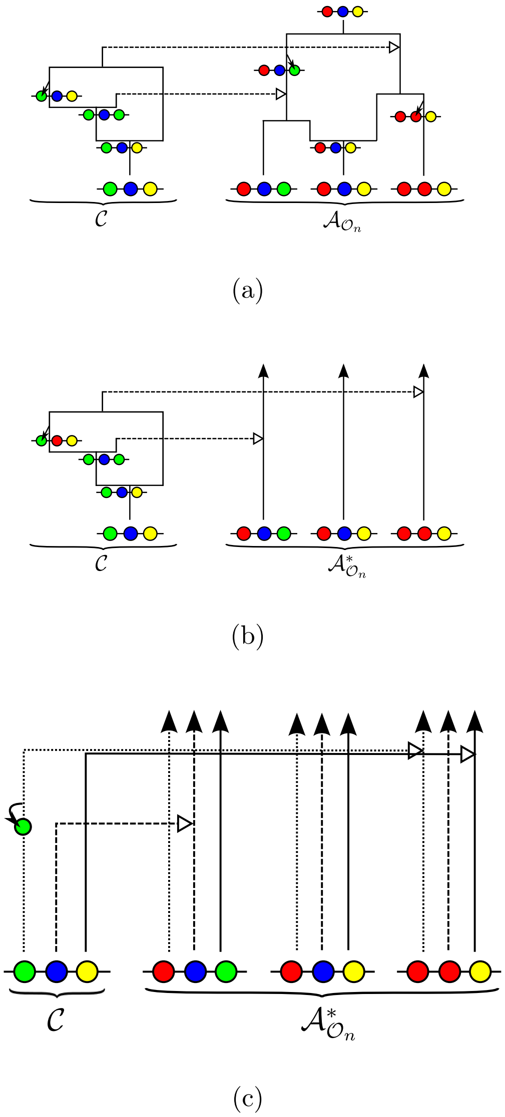
- *Mutation*: Each lineage undergoes mutation at locus ℓ with rate θ_ℓ according to $\mathbf{P}^{(\ell)}$.
- *Recombination*: Each lineage undergoes recombination at breakpoint $b \in B$ with rate ρ_b.
- *Coalescence*: Each pair of lineages coalesce with rate 2.
- *Absorption*: Each lineage is absorbed into a lineage of $\mathcal{A}^*_{\mathcal{O}_n}$ with rate 1.

The genealogical process described above completely characterizes a conditional sampling distribution (CSD), which Paul & Song (2010) denoted by $\hat{\pi}_{\text{PS}}$. Observe that the rate of absorption is the same as before, but the rates for mutation, recombination, and coalescence are each a factor of two larger than that mentioned earlier. Intuitively, this rate adjustment accounts for using the (inexact) trunk genealogy $\mathcal{A}^*_{\mathcal{O}_n}$, which remains static. Note that the adjustment follows as a mathematical consequence of the diffusion-generator approximation (De Iorio & Griffiths 2004a,b; Griffiths et al. 2008), and it is supported by the fact that the CSD $\hat{\pi}_{\text{PS}}$ has been shown to be exact for a one-locus model with parent-independent mutation (Paul & Song 2010).

It can be deduced from the diffusion-generator approximation that $\hat{\pi}_{\text{PS}}(\alpha \mid \mathcal{O}_n)$, the conditional probability of sampling an additional haplotype α given a set of previously sampled haplotypes $\mathcal{O}_n$, satisfies a recursion. Unfortunately, this recursion is computationally intractable to solve for even modest-sized data sets. To address this issue, Paul et al. (2011) proposed further approximations, described below, to obtain a CSD that admits efficient implementation, while retaining the accuracy of $\hat{\pi}_{\text{PS}}$.

3.2.2 Approximation 2 (Sequentially Markov CSD):

A given conditional genealogy $\mathcal{C}$ contains a *marginal conditional genealogy* (MCG) for each locus, where each MCG comprises a series of mutation events and the eventual absorption into a lineage of the trunk $\mathcal{A}^*_{\mathcal{O}_n}$. See Figure 3.2.1 for an illustration. The key insight (Wiuf & Hein 1999) is that we can generate the conditional genealogy as a *sequence* of MCGs across the sequence, rather than backwards in time. Although the sequential process is actually not Markov, it is well approximated (McVean & Cardin 2005; Marjoram & Wall 2006; Paul et al. 2011) by a Markov process using a two-locus transition density. Applying this approximation to $\hat{\pi}_{\text{PS}}$ yields the *sequentially Markov CSD* $\hat{\pi}_{\text{SMC}}$.


$\mathcal{C}$
$\mathcal{A}_{\mathcal{O}_n}$
$\mathcal{C}$
$\mathcal{A}^*_{\mathcal{O}_n}$
$\mathcal{C}$
$\mathcal{A}^*_{\mathcal{O}_n}$

(a)

(b)

(c)

Figure 3.1: Illustration of a conditional genealogy $\mathcal{C}$ for a three-locus model. The three loci of a haplotype are each represented by a solid circle, with the color indicating the allelic type at that locus. Mutation events, along with the locus and resulting haplotype, are indicated by small arrows. Recombination events, and the resulting haplotype, are indicated by branching events. Absorption events are indicated by dotted horizontal lines. (a) The true genealogy $\mathcal{A}_{\mathcal{O}_n}$ for the already observed sample $\mathcal{O}_n$. (b) Approximation by the trunk genealogy $\mathcal{A}^*_{\mathcal{O}_n}$; lineages in the trunk do not mutate, recombine, or coalesce. (c) *Marginal conditional genealogy* for each locus.

Conditional on the MCG $\mathcal{C}_{\ell-1}$ at locus $\ell - 1$, the MCG $\mathcal{C}_\ell$ at locus ℓ can be sampled by first placing recombination events onto $\mathcal{C}_{\ell-1}$ according to a Poisson process with rate $\rho_{(\ell-1,\ell)}$.

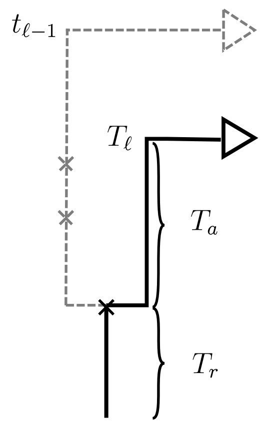


Figure 3.2: Illustration of the sequentially Markov approximation in which the absorption time T_ℓ at locus ℓ is sampled conditionally on the absorption time $T_{\ell-1} = t_{\ell-1}$ at the previous locus. In the marginal conditional genealogy $\mathcal{C}_{\ell-1}$ for locus $\ell - 1$, recombination breakpoints are realized as a Poisson process with rate $\rho_{(\ell-1,\ell)}$. If no recombination occurs, $\mathcal{C}_\ell$ is identical to $\mathcal{C}_{\ell-1}$. If recombination does occur, as in the example here, $\mathcal{C}_\ell$ is identical to $\mathcal{C}_{\ell-1}$ up to the time T_r of the most recent recombination event. At this point, the lineage at locus ℓ, independently of the lineage at locus $\ell - 1$, proceeds backwards in time until being absorbed into a lineage of the trunk. The absorption time at locus ℓ is $T_\ell = T_r + T_a$, where T_a is the remaining absorption time after the recombination event.

If no recombination occurs, $\mathcal{C}_\ell$ is identical to $\mathcal{C}_{\ell-1}$. If recombination does occur, $\mathcal{C}_\ell$ is identical to $\mathcal{C}_{\ell-1}$ up to the time T_r of the most recent recombination event. At this point, the lineage at locus ℓ, independently of the lineage at locus $\ell-1$, proceeds backwards in time until being absorbed into a lineage of the trunk. This transition mechanism for the Markov process is illustrated in Figure 3.2. McVean & Cardin (2005) use this approximation as well, while the transition process in Marjoram & Wall (2006) *does* allow the lineage to coalesce back into itself.

Given $\mathcal{C}_\ell$, mutations are superimposed onto it according to a Poisson process with rate θ_ℓ. The MCG is absorbed into a trunk lineage corresponding to some haplotype h, which specifies an "ancestral" allele $h[\ell]$. This allele is then propagated to the present according to the superimposed mutations and the transition matrix $\mathbf{P}^{(\ell)}$, thereby generating an allele at locus ℓ of the additional haplotype α. We refer to the associated distribution of alleles as the emission distribution.

The generative process described above for the SMCSD $\hat{\pi}_{\text{SMC}}$ can be formulated as an HMM, in which the hidden state at locus ℓ corresponds to the MCG $\mathcal{C}_\ell$ excluding mutation events: we denote the hidden state at locus ℓ in the HMM by $S_\ell = (T_\ell, H_\ell)$, where $T_\ell \in [0, \infty)$ is the absorption time and $H_\ell \in \mathcal{O}_n$ is the absorption haplotype. The emission at locus ℓ corresponds to the allele $\alpha[\ell]$. See Paul et al. (2011) for explicit expressions for the initial, transition, and emission densities in the case of a constant population size.

3.3 Incorporating variable population size

Here, we extend the SMCSD framework described in the previous section to incorporate variable population size. A history of relative effective population size is described by the function

$$\lambda(t) = \frac{N(t)}{N_{\text{ref}}}, \tag{3.1}$$

where $t \in [0, \infty)$, with $t = 0$ corresponding to the present time, N_{ref} is some reference effective population size, and $N(t)$ is the effective population size at time t in the past. The population-scaled recombination and mutation rates are defined with respect to N_{ref}. Specifically, for $b = (\ell - 1, \ell)$, we define $\rho_b = 4N_{\text{ref}}r_b$, where r_b denotes the recombination rate per generation per individual between loci $\ell - 1$ and ℓ; and $\theta_\ell = 4N_{\text{ref}}\mu_\ell$, where μ_ℓ denotes the mutation rate per generation per individual at locus ℓ.

3.3.1 Initial density:

In the case of a constant population size, the absorption time T_ℓ for locus ℓ follows an exponential distribution, but with a variable population size the absorption time is described by a non-homogeneous Markov chain. See Griffiths & Tavaré (1994) for a more thorough discussion of the coalescent with variable population size. As in the constant population size case, however, the prior distribution of absorption haplotype H_ℓ is still uniform over the observed haplotypes $\mathcal{O}_n$ in the trunk genealogy. In summary, the marginal density of the hidden state $s_\ell = (t, h)$ is given by

$$\zeta^{(\lambda)}(t, h) = \frac{n_h}{\lambda(t)} \exp\left(-n \int_0^t \frac{1}{\lambda(\tau)} d\tau\right), \tag{3.2}$$

where n_h denotes the number of haplotypes in $\mathcal{O}_n$ that are identical to haplotype h.

3.3.2 Transition density:

To obtain the transition density, we need to take into account recombination, which causes changes in the hidden state of our HMM. If no recombination occurs between loci $\ell - 1$ and ℓ (prior to $T_{\ell-1}$), then $s_\ell = s_{\ell-1}$. If a recombination event occurs between loci $\ell - 1$ and ℓ, the absorption time for locus ℓ will be $T_\ell = T_r + T_a$, where T_r is the time of recombination (which must be less than $T_{\ell-1}$ and T_ℓ) and T_a is the remaining additional time to absorption, as illustrated in Figure 3.2. To compute the transition density, we need to convolve the hidden variables T_r and T_a. Letting $b = (\ell - 1, \ell)$ for ease of notation, the transition density from $s_{\ell-1} = (t, h)$ to $s_\ell = (t', h')$ is given by

$$\phi^{(\lambda)}(s_\ell | s_{\ell-1}) = e^{-\rho_b t} \cdot \delta_{s_{\ell-1}, s_\ell} + \int_0^{\min(t,t')} \rho_b e^{-\rho_b t_r} \left[\frac{\zeta^{(\lambda)}(t', h')}{\int_{t_r}^\infty \zeta^{(\lambda)}(\tau) d\tau}\right] dt_r, \tag{3.3}$$

where $\zeta^{(\lambda)}(t', h')$ is defined in (3.2) and $\zeta^{(\lambda)}(\tau) := \sum_{h \in \mathcal{O}_n} \zeta^{(\lambda)}(\tau, h)$. Note that $\int_0^\infty \zeta^{(\lambda)}(\tau)\, d\tau = 1$.

3.3.3 Emission probability:

The probability of emitting allele a at locus ℓ (i.e., $\alpha[\ell] = a$) given hidden state $s_\ell = (t, h)$ is

$$\xi^{(\lambda)}(a|s_\ell) = e^{-\theta_\ell t} \sum_{m=0}^{\infty} \frac{1}{m!} (\theta_\ell t)^m \left[(\mathbf{P}^{(\ell)})^m\right]_{h[\ell],a}.$$

This is the same emission probability as in Paul et al. (2011), but when we discretize the state space in the following section we will have to take into account the effects of variable population size.

3.3.4 The sequentially Markov conditional sampling probability:

Using the initial, transition, and emission densities described above, we can write down an integral recursion for the forward probability $f^{(\lambda)}_{\text{SMC}}(\alpha[1:\ell], s_\ell)$ of observing the first ℓ alleles $\alpha[1], \dots, \alpha[\ell]$ and the state at locus ℓ being s_ℓ. For $2 \le \ell \le L$,

$$f^{(\lambda)}_{\text{SMC}}(\alpha[1:\ell], s_\ell) = \xi^{(\lambda)}(\alpha[\ell] \mid s_\ell) \cdot \int \phi^{(\lambda)}(s_\ell|s_{\ell-1}) f^{(\lambda)}_{\text{SMC}}(\alpha[1:\ell-1], s_{\ell-1})\, ds_{\ell-1}, \tag{3.4}$$

with base case

$$f^{(\lambda)}_{\text{SMC}}(\alpha[1], s_1) = \xi^{(\lambda)}(\alpha[1] \mid s_1) \cdot \zeta^{(\lambda)}(s_1).$$

Finally, the conditional probability of sampling an additional haplotype α having previously observed $\mathcal{O}_n = \{h_1, \cdots, h_n\}$ is given by

$$\hat{\pi}^{(\lambda)}_{\text{SMC}}(\alpha \mid \mathcal{O}_n) = \int f^{(\lambda)}_{\text{SMC}}(\alpha[1:L], s_L)\, ds_L. \tag{3.5}$$

As with the constant population size HMM, a backward algorithm can also be devised to compute $\hat{\pi}^{(\lambda)}_{\text{SMC}}(\alpha \mid \mathcal{O}_n)$, though we do not present it here.

3.4 Discretizing the state space

To efficiently evaluate the recursion (3.4) and the marginalization (3.5), we discretize the time component of the state space. We partition time (in units of $2N_{\text{ref}}$ generations) into d intervals, demarcated by

$$t_0 = 0 < t_1 < \cdots < t_d = \infty,$$

and assume that $\lambda(t)$ defined in (3.1) has a constant value λ_i in each interval $D_i := [t_{i-1}, t_i)$, for $i = 1, \dots, d$:

$$\lambda(t) = \sum_{i=1}^{d} \mathbf{1}(t_{i-1} \le t < t_i)\lambda_i, \tag{3.6}$$

where $\mathbf{1}(\cdot)$ is the indicator function. Using this piecewise constant $\lambda(t)$, we can write the HMM probabilities in a more workable form, as detailed below.

3.4.1 Initial probability:

For $t \in D_i$, (3.6) implies that the initial density (3.2) can be written as

$$\zeta^{(\lambda)}(t,h) = \frac{n_h}{\lambda_i} e^{-n(t-t_{i-1})/\lambda_i} \prod_{j=1}^{i-1} e^{-n(t_j-t_{j-1})/\lambda_j}.$$

To obtain the initial probability in the time-discretized model, we integrate over the time interval D_i to obtain

$$\hat{\zeta}^{(\lambda)}(D_i,h) = \int_{D_i} \zeta^{(\lambda)}(t,h)\, dt = \frac{n_h}{n} w^{(i)}, \tag{3.7}$$

where

$$w^{(i)} = \left[1 - e^{-n(t_i-t_{i-1})/\lambda_i}\right] \prod_{m=1}^{i-1} e^{-n(t_m-t_{m-1})/\lambda_m},$$

which corresponds to the probability that a lineage in the conditional genealogy gets absorbed into the trunk genealogy within the interval D_i.

3.4.2 Transition probability:

For the transition density from state $s_{\ell-1} = (t,h)$ to state $s_\ell = (t',h')$, we let i denote the time interval index such that $t \in D_i = [t_{i-1}, t_i)$ and let j denote the index such that $t' \in D_j = [t_{j-1}, t_j)$. After some simplification, the transition density (3.3) becomes

$$\phi^{(\lambda)}(s_\ell \mid s_{\ell-1}) = e^{-\rho_b t} \cdot \delta_{s_{\ell-1},s_\ell} + \frac{n_h}{\lambda_j} e^{-n(t'-t_{j-1})/\lambda_j} \left[\prod_{m=1}^{j-1} e^{-n(t_m-t_{m-1})/\lambda_m}\right] R(i,t;j,t'), \tag{3.8}$$

where $R(i,t;j,t')$ is defined in the Appendix.

To compute the transition probability in the time-discretized model, we use Bayes' rule and integrate the transition density function to obtain

$$\begin{aligned} \hat{\phi}^{(\lambda)}(D_j, h' \mid D_i, h) &= \frac{1}{\hat{\zeta}^{(\lambda)}(D_i,h)} \int_{D_j} \int_{D_i} \phi^{(\lambda)}(t',h' \mid t,h) \zeta^{(\lambda)}(t,h)\, dt\, dt' \\ &=: y^{(i)} \cdot \delta_{i,j} \delta_{h,h'} + z^{(i,j)} \cdot \frac{n_{h'}}{n}, \end{aligned}$$

where $\hat{\zeta}^{(\lambda)}(D_i,h)$ is defined in (3.7), and explicit formulas for $y^{(i)}$ and $z^{(i,j)}$ are provided in the Appendix. The first term arises from the case of no recombination, while the second term accounts for the case when recombination does occur. Note that $y^{(i)}$ and $z^{(i,j)}$ depend only on the time interval, not on the absorbing haplotype.

3.4.3 Emission probability:

Although thus far the emission density has not been affected by the population size being variable, discretizing time introduces dependence on the function $\lambda(t)$. Let a denote the emitted allele of the newly sampled haplotype α at locus ℓ. Using Bayes' rule again and then integrating over the absorption time interval gives

$$\hat{\xi}^{(\lambda)}(a|D_i, h) = \frac{1}{\hat{\zeta}^{(\lambda)}(D_i, h)} \int_{D_i} \xi^{(\lambda)}(a|t, h)\zeta^{(\lambda)}(t, h)\ dt = \sum_{m=0}^{\infty} v^{(i)}(m) \cdot [(\mathbf{P}^{(\ell)})^m]_{h[\ell],a}, \qquad (3.10)$$

where a formula for $v^{(i)}(m)$ is provided in the Appendix.

3.4.4 Discretizing time and grouping parameters:

To discover periods of population expansion or contraction with the SMCSD, it is necessary to specify a time discretization that has high resolution during such time periods. This is challenging in cases where we have little *a priori* knowledge of the demographic history. Ideally the (unknown) coalescence events would be distributed uniformly across the time intervals of our discretization; if very few coalescence events occur in an interval, the corresponding population size will often be overestimated, leading to run-away behavior. In our implementation, we employ a heuristic method, detailed in the Appendix, for choosing the discretization time points $t_1, \ldots, t_{d-1}$ based on the spacing of SNPs in the data, with the aim for coalescence events to be distributed evenly across the d time intervals. Alternatively, the user has the option of specifying their own discretization time points to achieve a desired resolution.

As noted by Li & Durbin (2011), allowing separate population size parameters during time intervals that contain too few expected coalescence events can lead to inaccurate estimates. Following their lead, we mitigate this problem by constraining a few consecutive time intervals to have the same population size.

3.5 Modifying the trunk genealogy

The trunk genealogy approximation in Paul & Song (2010) was derived by making an approximation in the diffusion process dual to the coalescent for a constant population size. It yields an accurate approximate CSD in the case of a population at equilibrium, and for parent-independent mutation models, the CSD becomes exact in the limit as the recombination rate approaches ∞. However, in the variable population size setting, we must modify the trunk genealogy approximation for the following reason: In the formulation described earlier, the trunk absorbs a lineage in the conditional genealogy $\mathcal{C}$ at the rate $ndt/\lambda(t)$ at time t. Our HMM uses this inverse dependence and the inferred distribution of absorption times to estimate the population size scaling function $\lambda(t)$. In reality, at time t the number of ancestral lineages is $n(t) \leq n$ and a lineage in $\mathcal{C}$ gets absorbed at rate $n(t)dt/\lambda(t)$.

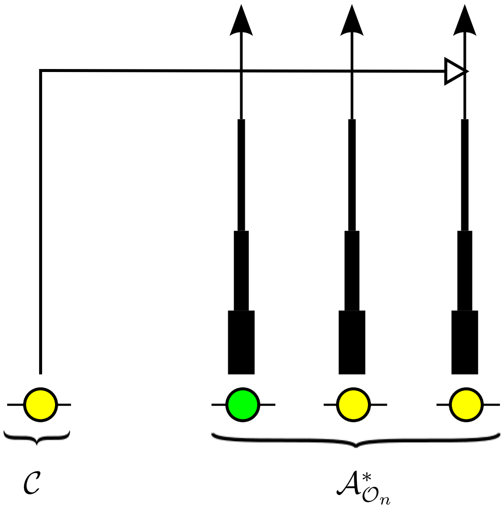


Figure 3.3: Illustration of the wedding cake genealogy approximation, in which the varying thickness of a lineage in $\mathcal{A}^*_{\mathcal{O}_n}$ schematically represents the amount of contribution to the absorption rate. As the figure depicts, the wedding cake genealogy never actually loses any of the n lineages, and absorption into any of the n lineages is allowed at all times; we are only modifying the absorption rate as a function of time.

Hence, assuming that the trunk genealogy contains n lineages in every time interval causes absorption events to occur too quickly, leaving the ancient population sizes over-estimated. We later provide empirical results which support this intuition (see Figure 3.8).

To remedy the problem described above, in our work we use the expected number of lineages in the trunk to modify the rate of absorption, while still forbidding mutation, recombination, and coalescence in the trunk genealogy. Let $A_n(t)$ denote the number of lineages at time t ancestral to a sample of size n at time 0. Under the coalescent, the probability distribution of $A_n(t)$ is known in closed form (Tavaré 1984), but using it directly to compute the expected number of lineages leads to numerically unstable results, due to alternating signs. However, one can obtain the following expression for the expectation (Tavaré 1984, equation 5.11) which is numerically stable:

$$\overline{n}(t) := \mathbb{E}[A_n(t)] = \sum_{i=1}^{n} \exp\left[-\binom{i}{2}\int_0^t \frac{1}{\lambda(\tau)}d\tau\right] \frac{n(n-1)\cdots(n-i+1)}{n(n+1)\cdots(n+i-1)}(2i-1). \tag{3.11}$$

For simplicity, we assume that throughout time interval $D_i = [t_{i-1}, t_i)$, there are $\overline{n}(t_{i-1})$ lineages, creating what we call a "wedding cake genealogy," as illustrated in Figure 3.3.

To modify the HMM formulas, we simply replace each n in (3.7), (3.9), and (3.10) with the appropriate $\overline{n}(\cdot)$ from (3.11), except in the ratio n_h/n multiplying $w^{(i)}$ in (3.7) and the ratio $n_{h'}/n$ multiplying $z^{(i,j)}$ in (3.9) (these ratios are kept intact to preserve the relative contributions of different haplotypes). Note that the trunk genealogy never actually loses any of the n lineages, and absorption into any of the n lineages is allowed at all times; we are only modifying the absorption rate as a function of time. In the case of two sequences (one trunk lineage and one additionally sampled lineage), $\overline{n}(t) = 1$ for all t, so the wedding cake approximation does not change the model. Making the number of lineages more accurate using this approximation improves our ability to estimate absorption times and therefore population sizes.

3.6 Population size inference with Expectation-Maximization

To utilize all our data in an exchangeable way, we use a "leave-one-out" approach where we leave each haplotype out in turn and perform the SMCSD computation. More precisely, we define the leave-one-out composite likelihood (LCL) as

$$L_{\text{LCL}}(\lambda; h_1, \dots, h_n) = \prod_{i=1}^{n} \hat{\pi}_{\text{SMC}}^{(\lambda)}(h_i | h_1, \dots, h_{i-1}, h_{i+1}, \dots, h_n). \tag{3.12}$$

Because we compute the conditional sampling probability through dynamic programming and the probability depends on the effective population sizes in complex ways, we cannot find the maximum-likelihood estimates analytically. Although direct optimization could be used, it is computationally expensive. Thus we employ an expectation-maximization (EM) algorithm to estimate the piecewise constant function $\lambda(t)$. Our current implementation assumes that the population-scaled recombination rates ρ_b and mutation rates θ_ℓ, as well as the mutation transition matrices $\mathbf{P}^{(\ell)}$, are given and fixed. For computational simplicity we currently assume that θ_ℓ and $\mathbf{P}^{(\ell)}$ are the same for each site ℓ, and ρ_b is the same for each pair of consecutive sites. The time discretization is fixed throughout the EM algorithm. The output of the algorithm is an estimated population size scaling factor λ_i for each interval $D_i = [t_{i-1}, t_i)$. To convert these scaling factors into diploid effective population sizes, one would need to multiply by N_{ref}. Similarly, the discretization times can be converted to years by multiplying them by $2N_{\text{ref}}g$, where g is an average number of years per generation.

The standard Baum-Welch algorithm gives an EM procedure for learning the parameters of an HMM in which the transition probabilities and emission probabilities are treated as unknown independent parameters. However, our HMM is more constrained than a general one, with $(dn)^2 + d|\Sigma|^2$ (where Σ is the alphabet of alleles) unknown probabilities $\hat{\phi}^{(\lambda)}(D_j, h' \mid D_i, h)$ and $\hat{\xi}^{(\lambda)}(\alpha[\ell] \mid D_i, h)$ that are functions of the d parameters $\lambda_1, \dots, \lambda_d$. During the E-step, we compute the matrix $[A_{ij}]$ of the expected number of D_i to D_j transitions. We also compute the matrix $[E_i(b)]$ of the expected number of times allele b is emitted from time interval i. Then, during the M-step we maximize the likelihood function

$$(\lambda_1^{(k+1)}, \dots, \lambda_d^{(k+1)}) = \operatorname*{argmax}_{\lambda^{(k)}} \prod_{i,j} [\hat{\phi}^{(\lambda^{(k)})}(D_j | D_i)]^{A_{ij}^{(k)}} \prod_{i,b} [\hat{\xi}^{(\lambda^{(k)})}(b | D_i)]^{E_i^{(k)}(b)},$$

where $\hat{\phi}^{(\lambda)}(D_j | D_i)$ and $\hat{\xi}^{(\lambda)}(b | D_i)$ refer to the transition and emission probabilities where we have marginalized over the absorption haplotype.

We initialize the algorithm with $\lambda_i = 1$ for all $i = 1, \ldots, d$. To compute $[A_{ij}]$ and $[E_i(b)]$, we use the forward and backward probabilities of our HMM. The exact details of making this step computationally efficient are provided in the Appendix. After the E-step, we use the Nelder-Mead optimization routine (Flanagan 2010) to update the parameters in the M-step. Because of local maxima in the likelihood surface, we run this optimization routine several times (≈ 10) with different starting conditions and then retain the estimates with the largest likelihood. In the analysis discussed in this paper, we ran the EM procedure for 20 iterations to obtain convergence. As pointed out by Li & Durbin (2011), running the EM procedure for many iterations often leads to over-fitting.

3.7 Results

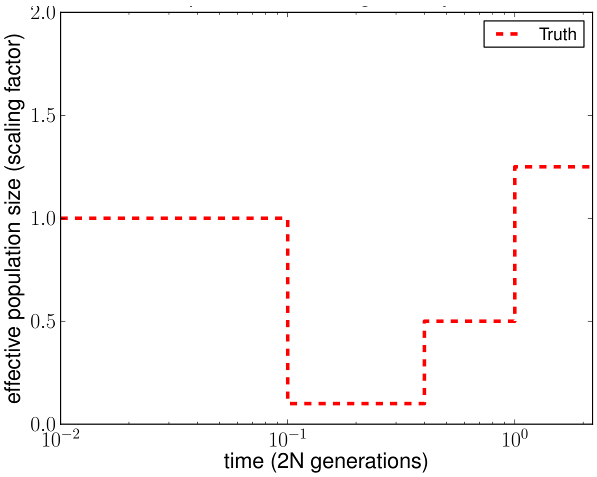
We compared the performance of our method, diCal, with that of PSMC (Li & Durbin 2011) on both simulated and real data. We compared diCal using an n-haplotype leave-one-out scheme (3.12) with PSMC using the same n haplotypes paired up sequentially (i.e. haplotype 1 paired with haplotype 2, haplotype 3 with haplotype 4, etc).

Unless stated otherwise, we used 16 discretization intervals and inferred 7 free population size parameters in both PSMC and diCal. In the notation introduced by Li & Durbin (2011), the pattern we used is "3+2+2+2+2+2+3," which means that the first parameter spans the first three discretization intervals, the second parameter spans the next two intervals, and so on. We found that grouping a few consecutive intervals to share the same parameter significantly improved the accuracy of estimates. For example, due to an insufficient number of coalescence events, the first and last intervals are particularly susceptible to runaway behavior if they are assigned their own free parameters, but grouping with their neighboring intervals prevented such pathological behavior. See the Supporting Information for further details of running PSMC and our method.

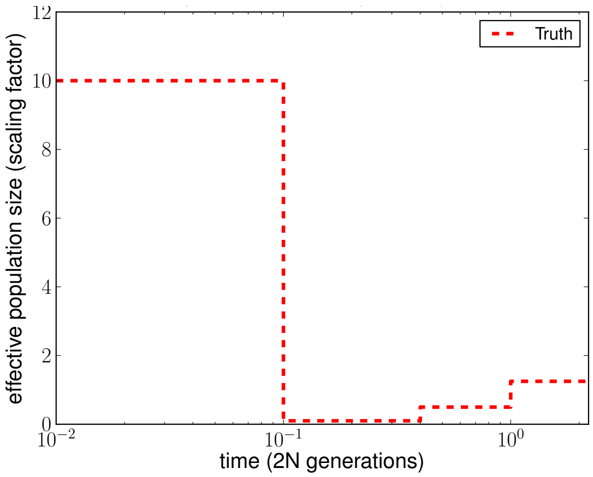
3.7.1 The accuracy of population size inference on simulated data:

We used `ms` (Hudson 2002) to simulate full ancestral recombination graphs (ARGs) under two different population histories, and then superimposed a quadra-allelic, finite-sites mutation process on the ARGs to generate sequence data over $\{A, C, G, T\}$. As illustrated in Figure 3.4, both histories contained bottlenecks in the moderately recent past. History S2 in addition contained a recent rapid population expansion relative to the ancient population size. For each history, we simulated 10 independent ARGs for $L = 10^6$ sites and 10 haplotypes, with the population-scaled recombination rate set to 0.01 per site in `ms`. To add mutations, we set the population-scaled mutation rate to 0.014 per site and used the quadra-allele mutation matrix described in the Supporting Information.

As shown in Figures 3.5 and 3.6, our method performed much better in the recent past than did PSMC. PSMC often had the type of runaway behavior shown in Figure 3.6, where it overestimated the most recent population size by over three orders of magnitude. We



(a)



(b)

Figure 3.4: Population size histories considered in our simulation study, with time $t = 0$ corresponding to the present. (a) History S1 containing a bottleneck. (b) History S2 containing a bottleneck followed by a rapid expansion.

Table 3.1: Goodness-of-fit for PSMC and diCal, averaged over 10 simulated data sets each with a sample of $n = 10$ haplotypes. The underlying population size histories are shown in Figure 3.4. The error metric used is a normalized integral of the absolute difference between the true history and the inferred history over time. These results demonstrate that diCal is substantially more accurate than the PSMC method.

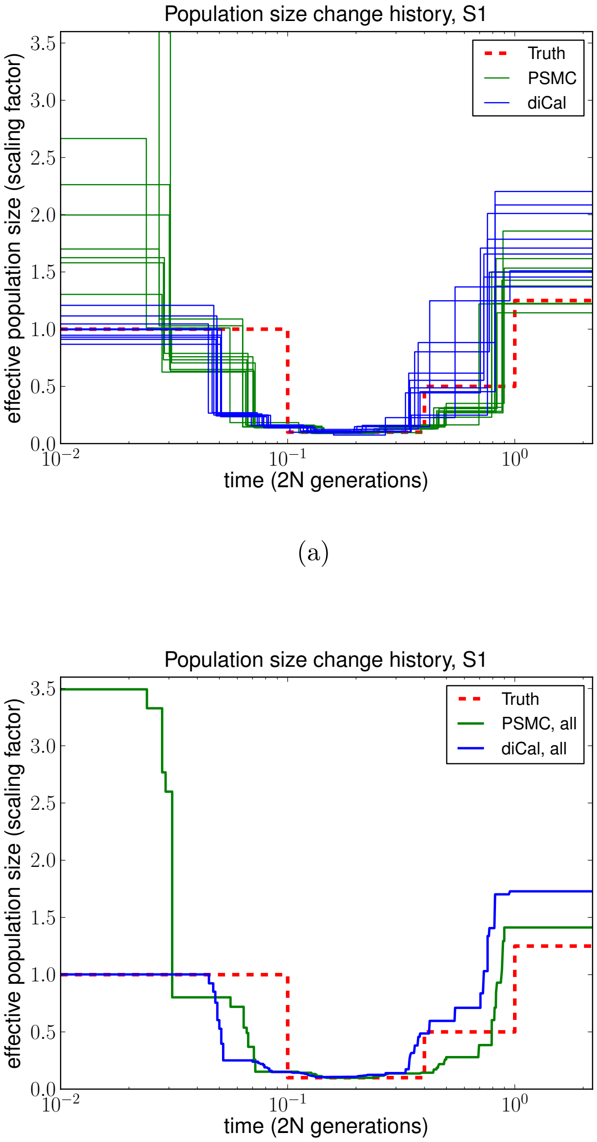
Simulated History	PSMC error	diCal error
S1	0.40328	0.10283
S2	0.71498	0.29992

note that our method began to lose accuracy for more ancient times, most likely because ancient absorption events in a 1 Mb region are few and sparsely distributed in time in the leave-one-out SMCSD computation. Both methods tend to smooth out sudden changes in population size, which is why the inferred recovery time from a bottleneck is more recent than it should be. To quantify the improvement in accuracy of our method over PSMC, we used an error metric described in Li & Durbin (2011), which is a normalized integral of the absolute difference between the true `ms` history and the inferred history over time. The results, summarized in Table 3.1, show that our method had a substantially lower overall error than PSMC.

For inference using diCal, we examined the impact of considering more haplotypes on the accuracy of population size estimation. In this study, we focused on history S1 and grouped adjacent parameters to fit roughly with population size change points for illustration purposes. Figure 3.7 shows qualitatively that increasing the sample size n makes our estimate of the recent population size more accurate. Intermediate sizes changed little with increasing n, and ancient sizes were somewhat variable depending on the distribution of coalescence events. Note that for $n = 2$, our method is very similar to PSMC; we compute the transition probabilities slightly differently, but the wedding cake approximation does not change the model in this case. We used the same error metric mentioned above to quantify the advantage of increasing the sample size. As shown in Table 3.2, the overall error decreased as the sample size increased, with improvement tapering around 8 to 10 haplotypes for this particular history.

3.7.2 The impact of the wedding cake genealogy approximation:

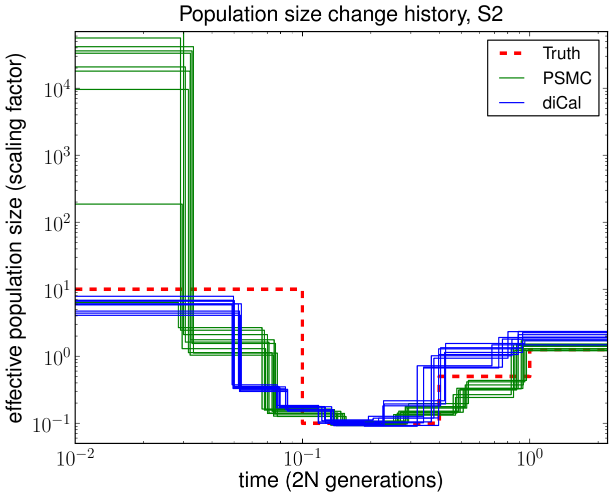
We examined the advantage of using the wedding cake genealogy approximation in the SMCSD computation, compared to assuming an unmodified trunk genealogy. Figure 3.8 illustrates that the unmodified trunk genealogy leads to overestimation of population sizes in the distant past, as discussed in Modifying the Trunk Genealogy. The wedding cake genealogy approximation, which adjusts the absorption rate by accounting for the expected number of ancestral lineages of the already observed sample, leads to a significant improvement in the accuracy of population size inference in the ancient past.



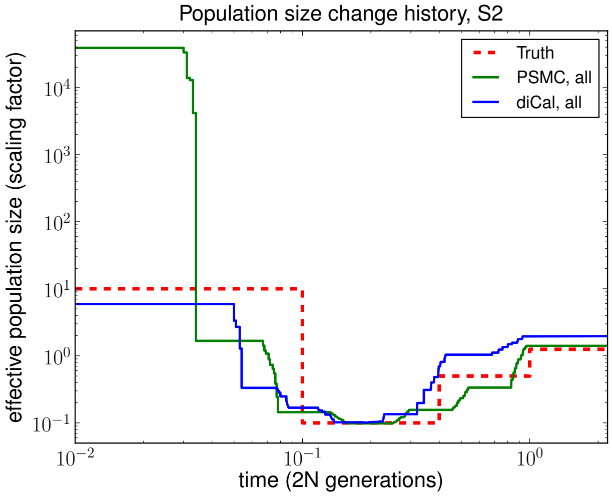
(a)

(b)

Figure 3.5: Results of PSMC and diCal on data sets simulated under history S1 with sample size $n = 10$ and four alleles (A,C,G,T). PSMC significantly overestimates the most recent population size, whereas we obtain good estimates up until the very ancient past. (a) Results for 10 different data sets. (b) Average over the 10 data sets.



(a)



(b)

Figure 3.6: Results of PSMC and diCal on data sets simulated under history S2 with sample size $n = 10$ and four alleles (A,C,G,T). The PSMC shows runaway behavior during the recent past, overestimating the most recent time by over three orders of magnitude on average. (a) Results for 10 different data sets. (b) Average over the 10 data sets.

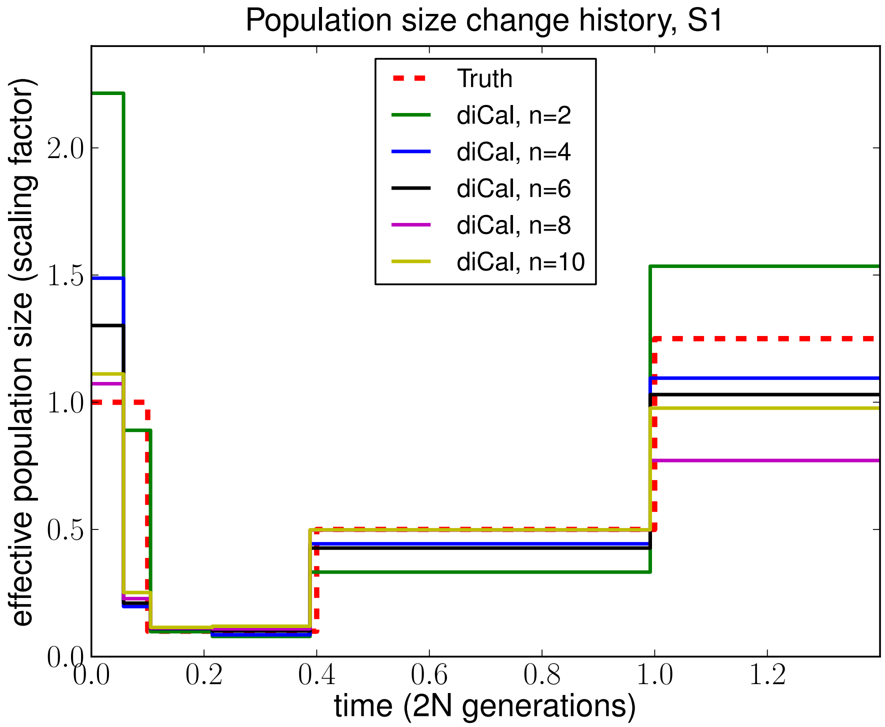


Figure 3.7: The effect of considering more haplotypes in diCal using the SMCSD-based leave-one-out likelihood approach. Data were simulated under population size history S1 with two alleles. In this study, we grouped adjacent parameters to fit roughly with population size change points for illustration purposes. This figure shows the increase in the accuracy of our method with an increasing sample size n. The recent sizes are the most dramatically affected, while intermediate sizes remain accurate even with few haplotypes.

Table 3.2: Goodness-of-fit for diCal on simulated bottlenecked history S1 for different sample sizes. We used the same error metric as in Table 3.1. As the sample size n increases, the error decreases, with global improvement tapering around 8 to 10 haplotypes.

Sample size n	diCal error
2	0.2914
4	0.1901
6	0.1446
8	0.0802
10	0.0899

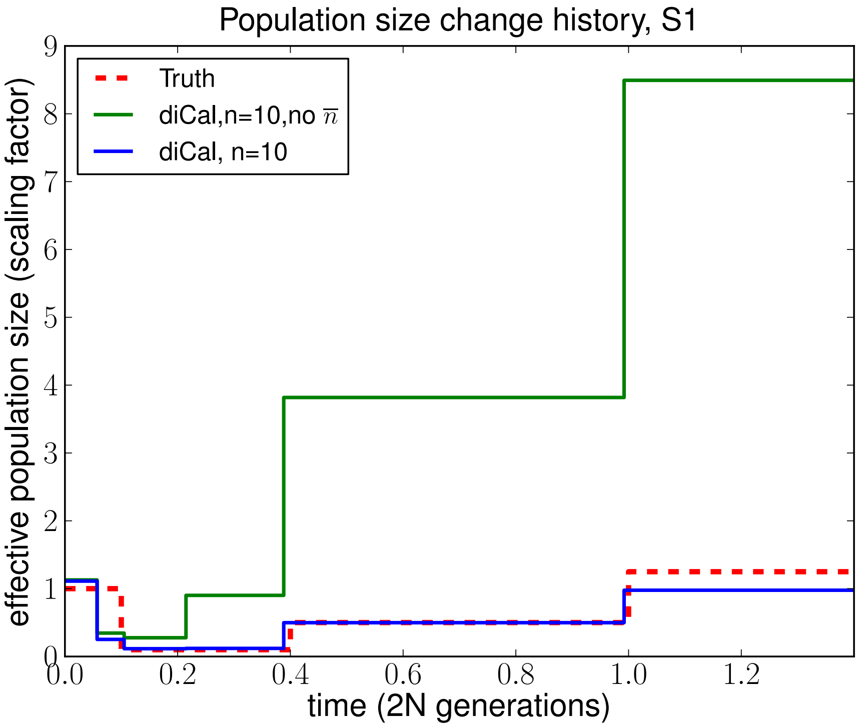


Figure 3.8: A comparison of the SMCSD-based leave-one-out likelihood approach in diCal using the wedding cake genealogy (blue line) with that using the unmodified trunk genealogy (green line). The results shown here are for $n = 10$ haplotypes simulated under history S1 with two alleles. Without the wedding cake genealogy, absorption of the left-out lineage into the trunk occurs too quickly, and the lack of absorption events in the mid to ancient past leads to substantial overestimation of the population sizes. Recent population sizes remain unaffected since during these times the absorption rates in the wedding cake genealogy and in the trunk genealogy are roughly the same. In this example, we grouped adjacent parameters to fit roughly with population size change points for illustration purposes.

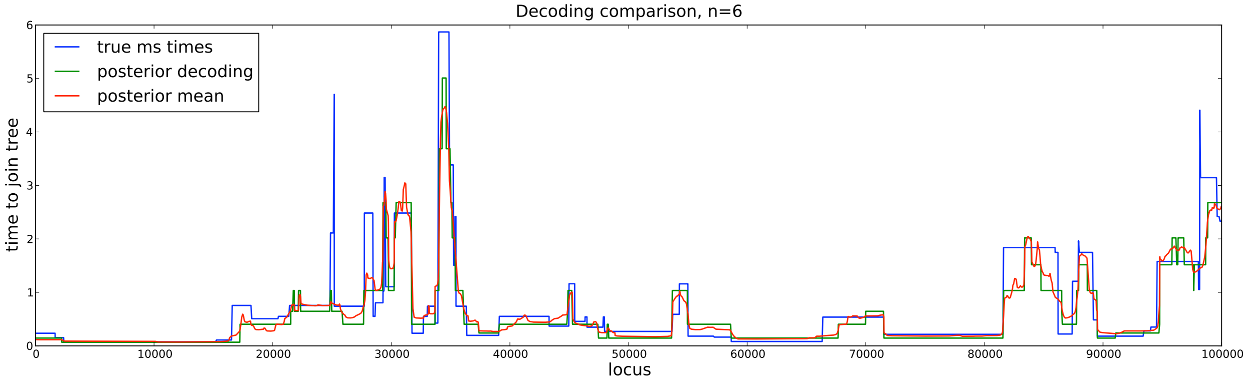


Figure 3.9: Estimated absorption times in diCal using the leave-one-out SMCSD method versus the true coalescence times for a 100 kb region. The data were simulated using ms for $n = 6$ haplotypes assuming a constant population size. The true coalescence time at each site, obtained from ms, was taken as the time the ancestral lineage of a left-out haplotype joined the rest of the coalescent tree at that site. The figure shows the true coalescence time for the nth haplotype and our corresponding inferred absorption times, obtained from the posterior decoding and the posterior mean. Our estimates generally track the true coalescence times closely.

3.7.3 The accuracy of estimated coalescence times:

To assess the accuracy of estimated coalescence times, we produced the posterior decoding and the posterior mean of the times that a left-out haplotype got absorbed into a wedding cake genealogy. The data were simulated under the full coalescent with recombination using ms assuming a constant population size. The true coalescence time at each site was taken as the time the left-out lineage joined the rest of the coalescent tree at that site. As shown in Figure 3.9, we found that our estimated absorption times closely tracked the true coalescence times.

3.7.4 Results on real data:

We applied our method to European (CEU) and African (YRI) subsamples from the 1000 Genomes Project (1000 Genomes Project 2012). To minimize potential confounding effects from natural selection, we chose a 3 Mb region on chromosome 1 with no genes and then used the middle 2 Mb for analysis. We used the human reference (version 36) to create a full multiple sequence alignment of 10 haplotypes (5 diploid individuals) for each of the CEU and YRI populations. Although we filtered out unphased individuals and sites, the final sequences are based on low-coverage short read data, so phasing and imputation errors could impact the accuracy of our coalescence time inference. We assumed a per-generation mutation rate of $\mu = 1.25 \times 10^{-8}$ per site, which is consistent with recent studies of *de novo* mutation in human trios (Roach et al. 2010; Awadalla et al. 2010; Kong et al. 2012), and a mutation transition matrix estimated from the human and the chimp reference genomes (shown in the Supporting Information). For simplicity, we assumed that the per-generation recombination rate r between consecutive bases is constant and equal to μ. The generation time was assumed to be 25 years. For a reference population size, we used $N_{\text{ref}} = 10,000$.

The results of PSMC and our method are shown in Figure 3.10. PSMC displayed runaway behavior and produced rather unrealistic results; see Figure 3.7.4, for which we truncated the y-axis at 40,000 for ease of comparison with Figure 3.7.4. The data set may be too small for PSMC to work accurately. We note that PSMC was able to produce more reasonable results on simulated data sets, probably because they were generated with much higher mutation and recombination rates, thus representing a larger genomic region for humans.

As shown in Figure 3.7.4, our method inferred that CEU and YRI had very similar histories in the distant past up until about 117 kya; discrepancies up to this point are most likely due to few observed ancient coalescence events with the leave-one-out approach. We inferred that the European population underwent a severe (out-of-Africa) bottleneck starting about 117 kya, with the effective population size dropping by a factor of about 12 from $\approx$ 28,000 to $\approx$ 2,250. Furthermore, at the level of resolution provided by our time discretization, our results suggest that the European population has recovered from the bottleneck to an average effective size of $\approx$ 12,500 for the past 16 thousand years.

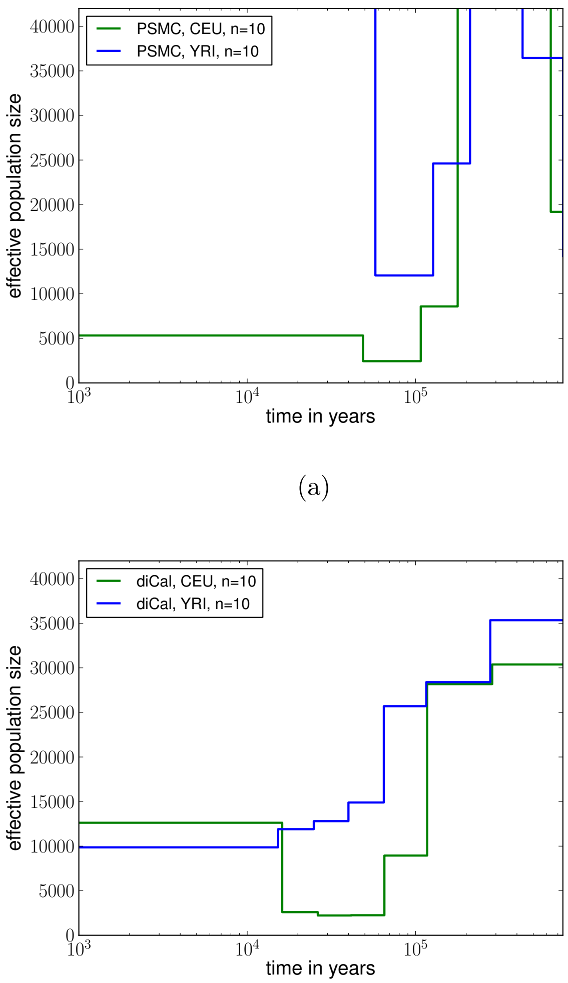
In contrast to previous findings [e.g., Li & Durbin (2011)], our method did not infer a significant drop in the YRI population size during the early out-of-Africa bottleneck phase in Europeans. Instead, the African effective population size seems to have decreased more gradually over time (possibly due to changes in structure) to an average effective size of $\approx$ 10,000 for the past 16 thousand years. We note that our results for real data are fairly robust to the choice of discretization, given that a sufficient number of coalescence events occur within each set of grouped intervals.

3.7.5 Runtime:

The runtime of our method is $O(Ld(d+n)n)$, where n is the number of sequences, L is the number of bases in each sequence, and d is the number of time discretization intervals; the runtime for each CSD computation is $O(Ld(d+n))$, and each sequence is left out in turn (although this step is parallelizable). The runtime for PSMC is $O(Ld^2P)$, where P is the number of pairs of sequences analyzed. In practice, PSMC can run much faster when consecutive sites are grouped into bins of size 100; a bin is considered heterozygous if it contains at least one SNP and homozygous otherwise. Creating a reasonable binning scheme for multiple sequences is less clear. We are currently exploring this avenue, which could significantly improve our runtime and potentially enable whole-genome analysis.

3.8 Discussion and future work

In this paper, we have generalized the recently developed sequentially Markov conditional sampling distribution framework (Paul et al. 2011) to accommodate a variable population



(a)

(b)

Figure 3.10: Variable effective population size inferred from real human data for European (CEU) and African (YRI) populations. For each population, we analyzed a 2 Mb region on chromosome 1 from 5 diploid individuals (10 haplotypes), assuming a per-generation mutation rate of $\mu = 1.25 \times 10^{-8}$ per site. (a) The results of PSMC, which had some runaway behavior and unrealistic results; the data set is probably too small for PSMC to work accurately. (b) The results of diCal. We inferred that the European population size underwent a severe bottleneck about 117 kya and recovered in the past 16,000 years to an effective size of $\approx$ 12,500. In contrast, our results suggest that the YRI population size did not experience such a significant drop during the early out-of-Africa bottleneck phase in Europeans.

size. One important new idea central to the success and accuracy of our method is the wedding cake genealogy approximation, which modifies the rate of absorption into the trunk by accounting for the varying number of lineages over time. On simulated data, we have shown that our method produces substantially more accurate estimates of the recent effective population size than does PSMC (Li & Durbin 2011).

Applying our method to a 2 Mb intergenic region of chromosome 1 from five Europeans and five Africans, sequenced as part of the 1000 Genomes Project, and using a per-generation mutation rate of $\mu = 1.25 \times 10^{-8}$ per site, we have inferred a severe (out-of-Africa) bottleneck in Europeans that began around 117 kya, with a drop in the effective population size by a factor of 12. In contrast, we have observed a much more mild population size decrease in the African population. We remark that our estimate of the timing of the bottleneck may not be very accurate, since we used only 16 discretization intervals and 7 free population size parameters. Furthermore, all of our inferred times and population sizes would be smaller by a factor of two if we had used $\mu = 2.5 \times 10^{-8}$. See Scally & Durbin (2012) for a more thorough discussion of how new mutation rate estimates are changing the way we view ancient population history. An earlier initial human dispersal out of Africa would fit with archaeological evidence of human artifacts dated at 74 kya in India and 64 kya in China (Scally & Durbin 2012).

During the recent past, our results demonstrate that the effective population size of Europeans has grown in the past 16,000 years, slightly surpassing the effective population size of Africans, which does not show a growth at this resolution. Recent studies (Gutenkunst et al. 2009; Gravel et al. 2011) suggest that the European population size recently grew much faster than the African population size, although the sample size we considered is not large enough to confirm this.

The main strength of our method is in the recent past. Large-scale sequencing studies (Coventry et al. 2010; Keinan & Clark 2012; Nelson et al. 2012) of a subset of genes suggest that humans underwent recent explosive population growth. Our method should be well equipped to infer such recent demographic histories, but we would need to consider a very large sample to accurately infer the rate of expansion and the time of onset. Because of issues of computational speed and memory footprint, our current implementation of the SMCSD computation can handle up to about 20 haplotypes and a few megabases, but we are in the process of exploring ways to increase the scalability. One way in which we should be able to reduce our runtime is by incorporating recently developed algorithms for blockwise HMM computation (Paul & Song 2012), which have been shown to result in a speed-up of several orders of magnitude for large data sets.

All the results in this article make use of a leave-one-out approach (3.12) instead of the well-used product of approximate conditionals (PAC) method proposed by Li & Stephens (2003). Briefly, the PAC approach utilizes the approximate likelihood $\hat{\pi}(h_{\sigma(1)})\hat{\pi}(h_{\sigma(2)}|h_{\sigma(1)}) \cdots \hat{\pi}(h_{\sigma(n)}|h_{\sigma(1)}, \ldots, h_{\sigma(n-1)})$, where $\hat{\pi}$ is an approximate conditional sampling distribution and σ is some permutation of $\{1, \ldots, n\}$. A well-known drawback of this approach is that different permutations may produce vastly different likelihoods. Li & Stephens suggested averaging the PAC likelihood over several random permutations to alleviate this problem and this strategy seems to work reasonably well in practice. In our work, we have avoided the problem by adopting the leave-one-out approach, which yields accurate estimates of population sizes for the recent past, but not as good results for the ancient past. Employing the PAC approach may produce accurate estimates for all times, but a challenge that needs to be addressed in the SMCSD framework is that the wedding cake genealogy, which is based on the *prior* expectation of the number of lineages, may not be accurate when there are few lineages, since coalescence times are more variable when they involve fewer lineages. We are working on improving the accuracy of the SMCSD computation by using the *posterior* absorption time distributions in a recursive fashion so that locus-specific absorption rates tailored to data can be used. This approach, together with the PAC model, should yield more accurate estimates of population sizes.

One factor that we have not investigated is the impact of variable recombination (and/or mutation) rates, although it is conceptually straightforward for our method to accommodate them. We have chosen not to incorporate recombination rate variation into our present implementation as it would make the method even more computationally expensive, since the transition probabilities would then be potentially different at each site. In addition, most fine-scale recombination maps (McVean et al. 2004; Chan et al. 2012; Fearnhead & Smith 2005; Crawford et al. 2004) are inferred under the assumption of a constant population size, which is exactly the assumption we are *not* making. We also note that Li & Durbin (2011) found that recombination hotspots did not impact their results significantly and that the important parameter is the average recombination rate.

A good choice of time discretization is critical to the performance of both diCal and PSMC. It is better to subdivide time more finely during periods with small population size than during periods with large population size when few coalescences occur. However, since the demography is what we are trying to infer, selecting an initial discretization is very difficult. Creating adaptive discretization schemes for coalescent HMMs is an important area of future research.

We have shown that posterior decodings of diCal enable accurate inference of coalescence times. Using this information, it should be possible to develop an efficient method of sampling marginal coalescent trees from the posterior distribution. We expect such local tree inference to have interesting applications, including genome-wide association studies and tests of selective neutrality.

The SMCSD framework has been recently extended (Steinrücken et al. 2013) to incorporate structured populations with migration. We are currently working on combining this extension with the work presented here to implement an integrated inference tool (to be incorporated into diCal) for general demographic models. Such a method could provide a detailed picture of the demographic history that created the diversity we see today in humans and other species.

Chapter 4

Decoding coalescent hidden Markov models in linear time

Originally published as: K. Harris, S. Sheehan, J.A. Kamm, and Y.S. Song. (2014) "Decoding coalescent hidden Markov models in linear time." *Research in Computational Molecular Biology* 100–114.

Coalescent Hidden Markov Models are versatile tools for inferring ancient population sizes, migration rates, divergence times, and other parameters such as mutation and recombination rates. As more loci, sequences, and hidden states are added to the model, however, the runtime of coalescent HMMs can quickly become prohibitive. Here we present a new algorithm for reducing the runtime of coalescent HMMs from quadratic in the number of hidden time states to linear, without making any additional approximations. Our algorithm can be incorporated into various coalescent HMMs, including the popular method PSMC for inferring variable effective population sizes. We implement this algorithm to speed up our demographic inference method diCal, which is equivalent to PSMC when applied to a sample of two haplotypes. We demonstrate that the linear-time method can reconstruct a population size change history more accurately than the quadratic-time method, given similar computation resources. We also apply the method to data from the 1000 Genomes project, inferring a high-resolution history of size changes in the European population.

4.1 Introduction

Over the past half-decade, the demographic inference literature has expanded so quickly that it contains many contradictory assertions. Estimates of the population divergence time between European and African humans range from 50 to 120 thousand years ago (kya), while estimates of the speciation time between polar bears and brown bears range from 50 kya to 4 million years ago (Hailer et al. 2012; Miller et al. 2012; Cahill et al. 2013). One reason that different demographic methods often infer conflicting histories is that they make different trade-offs between the mathematical precision of the model and scalability to larger input datasets. This is even true when comparing Coalescent Hidden Markov Models (HMMs) such as diCal and PSMC, which are much more similar to each other than to methods that infer demography from summary statistics (Gutenkunst et al. 2009; Palamara et al. 2012; Harris & Nielsen 2013) or Markov chain Monte Carlo (Gronau et al. 2011).

As described in Chapter 3, coalescent HMMs infer approximate Ancestral Recombination Graphs (ARGs), which are hypotheses about the genealogical relationships among the DNA sequences being analyzed. Exact inference of the posterior distribution of ARGs given data is a very challenging problem, the major reason being that the space of hidden states is infinite, parameterized by continuous coalescence times. In practice, when a coalescent HMM is implemented, time needs to be discretized and confined to a finite range of values. It is a difficult problem to choose an optimal time discretization that balances the information content of a dataset, the complexity of the analysis, and the desire to infer particular periods of history at high resolution. Recent demographic history is often of particular interest, but large sample sizes are needed to distinguish between the population sizes at time points that are very close together or very close to the present.

In a coalescent HMM under a given demographic model, optimal demographic parameters can be inferred using an expectation-maximization (EM) algorithm. The speed of this EM algorithm is a function of at least three variables: the length L of the genomic region being analyzed, the number n of sampled haplotypes, and the number d of states for discretized time. In most cases, the complexity is linear in L, but the complexity in n can be enormous because the number of distinct n-leaved tree topologies grows super-exponentially with n. PSMC and CoalHMM avoid this problem by restricting n to be very small, analyzing no more than four haplotypes at a time. diCal admits larger values of n by using a *trunk genealogy* approximation (see (Paul et al. 2011; Sheehan et al. 2013; Steinrücken et al. 2013) for details) which is derived from the diffusion process dual to the coalescent process, sacrificing information about the exact structure of local genealogies in order to analyze large samples which are informative about the recent past.

To date, all published coalescent HMMs have had quadratic complexity in d. This presents a significant limitation given that small values of d lead to biased parameter estimates (Mailund et al. 2011) and limit the power of the method to resolve complex demographic histories. PSMC is typically run with a discretization of size $d = 64$, but diCal and CoalHMM analyses of larger datasets are restricted to coarser discretizations by the cost of increasing the sample size. In this paper, we exploit the natural symmetries of the coalescent process to derive an alternate EM algorithm with linear complexity in d. The speedup requires no approximations to the usual forward-backward probabilities; we perform an exact computation of the likelihood in $O(d)$ time rather than $O(d^2)$ time using an augmented HMM. We implement the algorithms presented in this paper to speed up our published method diCal, which is equivalent to PSMC when the sample size is two, yielding results of the same quality as earlier work in a fraction of the runtime. We have included the speedup in the most recent version of our program diCal; source code can be downloaded at http://sourceforge.net/projects/dical/.

4.2 Linear-Time Computation of the Forward and Backward Probabilities

We consider a coalescent HMM $\mathcal{M}$ with hidden states $S_1, \ldots, S_L$ and observations $x = x_1, \ldots, x_L$. For PSMC, S_ℓ is the discretized time interval in which two homologous chromosomes coalesce at locus ℓ, while x_ℓ is an indicator for heterozygosity. In diCal, recall that the hidden state at locus ℓ is $S_\ell = (H_\ell, T_\ell)$, where $H_\ell \in \mathcal{H}$ denotes the haplotype in the trunk genealogy with which x coalesces at locus ℓ and $T_\ell \in \{1, \ldots, d\}$ denotes the discretized time interval of coalescence; the observation $x_\ell \in \mathcal{A}$ is the allele of haplotype x at locus ℓ. For $n = |\mathcal{H}| = 1$, diCal is equivalent to PSMC. In what follows, we present our algorithm in the context of diCal, but we note that the same underlying idea can be applied to other coalescent HMMs.

4.2.1 A linear-time forward algorithm

We use $f(x_{1:\ell}, (h, j))$ to denote the joint forward probability of observing the partial emitted sequence $x_{1:\ell} := x_1, \ldots, x_\ell$ and the hidden state $S_\ell = (h, j)$ at locus ℓ. The probability of transitioning from state (h', k) at locus ℓ to state (h, j) at locus $\ell + 1$ is denoted by $\phi(h, j \mid h', k)$, the stationary probability of state (h, i) is denoted $\zeta(h, i)$, and the emission probability of the observed allele $x_\ell = a$ given coalescence at $T_\ell = j$ onto haplotype h with allele $h_\ell = b$ at locus ℓ is denoted by $\xi(a \mid b, j)$. When ℓ is obvious from the context, we sometimes use $\xi(a \mid s) := \xi(a \mid h_\ell, j)$ for $s = (h, j)$. Explicit expressions for $\zeta(h, i)$, $\phi(h, j \mid h', k)$, and $\xi(a \mid b, j)$ are given in Appendix B.1.1.

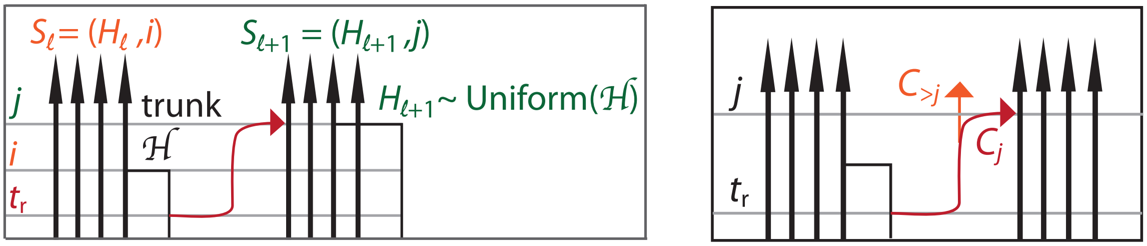
The forward probabilities are computed using the recursion

$$f(x_{1:\ell+1}, (h, j)) = \xi(x_{\ell+1}|h_{\ell+1}, j) \cdot \sum_{k=1}^{d} \sum_{h' \in \mathcal{H}} f(x_{1:\ell}, (h', k)) \cdot \phi(h, j|h', k), \tag{4.1}$$

which contains nd terms. Since there are also nd possibilities for $S_{\ell+1} = (h, j)$, it should naively take $O(n^2 d^2 L)$ time to compute the entire forward dynamic programming (DP) table $\{f(x_{1:\ell}, S_\ell)\}_{\ell=1}^{L}$. The key to achieving a speed-up is to factor (4.1) in a way that reflects the structure of the coalescent, exploiting the fact that many transitions between different hidden states have identical probabilities.

After a sampled lineage recombines at time t_r between loci ℓ and $\ell+1$, it will “float" backward in time from the recombination breakpoint until eventually coalescing with a trunk lineage chosen uniformly at random (Figure 4.1a). This implies that $\phi(h, j|h', k) = \phi(h, j|h'', k)$ whenever $h' \neq h$ and $h'' \neq h$, and exploiting this symmetry allows the forward table to be computed in $O(nd^2 L)$ time. This speed-up was already implemented in the algorithm described in Paul *et al.* (Paul et al. 2011).

Another symmetry of the transition matrix, not exploited previously, can be found by decomposing the transition from locus ℓ to locus $\ell + 1$ as a sequence of component events.


$S_\ell = (H_\ell, i)$
$S_{\ell+1} = (H_{\ell+1}, j)$
j
i
t_r
trunk
$\mathcal{H}$
$H_{\ell+1} \sim \text{Uniform}(\mathcal{H})$
j
$C_{>j}$
C_j
t_r

(a) (b)

Figure 4.1: (a). Here, we illustrate a transition from hidden state $S_\ell = (h_n, i)$ to hidden state $S_{\ell+1} = (h_k, j)$ that proceeds via recombination at time t_r. The probability of this transition does not depend on the identity of the haplotype h_k.(b). As a recombined lineage floats through time interval j, it can either coalesce with the trunk (event C_j) or keep floating (event $C_{>j}$) and eventually coalesce with the trunk in a more ancient time interval.

In particular, let R_i be the event that a recombination occurs during time interval i, and let $\overline{R}$ be the event that no recombination occurs between ℓ and $\ell + 1$. Then we have that

$$\begin{aligned}\phi((h,j) \mid (h',k)) = \frac{1}{n} \sum_{i=1}^{\min(j,k)} \big(&\mathbb{P}(R_i, T_{\ell+1} = j \mid T_\ell = k) \\ &+ \mathbf{1}_{\{(h,j)=(h',k')\}} \mathbb{P}(\overline{R} \mid T_\ell = k)\big)\,, \end{aligned} \tag{4.2}$$

where $\mathbf{1}_E = 1$ if the event E is true or 0 otherwise. The factor $1/n$ corresponds to the probability that the sampled lineage coalesces with haplotype $h \in \mathcal{H}$ in the trunk genealogy.

If a recombination occurs in time interval i, the sampled lineage will start to "float" freely back in time until it either coalesces in i or floats into the next time interval $i+1$ (Figure 4.1b). Specifically, we let $C_{>i}$ denote the event where the sampled lineage recombines at or before i and floats into $i + 1$, and C_i denote the event where the recombined lineage coalesces back in interval i. Noting that $\mathbb{P}(R_i, C_i \mid T_\ell = i')$ and $\mathbb{P}(R_i, C_{>i} \mid T_\ell = i')$ are independent of i' whenever $i' > i$, and that coalescence happens as a Markov process backwards in time, we obtain

$$\begin{aligned}\mathbb{P}(R_i, T_{\ell+1} = j \mid T_\ell = k) = {}& \mathbf{1}_{i=j=k} \cdot \mathbb{P}(R_i, C_i \mid T_\ell = i) \\ &+ \mathbf{1}_{i=j<k} \cdot \mathbb{P}(R_i, C_i \mid T_\ell > i) \\ &+ \mathbf{1}_{i=k<j} \cdot \mathbb{P}(R_i, C_{>i} \mid T_\ell = i) \cdot \prod_{k=i}^{j-1} \mathbb{P}(C_{>k+1} \mid C_{>k}) \\ &+ \mathbf{1}_{i<\min(j,k)} \cdot \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) \cdot \prod_{k=i}^{j-1} \mathbb{P}(C_{>k+1} \mid C_{>k}). \end{aligned} \tag{4.3}$$

Explicit formulas for the above terms are provided in Appendix C.1.

By combining (4.2) with (4.3) and then collecting terms in (4.1), we can remove the sum over $T_\ell = k$ when computing $f(x_{1:\ell+1}, S_{\ell+1})$. In particular, we define additional forward probabilities

$$f\left(x_{1:\ell}, T_\ell = k\right) := \mathbb{P}(x_{1:\ell}, T_\ell = k) = \sum_{h' \in \mathcal{H}} f(x_{1:\ell}, S_\ell = (h', k)), \tag{4.4}$$

$$f\left(x_{1:\ell}, T_\ell > k\right) := \mathbb{P}(x_{1:\ell}, T_\ell > k) = \sum_{k'=k+1}^{d} \sum_{h' \in \mathcal{H}} f(x_{1:\ell}, S_\ell = (h', k')), \tag{4.5}$$

$$\begin{aligned} f\left(x_{1:\ell}, R_{\leq j}, C_{>j}\right) &:= \sum_{i=1}^{j} \mathbb{P}(x_{1:\ell}, R_i, C_{>i}, \ldots, C_{>j}) \\ &= \sum_{i=1}^{j} \left\{ \left[\prod_{i'=i+1}^{j} \mathbb{P}(C_{>i'} \mid C_{>i'-1}) \right] \right. \\ \times \Big[f\left(x_{1:\ell}, T_\ell = i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell = i) &+ f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) \Big] \bigg\}. \end{aligned} \tag{4.6}$$

Then, (4.1) can be written as

$$\begin{aligned} f(x_{1:\ell+1}, (h, j)) \;=\; \xi(x_{\ell+1} \mid h_{\ell+1}, j) \cdot \Bigg[& \frac{1}{n} f\left(x_{1:\ell}, R_{\leq j-1}, C_{>j-1}\right) \mathbb{P}(C_j | C_{>j-1}) \\ & + \frac{1}{n} f\left(x_{1:\ell}, T_\ell > j\right) \mathbb{P}(R_j, C_j \mid T_\ell > j) \\ & + \frac{1}{n} f\left(x_{1:\ell}, T_\ell = j\right) \mathbb{P}(R_j, C_j \mid T_\ell = j) \\ & + f(x_{1:\ell}, (h, j)) \mathbb{P}(\overline{R} \mid T_\ell = j) \Bigg]. \end{aligned}$$

This can be seen by noting that the first three terms in the sum correspond to the terms for $i < j$, $i = j < k$, and $i = j = k$, respectively when putting together (4.1) and (4.2). Alternatively, (4.7) follows from directly considering the probabilistic interpretation of the terms $f(x_{1:\ell}, *)$ as given by (4.4), (4.5), and (4.6).

The required values of $f\left(x_{1:\ell}, R_{\leq i}, C_{>i}\right)$ and $f\left(x_{1:\ell}, T_\ell > i\right)$ can be computed recursively using

$$\begin{aligned} f\left(x_{1:\ell}, T_\ell > i\right) &= f\left(x_{1:\ell}, T_\ell > i + 1\right) + f\left(x_{1:\ell}, T_\ell = i + 1\right), \\ f\left(x_{1:\ell}, R_{\leq i}, C_{>i}\right) &= f\left(x_{1:\ell}, R_{\leq i-1}, C_{>i-1}\right) \mathbb{P}(C_{>i} | C_{>i-1}) \\ &\quad + f\left(x_{1:\ell}, T_\ell = i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell = i) \\ &\quad + f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i), \end{aligned}$$

with the base cases

$$f\left(x_{1:\ell}, T_\ell > d\right) = 0,$$

$$\begin{aligned} f\left(x_{1:\ell}, R_{\leq 1}, C_{>1}\right) &= f\left(x_{1:\ell}, T_\ell > 1\right) \mathbb{P}(R_1, C_{>1} \mid T_\ell > 1) \\ &\quad + f\left(x_{1:\ell}, T_\ell = 1\right) \mathbb{P}(R_1, C_{>1} \mid T_\ell = 1). \end{aligned}$$

Hence, using the recursions (4.7), (4.8), and (4.9), it is possible to compute the entire forward DP table $\{f(x_{1:\ell}, S_\ell)\}_{\ell=1}^L$ exactly in $O(ndL)$ time.

4.2.2 A linear-time backward algorithm

The backward DP table $\{b(x_{\ell+1:L} \mid S_\ell)\}$ can be also computed in $O(ndL)$ time. Given the linear-time forward algorithm discussed in the previous section, the easiest way to compute the backward DP table is as follows: Let $x^{(r)} = x_1^{(r)}, x_2^{(r)}, \ldots, x_L^{(r)} = x_L, x_{L-1}, \ldots, x_1$ denote the reversed x and let $S_\ell^{(r)}$ denote the hidden states for the HMM generating $x^{(r)}$. Then, since the coalescent is reversible along the sequence,

$$b(x_{\ell+1:L}^{(r)} \mid s) = \frac{\mathbb{P}(x_{\ell+1:L}^{(r)}, S_\ell = s)}{\zeta(s)} = \frac{\mathbb{P}(x_{\ell:L}^{(r)}, S_\ell = s)}{\xi(x_\ell^{(r)} \mid s)\zeta(s)} = \frac{f(x_{1:L-\ell+1}^{(r)}, S_{L-\ell+1}^{(r)} = s)}{\xi(x_\ell^{(r)} \mid s)\zeta(s)}.$$

4.3 Linear-Time EM via an Augmented HMM

The primary application of PSMC and diCal is parameter estimation, specifically the estimation of demographic parameters such as changing population sizes. This is done through a maximum likelihood framework with the expectation maximization (EM) algorithm. In this section, we describe how to speed up the EM algorithm to work in linear time.

4.3.1 The standard EM algorithm with $O(d^2)$ time complexity

Let Θ denote the parameters we wish to estimate, and $\hat{\Theta}$ denote the maximum likelihood estimate:

$$\hat{\Theta} = \arg\max_{\Theta'} \mathcal{L}(\Theta') = \arg\max_{\Theta'} \mathbb{P}_{\Theta'}(x_{1:L}).$$

To find $\hat{\Theta}$, we pick some initial value $\Theta^{(0)}$, and then iteratively solve for $\Theta^{(t)}$ according to

$$\Theta^{(t)} = \arg\max_{\Theta'} \mathbb{E}_{S_{1:L};\Theta^{(t-1)}}[\log \mathbb{P}_{\Theta'}(x_{1:L}, S_{1:L}) \mid x_{1:L}],$$

where $S_{1:L} := S_1, \ldots, S_L$. The sequence $\Theta^{(0)}, \Theta^{(1)}, \ldots$ is then guaranteed to converge to a local maximum of the surface $\mathcal{L}(\Theta)$.

Since $(x_{1:L}, S_{1:L})$ forms an HMM, the joint likelihood $\mathbb{P}(x_{1:L}, S_{1:L})$ can be written as

$$\mathbb{P}_{\Theta'}(x_{1:L}, S_{1:L}) = \zeta_{\Theta'}(S_1) \left[\prod_{\ell=1}^{L} \xi_{\Theta'}(x_\ell \mid S_\ell)\right] \left[\prod_{\ell=2}^{L} \phi_{\Theta'}(S_\ell \mid S_{\ell-1})\right].$$

Letting $\mathbb{E}[\#\ell : E \mid x_{1:L}]$ denote the posterior expected number of loci where event E occurs, and $\pi(x) := \mathbb{P}(x) = \sum_s f(x_{1:L}, s)$ denote the total probability of observing x, we then have

$$\begin{aligned}
&\mathbb{E}_{S_{1:L};\Theta}\big[\log \mathbb{P}_{\Theta'}(x_{1:L}, S_{1:L})\big|x_{1:L}\big] \\
=&\sum_s (\log \zeta_{\Theta'}(s))\, \mathbb{P}_\Theta(S_1 = s|x_{1:L}) \\
&\quad + \sum_{(h,i)} \sum_{a,b\in\mathcal{A}} (\log \xi_{\Theta'}(a|b,i))\, \mathbb{E}_\Theta\big[\#\ell : \{S_\ell = (h,i), h_\ell = b, x_\ell = a\}\big|x_{1:L}\big] \\
&\quad + \sum_{s,s'} (\log \phi_{\Theta'}(s' \mid s))\, \mathbb{E}_\Theta\big[\#\ell : \{S_{\ell-1} = s, S_\ell = s'\}\big|x_{1:L}\big] \\
=&\frac{1}{\pi_\Theta(x)}\Bigg[\sum_s (\log \zeta_{\Theta'}(s))\, f_\Theta(x_1, s) b_\Theta(x_{2:L}|s) \\
&\quad + \sum_{(h,i)} \sum_{a,b\in\mathcal{A}} (\log \xi_{\Theta'}(a|b,i)) \sum_{\substack{\ell : x_\ell = a \\ h_\ell = b}} f_\Theta(x_{1:\ell}, (h,i)) b_\Theta(x_{\ell+1:L}|h,i) \\
&+ \sum_{s,s'} (\log \phi_{\Theta'}(s' \mid s)) \left(\sum_{\ell=1}^{L-1} f_\Theta(x_{1:\ell}, s)\phi_\Theta(s' \mid s)\xi_\Theta(x_{\ell+1} \mid s') b_\Theta(x_{\ell+2:L} \mid s')\right)\Bigg].
\end{aligned} \tag{4.10}$$

Note that we have to compute the term $\sum_\ell f_\Theta(x_{1:\ell}, s)\phi_\Theta(s' \mid s)\xi_\Theta(x_{\ell+1} \mid s')\ b_\Theta(x_{\ell+2:L} \mid s')$ for every pair of states s, s', which makes computing the EM objective function quadratic in the number d of discretization time intervals, despite the fact that we computed the forward and backward tables in linear time.

4.3.2 A linear-time EM algorithm

By augmenting our HMM to condition on whether recombination occurred between loci ℓ and $\ell + 1$, the EM algorithm can be sped up to be linear in d. We now describe this augmented HMM. Let $\mathcal{M}$ denote our original HMM, with states $S_{1:L}$ and observations $x_{1:L}$. Between loci ℓ and $\ell + 1$, define

$$\mathcal{R}_{l,l+1} = \begin{cases} \overline{R}, & \text{if no recombination,} \\ R_i, & \text{if recombination occurred at time } i. \end{cases}$$

Now let $S_1^* = S_1$, and $S_\ell^* = (\mathcal{R}_{\ell-1,\ell}, S_\ell)$ for $\ell > 1$. We let $\mathcal{M}^*$ be the HMM with hidden variables $S_{1:L}^* = S_1^*, \ldots, S_L^*$, observations $x_{1:L}$, transition probabilities $\mathbb{P}(S_\ell^* \mid S_{\ell-1}^*) = \mathbb{P}(S_\ell^* \mid S_{\ell-1})$, and emission probabilities $\mathbb{P}(x_\ell \mid S_\ell^*) = \mathbb{P}(x_\ell \mid S_\ell)$. Note that the probability of observing the data is the same under $\mathcal{M}$ and $\mathcal{M}^*$, i.e.,

$$\mathcal{L}(\Theta) = \mathbb{P}_\Theta(x_{1:L} \mid \mathcal{M}) = \mathbb{P}_\Theta(x_{1:L} \mid \mathcal{M}^*),$$

and so we may find a local maximum of $\mathcal{L}(\Theta)$ by applying the EM algorithm to the augmented HMM $\mathcal{M}^*$, instead of to the original HMM $\mathcal{M}$.

To compute the EM objective function for $\mathcal{M}^*$, we start by noting that the joint likelihood is

$$\mathbb{P}(x_{1:L}, S^*_{1:L}) = \zeta(S_1)\left[\prod_{\ell=1}^{L} \xi(x_\ell \mid S_\ell)\right]\left[\prod_{\ell:\mathcal{R}_{\ell,\ell+1}=\overline{R}} \mathbb{P}(\overline{R} \mid T_\ell)\right] \times \left[\prod_{i=1}^{d} \prod_{\ell:\mathcal{R}_{\ell,\ell+1}=R_i} \mathbb{P}(R_i, T_{\ell+1} \mid T_\ell)\right] \left(\frac{1}{n}\right)^{\#\ell:\mathcal{R}_{\ell,\ell+1}\neq\overline{R}}, \tag{4.11}$$

where we decomposed the joint likelihood into the initial probability, the emission probabilities, the transitions without recombination, and the transitions with recombination. We note that the initial probability can be decomposed as

$$\zeta(S_1 = (h, j)) = \frac{1}{n}\left[\prod_{i=1}^{j-1} \mathbb{P}(C_{>i} \mid C_{>i-1})\right]\mathbb{P}(C_j \mid C_{>j-1}), \tag{4.12}$$

and from (4.3), we decompose the product of transition recombination probabilities as

$$\begin{aligned}\prod_{i=1}^{d} \prod_{\ell:\mathcal{R}_{\ell,\ell+1}=R_i} \mathbb{P}(R_i, T_{\ell+1} \mid T_\ell) = \prod_{i=1}^{d} \Bigg\{ & \Bigg[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_\ell=T_{\ell+1}=i}} \mathbb{P}(R_i, C_i \mid T_\ell = i)\Bigg] \\ & \times \Bigg[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_\ell>T_{\ell+1}=i}} \mathbb{P}(R_i, C_i \mid T_\ell > i)\Bigg]\Bigg[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_{\ell+1}>T_\ell=i}} \mathbb{P}(R_i, C_{>i} \mid T_\ell = i)\Bigg] \\ & \times \Bigg[\prod_{\substack{\ell:\mathcal{R}_{\ell,\ell+1}=R_i \\ T_\ell,T_{\ell+1}>i}} \mathbb{P}(R_i, C_{>i} \mid T_\ell > i)\Bigg]\Bigg[\prod_{\substack{\ell:T_{\ell+1}>i \\ \mathcal{R}_{\ell,\ell+1}\in R_{<i}}} \mathbb{P}(C_{>i} \mid C_{>i-1})\Bigg] \\ & \times \Bigg[\prod_{\substack{\ell:T_{\ell+1}=i \\ \mathcal{R}_{\ell,\ell+1}\in R_{<i}}} \mathbb{P}(C_i \mid C_{>i-1})\Bigg]\Bigg\},\end{aligned} \tag{4.13}$$

where $R_{<i} := \cup_{j<i} R_j$. Figure 4.2 shows a graphical representation for the transitions of $\mathcal{M}^*$.

By plugging (4.12) and (4.13) into (4.11), then taking the posterior expected logarithm of (4.11), we obtain the EM objective function for $\mathcal{M}^*$:

$$\mathbb{E}_{S^*_{1:L};\Theta}\left[\log \mathbb{P}_{\Theta'}(x_{1:L}, S^*_{1:L}) \mid x_{1:L}\right] = -L \log n + \sum_{i=1}^{d} q_i(\Theta, \Theta'), \tag{4.14}$$

where

$$q_i(\Theta, \Theta') := \sum_{a,b\in\mathcal{A}} \left[\frac{\log \xi_{\Theta'}(a \mid b, i)}{\pi_\Theta(x)} \sum_{\ell:x_\ell=a} \sum_{h:h_\ell=b} f_\Theta(x_{1:\ell}, (h, i)) b_\Theta(x_{\ell+1:L} \mid (h, i))\right]$$

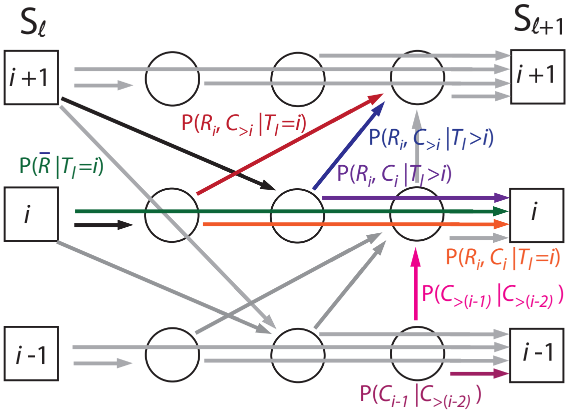
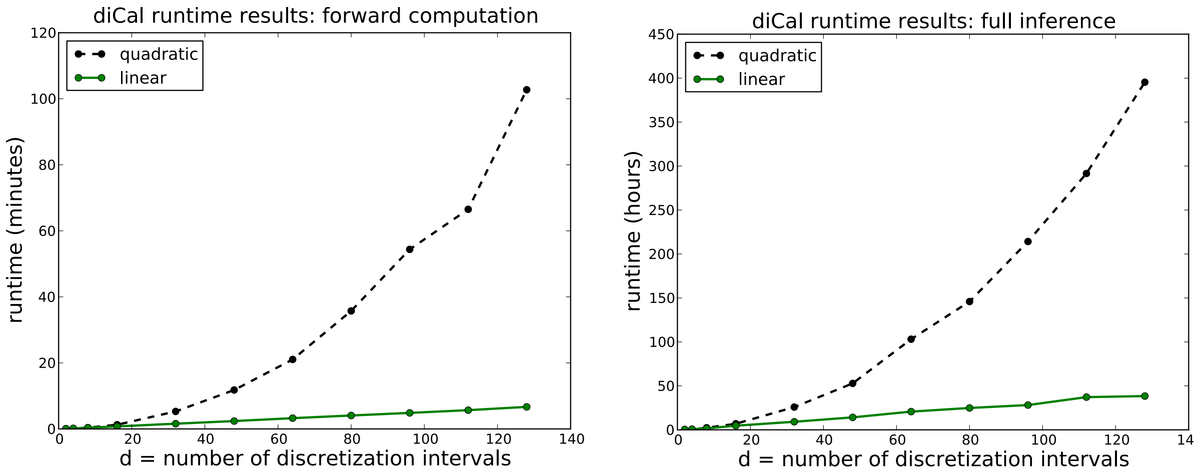


Figure 4.2: This diagram illustrates the flow of transition probabilities through the augmented HMM. Lineages may transition between different coalescence times at loci ℓ and $\ell+1$ by recombining and passing through the floating states represented by circles. Each interval contains three distinct floating states to capture the the dependence of recombination and coalescence probabilities on whether any of these events occur during the same time interval.

$$
\begin{aligned}
&+ \big(\log \mathbb{P}_{\Theta'}(\overline{R} \mid T = i) + \log n\big)\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = \overline{R}, T_\ell = i\} \mid x_{1:L}\Big] \\
&+ (\log \mathbb{P}_{\Theta'}(R_i, C_i \mid T = i))\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell = T_{\ell+1} = i\} \mid x_{1:L}\Big] \\
&+ (\log \mathbb{P}_{\Theta'}(R_i, C_i \mid T > i))\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > T_{\ell+1} = i\} \mid x_{1:L}\Big] \\
&+ (\log \mathbb{P}_{\Theta'}(R_i, C_{>i} \mid T = i))\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_{\ell+1} > T_\ell = i\} \mid x_{1:L}\Big] \\
&+ (\log \mathbb{P}_{\Theta'}(R_i, C_{>i} \mid T > i))\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > i, T_{\ell+1} > i\} \mid x_{1:L}\Big] \\
&+ (\log \mathbb{P}_{\Theta'}(C_{>i} \mid C_{>i-1}))\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} \in R_{<i}, T_{\ell+1} > i\} \mid x_{1:L}\Big] \\
&+ (\log \mathbb{P}_{\Theta'}(C_i \mid C_{>i-1}))\, \mathbb{E}_\Theta\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} \in R_{<i}, T_{\ell+1} = i\} \mid x_{1:L}\Big] \\
&+ \mathbb{P}_\Theta(T_1 > i \mid x_{1:L}) + \mathbb{P}_\Theta(T_1 = i \mid x_{1:L}).
\end{aligned}
\tag{4.15}
$$

The computation time for each of the posterior expectations $\mathbb{E}_\Theta[\#\ell : * \mid x_{1:L}]$ and $\mathbb{P}_\Theta(T_1 \mid x_{1:L})$ does not depend on d; full expressions are listed in Appendix C.2. Hence, the number of operations needed to evaluate (4.14) is linear in d.

We note another attractive property of (4.14). By decomposing the EM objective function into a sum of terms $q_i(\Theta, \Theta')$, we obtain a natural strategy for searching through the parameter space. In particular, one can attempt to find the $\arg\max_{\Theta'}$ of (4.14) by optimizing the $q_i(\Theta, \Theta')$ one at a time in i. In fact, for the problem of estimating changing population sizes, $q_i(\Theta, \Theta')$ depends on Θ' almost entirely through a single parameter (the population size λ'_i in interval i), and we pursue a strategy of iteratively solving for λ'_i while holding the other coordinates of Θ' fixed, thus reducing a multivariate optimization problem into a



(a)

(b)

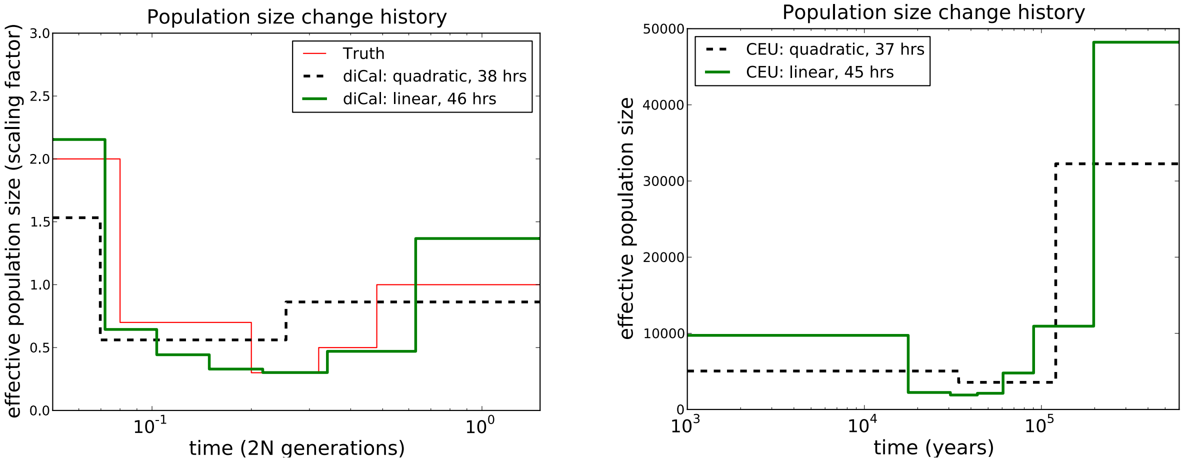
Figure 4.3: Runtime results on simulated data with $L = 2$ Mb and 2 haplotypes, for varying number d of discretization intervals. 4.4 Runtime results (in minutes) for the forward computation. 4.4 Runtime results (in hours) for the entire EM inference algorithm (20 iterations) extrapolated from the time for one iteration.

sequence of univariate optimization problems.

Although both the linear and quadratic EM procedures are guaranteed to converge to local maxima of $\mathcal{L}(\Theta)$, they may have different rates of convergence, and may converge to different local maxima. The search paths of the two EM algorithms may differ for two reasons: first, the intermediate objective functions (4.10) and (4.14) are not equal, and secondly, as discussed above, we use different search strategies to find the optima of (4.10) and (4.14). We have no proven guarantee that either search should perform better than the other, but our observations indicate that the linear-time EM algorithm typically converges to a value of Θ with a equal or higher value of $\mathcal{L}(\Theta)$ than the quadratic-time algorithm, in a fraction of the time (see Figure 4.5 for an example).

4.4 Results

To confirm the decrease in runtime, we ran the linear-time diCal method on simulated data with $L = 2$ Mb of loci and 2 haplotypes (in which case diCal is equivalent to PSMC), using $d = 2, 4, 8, 16, 32, 48, 64, 80, 96, 112, 128$ discretization intervals. To simulate the data, we used `ms` (Hudson 2002) with a population-scaled recombination rate $\rho = 0.0005$ to generate an ARG, and then added mutations using a population-scaled mutation rate of $\theta = 0.0029$ and a finite-sites mutation matrix described by Sheehan et al. (2013). Figure 4.4 shows the time required to compute the table of forward probabilities. We also measured the time



(a)

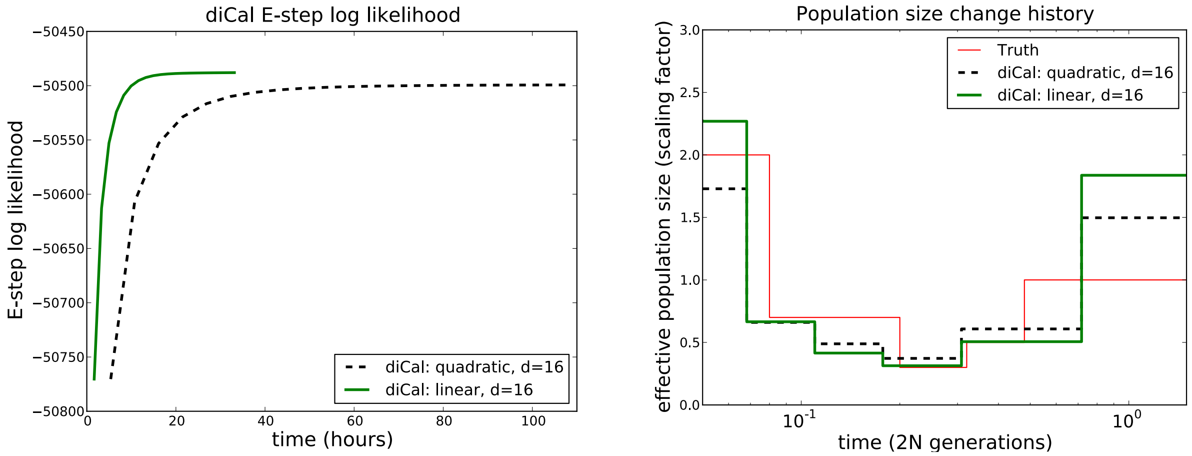
(b)

Figure 4.4: Effective population size change history results. The speedup from the linear method allows us to use a finer discretization ($d = 21$) than the quadratic method ($d = 9$) for about the same amount of runtime. 4.4 Results on simulated data with $L = 2$ Mb and 10 haplotypes. Using the quadratic method with $d = 9$, the error was 0.148. Using the linear method with $d = 21$, the error dropped to 0.079. 4.4 Results on 10 European haplotypes over a 2 Mb region of chromosome 1. The out-of-Africa bottleneck is very apparent with $d = 21$, but is not as well characterized for $d = 9$.

required for one EM iteration and then extrapolated to 20 iterations to approximate the time required to estimate an effective population size history (Figure 4.4). In both figures, the linear runtime of our new algorithm is apparent and significantly improves our ability to increase the number of discretization intervals.

To assess the gain in accuracy of population size estimates that is afforded by more discretization intervals, we ran both the linear- and quadratic-time methods on simulated data with 10 haplotypes and $L = 2$ Mb. The conditional sampling distribution was used in a leave-one-out composite likelihood approach (Sheehan et al. 2013) in this experiment. To run each method for roughly the same amount of time (≈ 40 hours), we used $d = 9$ for the quadratic method and $d = 21$ for the linear method. For both methods, we ran the EM for 20 iterations and inferred $d/3$ size change parameters. As measured by the PSMC error function, which integrates the absolute value of the difference between the true size function and the estimated size function (Li & Durbin 2011), larger values of d permit the inference of more accurate histories.

We also ran our method on 10 CEU haplotypes (Utah residents of European descent) sequenced during Phase I of the the 1000 Genomes Project (1000 Genomes Project 2012) (Figure 4.4). We can see that for the quadratic method with $d = 9$, we are unable to fully characterize the out-of-Africa bottleneck. In the same amount of computational time, we



(a)

(b)

Figure 4.5: Results on simulated data, using the same discretization for the linear and quadratic methods. Each method was run for 20 iterations. 4.4 The log likelihood of the EM algorithm, plotted against time, for both the linear and quadratic methods. 4.4 Population size change history results for the linear and quadratic methods, run with the same discretization using $d = 16$ and estimating 6 parameters.

can run the linear method with $d = 21$ and easily capture this feature. The disagreement in the ancient past between the two methods is most likely due to diCal's lack of power in the ancient past when there are not many coalescence events. Using a leave-one-out approach with 10 haplotypes, the coalescence events in the ancient past tend to be few and unevenly spaced, resulting in a less confident inference.

The runtime of the full EM algorithm depends on the convergence of the M-step, which can be variable. Occasionally we observed convergence issues for the quadratic method, which requires a multivariate optimization routine. For the linear method, we used the univariate Brent optimization routine from Apache Math Commons (`http://commons.apache.org/proper/commons-math/`), which converges quickly and to a large extent avoids local maxima.

To examine the convergence of the two EM algorithms, we ran the linear and quadratic methods on the simulated data with 10 haplotypes and the same number of intervals $d = 16$. We examine the likelihoods in Figure 4.4. The linear method reaches parameter estimates of higher likelihood, although it is unclear whether the two methods have found different local maxima, or whether the quadratic method is approaching the same maximum more slowly. Figure 4.4 shows the inferred population sizes for each method, which although similar, are not identical.

We have also looked at the amount of memory required for each method, and although the difference is small, the linear method does require more memory to store the augmented forward and backward tables. A more thorough investigation of memory requirements will be important as the size of the data continues to increase.

4.5 Discussion

The improvement to diCal described in this paper will enable users to analyze larger datasets and infer more detailed demographic histories. This is especially important given that large datasets are needed to distinguish between histories with subtle or recent differences. By using samples of 10 haplotypes rather than 2, diCal v1.0 (Sheehan et al. 2013) was able to distinguish between histories that diverged from each other less than 0.1 coalescent time units ago, in which period PSMC tends to exhibit runaway behavior and hence cannot produce reliable population size estimates. The faster algorithm described here can handle samples of 30 haplotypes with equivalent computing resources. Our results indicate that this improves the method's ability to resolve rapid, recent demographic shifts.

In organisms where multiple sequenced genomes are not available, the resources freed up by $O(d)$ HMM decoding could be used to avoid grouping sites into 100-locus bins. This binning technique is commonly used to improve the scalability of PSMC, but has the potential to downwardly bias coalescence time estimates in regions that contain more than one SNP per 100 bp.

In general, it is a difficult problem to choose the time discretization that can best achieve the goals of a particular data analysis, achieving high resolution during biologically interesting time periods without overfitting the available data. Sometimes it will be more fruitful to increase the sample size n or sequence length L than to refine the time discretization; an important avenue for future work will be tuning L, n, and d to improve inference in humans and other organisms.

Another avenue for future work will be to develop augmented HMMs for coalescent models with population structure. Structure and speciation have been incorporated into several versions of CoalHMM and diCal, and the strategy presented in this paper could be used to speed these up, though a more elaborate network of hidden states will be required. We are hopeful that our new technique will help coalescent HMMs keep pace with the number and diversity of genomes being sequenced and tease apart the demographic patterns that differentiated them.

Chapter 5

Error-prone polymerase activity causes multinucleotide mutations in humans

Originally published as: K. Harris and R. Nielsen. (2014) “Error-prone polymerase activity causes multinucleotide mutations in humans." *Genome Research* 24 (9) 1445–1454.

One near-universal feature of IBS tract length spectra is an excess of very short tracts (less than 10 base pairs long). The mutational clustering that creates these excess short tracts cannot be explained by demographic processes. Here, we examine clustered mutations that are segregating in a set of 1,092 human genomes and demonstrate that these clusters can be explained by multinucleotide mutations (MNMs), complex events that generate SNPs at multiple sites in a single generation. MNMs have the potential to accelerate the pace at which single genes evolve and to confound studies of demography and selection that assume all SNPs arise independently. About 2% of human genetic polymorphisms have been hypothesized to arise via MNMs, and we show that the signature of MNM becomes enriched as large numbers of individuals are sampled. We estimate the percentage of linked SNP pairs that were generated by simultaneous mutation as a function of the distance between affected sites and show that MNMs exhibit a high percentage of transversions relative to transitions, findings that are reproducible in data from multiple sequencing platforms and cannot be attributed to sequencing error. Among tandem mutations that occur simultaneously at adjacent sites, we find an especially skewed distribution of ancestral and derived alleles, with GC → AA, GA → TT and their reverse complements making up 27% of the total. These mutations have been previously shown to dominate the spectrum of the error-prone polymerase Pol ζ, suggesting that low-fidelity DNA replication by Pol ζ is at least partly responsible for the MNMs that are segregating in the human population. We develop statistical estimates of MNM prevalence that can be used to correct phylogenetic and population genetic inferences for the presence of complex mutations.

5.1 Introduction

The coalescent model predicts that mutations should be somewhat clustered together in the genome. Genomic regions with recent coalescence times are predicted to have a low density of mutations, while regions with ancient coalescence times are predicted to harbor more variation. As genealogy recombines and changes along the sequence, mutation density should also change over a length scale of thousands to millions of base pairs, and these changes are the signal that diCal and the IBS tract method harness to infer demographic history.

To improve the accuracy of population genetic inference from the spacing between SNPs, it will be important to assess the validity of standard assumptions about the mutational process. One such assumption is that mutations occur independently conditional on the genealogical history of the data; however, there are numerous lines of evidence that 1–5% of SNPs in diverse eukaryotic organisms are produced by multinucleotide mutation events (MNMs) that create two or more SNPs simultaneously. If simultaneously generated mutations are regarded as independent during population genetic analysis, the ages of the clustered variants will be overestimated. This could be important not only for the inference of demographic histories, but also for other endeavors such as the detection of long-term balancing selection. Closely spaced SNPs with ancient times to common ancestry can provide evidence that genetic diversity has been maintained by natural selection (Leffler et al. 2013; Charlesworth 2006; Ségurel et al. 2012), and simultaneous mutations have the potential to distort or mimic these signals.

One line of evidence for MNM comes from *de novo* mutations that occur in populations of laboratory organisms including *Drosophila melanogaster* (Keightley et al. 2009; Schrider et al. 2013), *Arabidopsis thaliana* (Ossowski et al. 2010), *Caenorhabditis elegans* (Denver et al. 2004, 2009), and *Saccharomyces cerevisiae* (Lynch et al. 2008), as well as *de novo* mutations detected by looking at human parent-child-trios (Schrider et al. 2011). The human *de novo* mutation rate per base per generation is somewhere between 1.0×10^{-8} and 2.5×10^{-8} (1000 Genomes Project 2010); assuming that mutations occur independently, it should be exceedingly rare to find two mutations within 100 kB of each other. Contrary to this expectation, trios show consistent evidence of mutations occurring in pairs ranging from 2 bp to tens of kb apart.

In yeast, there is additional evidence that MNMs are created by the activity of DNA Polymerase zeta (Pol ζ), an error-prone translesion polymerase that extends DNA synthesis past mismatches and damage-induced lesions (Sakamoto et al. 2007; Stone et al. 2012). Pol ζ is also responsible for MNMs that occur during somatic hypermutation of the variable regions of mouse immunoglobulins (Daly et al. 2012; Saribasak et al. 2012). These results were established by knocking out Pol ζ in mutant yeast strains and adult mouse cells, but it has not been possible to knock out Pol ζ in live mice without destroying their embryonic viability (Bemark et al. 2000; Esposito et al. 2000; Wittschieben et al. 2000). For this reason, there is no direct experimental evidence that Pol ζ creates heritable MNMs in higher eukaryotes.

Clusters of *de novo* mutations are not the only line of evidence for heritable MNM in eukaryotes. Additional evidence can be found in patterns of linkage disequilibrium (LD) between older SNPs that segregate in natural populations. Schrider et al. (2011) and Terekhanova et al. (2013) examined pairs of nearby SNPs in phased human haplotype data and found that the two derived alleles occurred more frequently on the same haplotype than on different haplotypes. When two mutations occur independently, their derived alleles should occur on the same haplotype only 50% of the time; in contrast, MNM should always produce mutation pairs with the two derived alleles on the same haplotype. Using a different counting argument, Hodgkinson & Eyre-Walker (2010) also concluded that many SNP pairs occurring at adjacent sites were generated by a simultaneous mutational mechanism . They noted that adjacent linked SNPs outnumber SNPs 2 bp apart by a factor of two, when the two types of pairs should have equal frequency under the assumption of independent mutation.

To gather more data about the MNM process, it will be impractical to rely on *de novo* mutations and essential to harness LD information. Although it is easiest to classify a pair of SNPs as an MNM when the mutations are observed *de novo*, eukaryotes have low enough mutation rates that fewer than 1 MNM per genome is expected to occur each generation on average. Motivated by this, we use an LD-based approach to identify signatures of MNM in the 1000 Genomes Phase I data, a public repository of 1,098 phased human genomes (1000 Genomes Project 2012). This repository is 100-fold larger than the datasets previously scrutinized for evidence of MNM, and its size confers new power to characterize the MNM spectrum.

In agreement with earlier studies of MNM, we find that patterns of LD between close-together SNPs are incompatible with mutational independence. However, the patterns are consistent with a simple mixture of independent and MNMs. We leverage the size of the 1000 Genomes dataset to make several novel discoveries about MNMs: firstly, that they are enriched for transversions, with a transition: transversion ratio of 1:1 in contrast to the 2:1 genomewide average. Second, we find that linked mutations in humans are enriched for the same allelic types recorded by Stone et al. (2012) in lines of yeast that have nucleotide excision repair (NER) deficiencies and thus rely heavily on Pol ζ for translesion synthesis. These frequent MNMs include the dinucleotide mutations GA → TT and GC → AA as well as mutations at non-adjacent sites that produce homogeneous AA/TT derived allele pairs. Such patterns are unlikely to result from errors in the DNA sequencing process and instead suggest that normal human Pol ζ activity generates at least some of the same MNMs that are produced by Pol ζ in NER-deficient yeast (Stone et al. 2012).

5.2 Results

Simultaneous mutations can be observed directly when they occur *de novo* in offspring that have been sequenced along with their parents. In addition, many more MNMs can be inferred from linkage in data from unrelated individuals. Schrider et al. (2011) previously invoked simultaneous mutations to explain LD patterns in a phased diploid genome, observing that SNPs less than 10 bp apart were disproportionately likely to have their derived alleles lie on the same haplotype . In the spirit of this approach, we looked at the prevalence of neighboring SNPs in the 1000 Genomes Phase I data that occur in perfect LD, meaning that the two derived alleles occur in the exact same subset of the 2,184 sequenced haplotypes. We hereafter define a pair of *close LD SNPs* to be a pair occurring less than 100 bp apart in perfect LD.

5.2.1 Excess nearby SNPs in LD

We counted 35,620 pairs of close LD SNPs in the 1000 Genomes Phase I data with both sites passing genotype quality control and with a consistent ancestral state identifiable from a human/chimp/orang/macaque reference alignment (see Methods). Simultaneous mutations should always create SNPs in perfect LD, but we also expect some independent mutations to create SNPs in perfect LD, and we quantified this expectation by simulating data under a Poisson process model of independent mutation and recombination implemented in `ms` (Hudson 2002). We simulated a total of 4.8×10^8 bp from an alignment of 2,184 haplotypes under a realistic human demographic model (Harris & Nielsen 2013) and recovered 36,991 close LD SNP pairs. For comparison, we also simulated 1.8×10^8 bp of data under the standard neutral coalescent with constant effective population size $N = 10,000$, recovering 36,202 close LD SNP pairs.

As shown in Figure 5.1, the distribution of distances between close LD SNPs were quite different in the simulated versus real data, with the real data containing about 5-fold more adjacent SNPs in LD and a decaying excess of SNPs separated by up to 20 bp in LD. In contrast, the simulations under different demographic models produced similar distributions of close LD SNPs.

Under the coalescent with independent mutation, the abundance of SNP pairs L bp apart in LD should decline approximately exponentially with L for small values of L (see supporting information section D.1), and we find this to hold for the simulated data in Figure 5.1. In contrast, the optimal least-squares exponential fit is a poor approximation to the abundance distribution of close LD SNP pairs in the 1000 Genomes data, which we denote by $N_{\text{LD}}(L)$. A possible explanation is that close linked SNPs are produced by a mixture of two processes, a point-mutation process that is accurately modeled by the coalescent and an MNM process that is not.

5.2.2 Closely linked SNPs have unusual transition/transversion frequencies

To our knowledge, no previous work has addressed whether MNMs have the same transition: transversion ratio as ordinary mutations. However, there is abundant evidence that different DNA polymerases produce mutations with different frequencies of ancestral and derived alleles. To investigate this question, we measured the fractions of linked SNP pairs

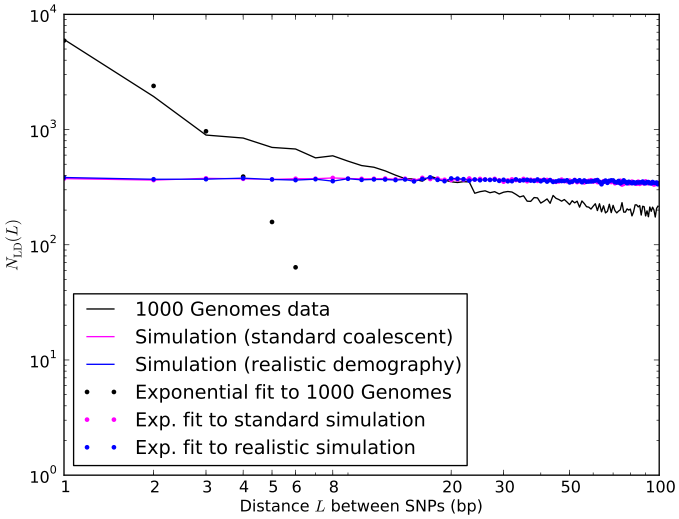


Figure 5.1: **Nearby SNPs in LD: 1000 Genomes Phase I data vs. simulation under mutational independence.** When we simulated 2,184 haplotypes under a realistic demographic model, we observed about 37,000 SNP pairs in LD separated by less than 100 bp in a sample of total length 4.8×10^8 bp. Their spacing was distributed almost uniformly between 1 and 100 bp. Among these pairs, the spacing between SNPs was distributed almost uniformly. We observed much less uniformity in the distribution of distances between SNP pairs in LD in the 1000 Genomes data, with an extreme excess of SNPs in LD at 1–2 bp and a less extreme excess of SNPs at distances up to 20 bp apart. (Note that the axes are logarithmically scaled, making exponential curves appear concave downward).

at distance L that are composed of transitions, transversions, and mixed pairs (one transition plus one transversion). We denote these fractions $f_{\mathrm{ts}}^{\mathrm{LD}}(L)$, $f_{\mathrm{tv}}^{\mathrm{LD}}(L)$, and $f_{\mathrm{m}}^{\mathrm{LD}}(L)$. We also measured the analogous fractions $f_{\mathrm{ts}}^{\text{non-LD}}(L)$, $f_{\mathrm{tv}}^{\text{non-LD}}(L)$, and $f_{\mathrm{m}}^{\text{non-LD}}(L)$ of transitions, transversions, and mixed pairs among SNPs not found in perfect LD.

In human genetic variation data, transitions are approximately twice as common as transversions (Kimura 1980). If the two mutation types of a SNP pair were chosen independently, we would therefore expect that $f_{\mathrm{ts}} = 0.66^2 = 0.44$, $f_{\mathrm{tv}} = 0.33^2 = 0.11$, and $f_{\mathrm{m}} = 2 \times 0.66 \times 0.33 = 0.45$. These predictions are very close to $f_{\mathrm{ts}}^{\text{non-LD}}(L)$, $f_{\mathrm{tv}}^{\text{non-LD}}(L)$, and $f_{\mathrm{m}}^{\text{non-LD}}(L)$ for L between 2 and 100. (Figure 5.2). For $L = 1$, $f_{\mathrm{ts}}^{\text{non-LD}}(L)$ is larger than expected because of the elevated transition rate at both positions of CpG sites.

Among mutations in perfect LD, we found that $f_{\mathrm{ts}}^{\mathrm{LD}}(L)$, $f_{\mathrm{tv}}^{\mathrm{LD}}(L)$, and $f_{\mathrm{m}}^{\mathrm{LD}}(L)$ deviate dramatically from the expectation of mutational independence, adding support to the idea that many such SNPs are produced by a nonstandard mutational process. The frequency of transversion pairs declines with L; we found that 36.7% of SNP pairs in LD at adjacent sites consisted of two transversions, compared to 11.1% of SNP pairs in LD at a distance

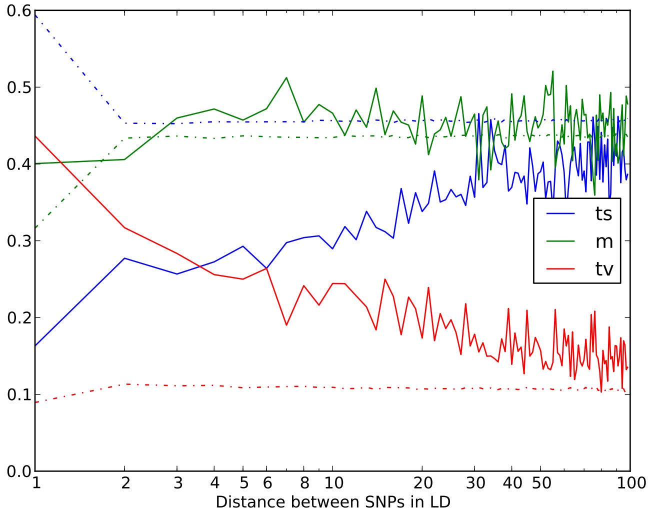


Figure 5.2: **The relationship between LD and the transition:transversion ratio.** In this figure, the black solid line plots the fraction of SNP pairs in LD that consist of 2 transitions. The fraction increases quickly as a function of the distance L between SNPs, asymptotically approaching the fraction of SNP pairs not in LD that consist of 2 transitions (black dashed line). The fraction of SNP pairs not in LD that consist of 2 transitions is nearly constant as a function of L except for an excess of adjacent transition pairs resulting from double mutation at CpG sites. Although transversion pairs make up just over 10% of unlinked SNP pairs, they account for over 40% of adjacent SNPs in perfect LD and about 20% of SNPs in LD at a distance of 10 bp apart.

of 100 bp and 10.7% of SNP pairs not in LD. These numbers are not just incompatible with a transition: transversion ratio of 2:1, but are also incompatible with two neighboring SNP types being assigned independently. If the SNP types were assigned independently, it should hold that $\sqrt{f_{\mathrm{ts}}(L)} + \sqrt{f_{\mathrm{tv}}(L)} = 1$, an assumption that is violated for small values of L. We also found excess close LD transversions in human data sequenced by Complete Genomics (Supporting Figure D.1), suggesting that this pattern is not an artifact of the Illumina sequencing platform or the 1000 Genomes SNP-calling pipeline.

5.2.3 Estimating the fraction of perfect LD SNPs that are MNMs

Schrider, *et al.* previously estimated the abundance of MNMs using the following analysis of a phased diploid genome: for distances L ranging from 1 to 20 bp, they counted heterozygous sites L bp apart where the derived alleles lay on the same haplotype and could potentially have arisen due to MNM. They compared this quantity, $S(L)$, to the number $D(L)$ of heterozygotes L bp apart with the derived alleles on different haplotypes. If all mutations arise independently, $S(L)$ and $D(L)$ are expected to be equal, leading them to propose $S(L) - D(L)$ as an estimate of the number of MNMs spanning L bp. We repeated this analysis on the 1000 Genomes data, subsampling each possible pair H from among the 2,184 phased haplotypes. For each L between 1 and 100 bp, we obtained counts $S_{\mathrm{ts}}^{H}(L)$, $S_{\mathrm{m}}^{H}(L)$, and $S_{\mathrm{tv}}^{H}(L)$ of transitions, mixed pairs, and transversions L bp apart where one haplotype carried the two ancestral alleles and the other haplotype carried the two derived alleles. Similarly, we obtained counts $D_{\mathrm{ts}}^{H}(L)$, $D_{\mathrm{m}}^{H}(L)$, and $D_{\mathrm{tv}}^{H}(L)$ where the derived alleles occurred on opposite haplotypes of H. Adding up these counts over all haplotype pairs subsampled from the 1000 Genomes data, we obtained global counts $S_t(L)$ and $D_t(L)$ for each pair type t. The quantity $(S_{\mathrm{tv}}(L) - D_{\mathrm{tv}}(L))/(S(L) - D(L))$, a direct estimate of the fraction of MNMs that are transversion pairs, is consistently slightly higher than $f_{\mathrm{LD}}^{\mathrm{tv}}(L)$ (Supporting Figure D.2), as expected if close linked SNP pairs are a mixture of MNMs and linked independent mutations.

We were able to use $S_{\mathrm{ts}}(L) - D_{\mathrm{ts}}(L)$, $S_{\mathrm{m}}(L) - D_{\mathrm{m}}(L)$, and $S_{\mathrm{tv}}(L) - D_{\mathrm{tv}}(L)$ to estimate the abundance of MNMs relative to perfect LD SNPs. Our simulations indicate that fewer than 1% of MNMs 100 bp apart should be ultimately broken up by recombination (Supplementary Table **??**); guided by this, we assume that MNMs are a subset of perfect LD SNP pairs. To make this assumption robust to phasing and genotyping error, we relax the definition of perfect LD to include site pairs where at most 2% of samples carry a discordant genotype (see Methods). For each linked SNP pair, we count the number of subsampled haplotype pairs for which exactly one lineage contains the two derived alleles. Adding up these counts over all perfect LD SNPs, we obtain a count $S^{(\mathrm{LD})}(L)$ that is strictly less than $S(L)$. We estimate that $m(L) = (S(L) - D(L))/S^{(\mathrm{LD})}(L)$ is the fraction of perfect LD SNP pairs created by MNM. Similarly, $m_{\mathrm{tv}}(L) = (S_{\mathrm{tv}}(L) - D_{\mathrm{tv}}(L))/S_{\mathrm{tv}}^{(\mathrm{LD})}(L)$ is the fraction of perfect LD transversions created by MNM. The results indicate that more than 90% of SNPs in perfect LD at adjacent sites are MNMs (Figure 5.3). At a distance of 5 bp between sites, 60% of perfect LD transversions are predicted to be MNMs, in contrast to 40% of perfect LD transitions and mixed pairs. At 100 bp between sites, about 35% of perfect LD pairs appear to be MNMs, a figure that is similar across transitions and transversions. We calculate that MNMs spanning 1–100 bp account for 1.8% percent of new point mutations (see Methods). Section D.3 describes how to simulate data containing 1.8% MNMs with realistic spacings of 1–100 bp.

By construction, $m(L)$ should accurately estimate the fraction of MNMs among the close LD SNPs that are polymorphic in a single diploid genome. This might be different from the absolute fraction of 1000 Genomes close LD SNPs that are MNMs, because these contain a higher proportion of rare alleles. However, $m(L)$ is the more relevant statistic to the prevalence of MNMs in smaller datasets that many readers will be concerned with.

The 1000 Genomes data contains many SNP pairs that lie in perfect LD at distances of more than 100 bp apart. Although their transition/transversion ratios are close to the genomewide average, values of $m(L)$ suggest that more than 25% of these are MNMs (Supplementary Table D.1). Although MNMs spanning 10,000 bp appear to be rare events, 10-fold rarer than MNMs spanning only 100 bp, they appear only about 4-fold rarer than independent mutations occurring in perfect LD at 10,000 bp, making it possible to infer their

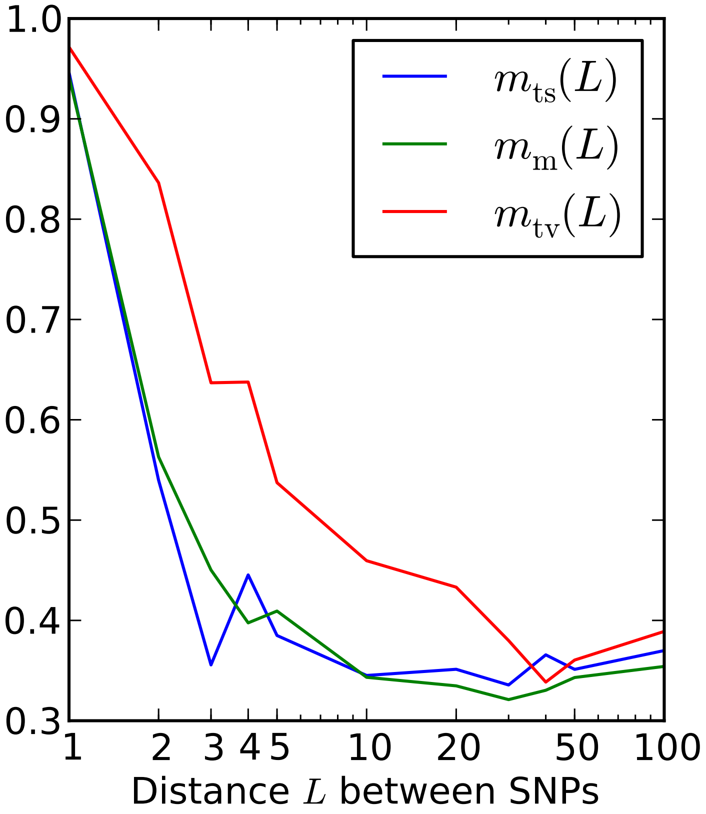


Figure 5.3: **The fraction of SNPs in perfect LD caused by MNM.** The dotted curve plots our estimate of the fraction of transversions in perfect LD L bp apart that were caused by simultaneous mutation. It is uniformly higher than our corresponding estimates for mixed pairs and transitions (dashed and solid lines).

distribution in the genome with high precision.

The large sample size of the 1000 Genomes data not only ensures that a huge number of rare mutations can be observed, but also ensures that independent mutations occur in perfect LD much less often than in samples of fewer individuals. The reason for this is illustrated in Figure 5.4: if two mutations occurred at different time points on the genealogical tree of an entire population, sampling more individuals increases the probability of sampling one who carries the older mutation and not the younger one. To test this prediction, we counted SNP pairs that appear to be in perfect LD in smaller subsets of the 1000 Genomes data. As proved in section D.2 of the supporting information, the genealogies of large samples are dominated by shorter branches, on average, than the genealogies of smaller samples, implying that the percentage of perfected LD SNPs caused by MNM should be an increasing function of the number of sampled lineages. This implies that the abundance of transversions relative to transitions should also increase with the number of sampled lineages.

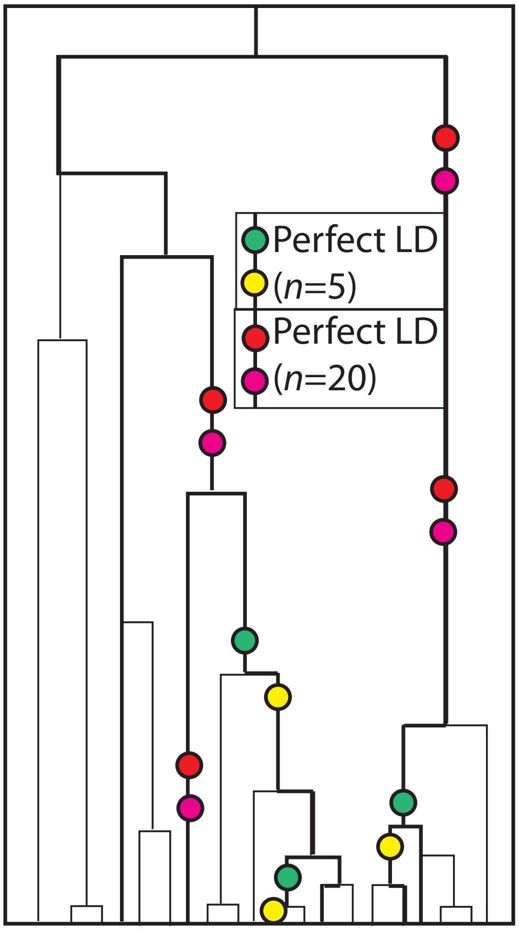


Figure 5.4: **Independent mutations in perfect LD.** This figure depicts a 20-lineage coalescent tree with a 5-lineage subsample highlighted in bold. Light grey circles represent mutation pairs that appear in perfect LD only in the 5-lineage sample. In contrast, dark grey circles represent pairs of independent mutations that occur in perfect LD in the entire 20-lineage sample. These pairs are concentrated on the longest branches of the tree that are often ancestral to many lineages, making their site frequency spectrum enriched for high frequencies.

In pairs of adjacent perfect LD SNPs, we find that the percentage of transversion pairs increases very quickly with the number of lineages, making up 27% of the total when only 2 haplotypes are sampled and nearly 40% of the total when all 2,184 haplotypes are sampled (Figure 5.5). For perfect LD SNPs that occur 100–200 bp apart, the percentage of transversion pairs increases much more slowly than for adjacent perfect LD SNPs. However, it is still 10% higher in a sample of 2,184 haplotypes than in samples of 2 to 1,000 haplotypes (Figure 5.5).

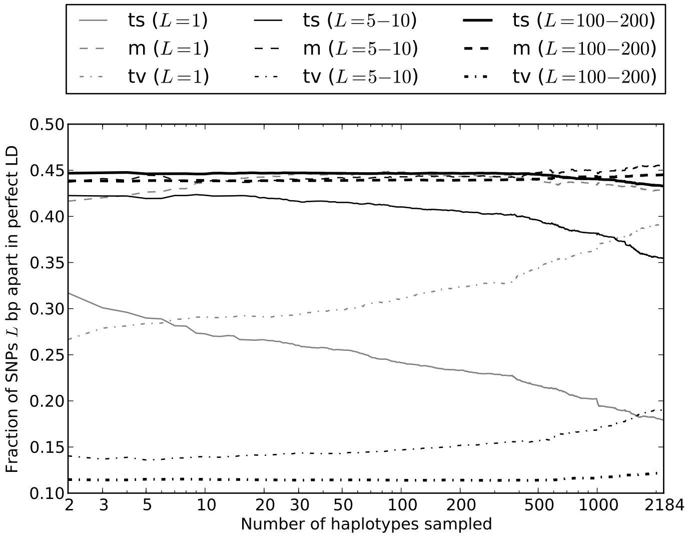


Figure 5.5: **Enrichment of transversion pairs and MNMs with increasing sample size.** We generated subsamples of the 1000 Genomes data containing 2–2,184 haplotypes and computed the percentages of transversion pairs, transition pairs, and mixed pairs for perfect LD SNPs in each dataset. As the number of sampled haplotypes increases, the percentage of perfect LD SNPs that are MNMs should increase, leading to an increase in the frequency of transversions and a decrease in the frequency of transitions. This effect is most apparent when the SNPs are adjacent (1 bp apart) or very close (5–10 bp apart). However, perfect LD SNPs that lie 100–200 bp apart display the same pattern, indicating that MNMs spanning 100–200 bp are much less common but are still evident in samples of many lineages.

5.2.4 Clustering of simultaneous mutations

Mutation-accumulation experiments have reported MNMs spanning long genomic distances (Denver et al. 2009; Keightley et al. 2009; Schrider et al. 2011, 2013), and yeast studies have suggested a possible mechanism for their formation. Roberts, *et al.* reported that double-strand breakage and subsequent repair can create sparse clusters of mutations spanning a megabase or more, with a mean spacing of 3,000 bp between simultaneous mutation events (Roberts et al. 2012). We found evidence for higher-order mutational clustering by counting groups of mutations in perfect LD with fewer than 1000 bp between each adjacent pair and plotting the distribution of cluster size, which ranged from 2 to 31 SNPs. The distribution had a fatter tail than the distribution of perfect LD clusters in an equivalent amount of simulated data, where the largest cluster contained 23 perfect LD SNPs (Supplementary Figure D.3).

5.2.5 The effect of complex mutation on the site frequency spectrum

In addition to showing that large samples contain fewer linked independent mutations than smaller samples, Figure 5.4 illustrates that linked independent mutations should be enriched for high frequencies relative to the site frequency spectrum (SFS) of ordinary mutations. High frequency mutations tend to occur on the longest branches of a genealogical tree, whereas low frequency mutations are scattered across many short branches that are each less likely to be hit with two separate mutations. Simulations confirm that linked independent mutations are biased toward high frequencies, with 6-fold fewer singletons and doubletons than the SFS of the dataset they come from (Figure 5.6F). In contrast, MNMs should have the same SFS as ordinary point mutations as long as they are not affected differently by natural selection.

Given a mixture of simultaneous and independent mutations, the SFS should be a linear combination of the site frequency spectra of independent and simultaneous linked mutations. The more heavily the mixture is weighted toward independent mutations, the more the SFS should be skewed toward high frequencies. In agreement with our inference that MNMs contain a high percentage of transversions, we observe that perfect LD transversions have lower frequencies on average than other perfect LD SNP pairs. In addition, far-apart perfect LD SNPs have higher frequencies than close-together pairs on average (Figure 5.6).

Using the empirical spectra of linked versus unlinked mutations, we devised a second method for estimating the fraction of perfect LD SNPs that are MNMs. For each mutation pair type (ts/m/tv), we compute the site frequency spectrum $\mathbf{S}(L)$ of perfect LD SNPs L bp apart. We also computed a SFS $\mathbf{S}_{\text{global}}$ from the entire set of SNPs in the sample. It is not possible to measure the spectrum $\mathbf{S}_{\text{indept-LD}}$ of linked independent mutations directly, and so we numerically optimized the entries of this spectrum jointly with mixture model coefficients $c_{\text{ts}}(L), c_{\text{m}}(L)$, and $c_{\text{tv}}(L)$ between 0 and 1, one for each distance L and mutation pair type t. We treated all entries of $\mathbf{S}_{\text{indept-LD}}$ as unknown free parameters and used the BFGS algorithm to minimize the following squared error residual $\mathbf{D}$:

$$\mathbf{D} = \sum_{i=2}^{n} \sum_{t \in \{\text{ts, m, tv}\}} \left(c_t(L) \times \mathbf{S}_{\text{global}}[i] + (1 - c_t(L)) \times \mathbf{S}_{\text{indept-LD}}[i] - \mathbf{S}_t(L)[i]\right)^2 \tag{5.1}$$

This has the effect of fitting each spectrum $\mathbf{S}_t(L)$ to the linear combination $c_t(L) \times \mathbf{S}_{\text{global}} + (1 - c_t(L)) \times \mathbf{S}_{\text{indept-LD}}$ of MNMs and linked independent mutations. The sum over i starts at 2 to exclude singletons because they cannot be phased. Assuming that $\mathbf{S}_{\text{global}}$ is a good estimate of the SFS of MNMs, we take $c_t(L)$ to be an estimate of the fraction of MNMs among linked SNPs of type t at distance L. In this way, we obtain estimates similar to the $m_t(L)$ estimates that we obtained earlier by measuring the excess of same-lineage of derived alleles (Figure 5.7). We find that $c_t(L)$ is larger than $m_t(L)$ for $L < 3$ and $L > 50$, but smaller than $m_t(L)$ at intermediate distances. The discrepancy might stem from noise in the data, but might also reflect a meaningful difference between the definitions of the two

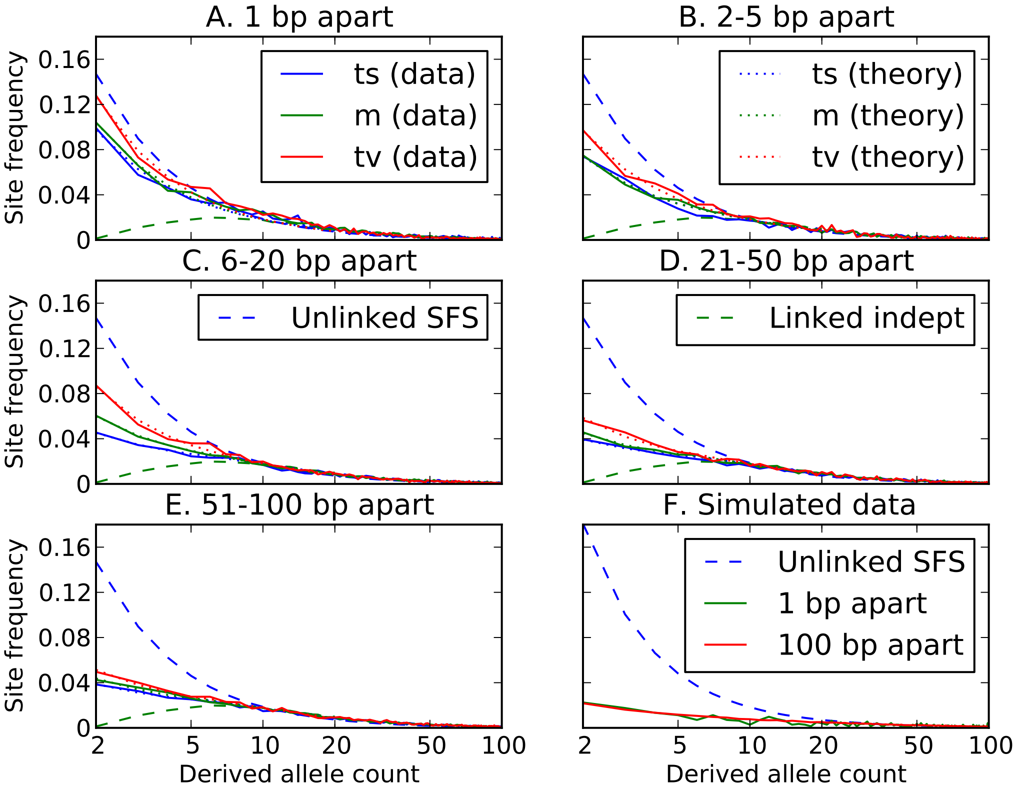


Figure 5.6: **Site frequency spectra of perfect LD mutations.** Each of the first five panels contains site frequency spectra of transitions, mixed pairs, and transversions found in perfect LD in the 1000 Genomes data. Singletons are excluded because they cannot be phased and therefore perfect LD status cannot be determined. For comparison, each panel contains the population-wide SFS of unlinked SNPs as well as the inferred SFS of linked independent mutations. SNP pairs are binned according to the distance between them, showing that close-together SNPs and transversions have spectra closer to the population SFS, while far-apart SNPs and transitions appear more weighted toward linked independent mutations. The theory points plotted alongside the data are the frequency spectra predicted by Equation (5.1) for each length and pair type category, assuming that the green dotted line depicts the correct SFS of linked independent mutations and that Figure 5.7 shows the correct MNM percentages in each category. For comparison, Panel F shows a population SFS and perfect LD frequency spectra obtained from data simulated under a human demographic model. In the simulated data, there is no difference between the frequency spectra of linked independent mutations that lie 1 bp apart versus 100 bp apart.

statistics. While $m(L)$ estimates the prevalence of MNMs among close LD SNPs that are heterozygous in a single diploid, $c(L)$ estimates the prevalence of MNMs among all close LD

SNPs present in the 1000 Genomes data.

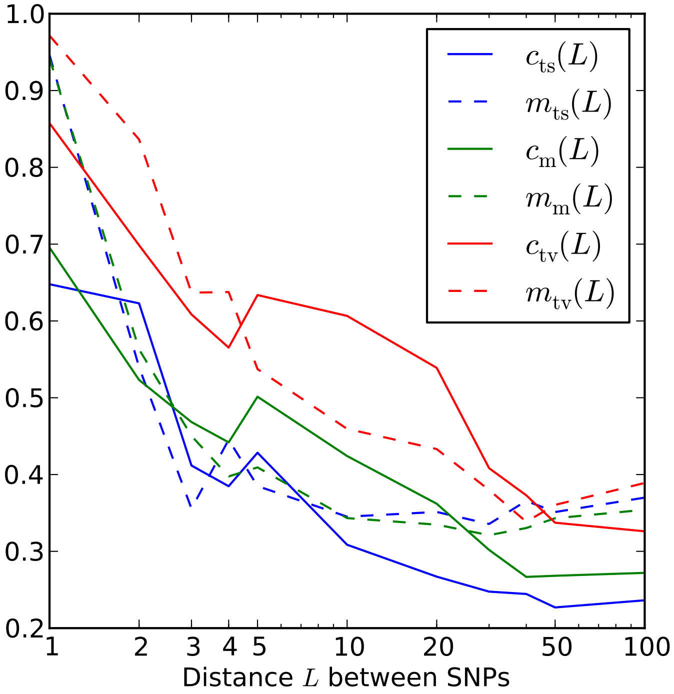


Figure 5.7: **Two estimates of MNM prevalence.** Here, the black lines plot $c_t(L)$, our SFS-based estimate of the fraction of perfect LD mutations caused by MNM. For comparison, grey lines plot the estimate $m_t(L)$ that is based on the excess of same-lineage derived alleles over different-lineage derived alleles in subsampled haplotype pairs.

5.2.6 Evidence for error-prone synthesis by Polymerase ζ

One mechanism that is known to generate MNMs *in vivo* is error-prone lesion bypass by Polymerase ζ, an enzyme found in all eukaryotes with the unique ability to extend primers with terminal mismatches (Gan et al. 2008; Waters et al. 2009). At a replication fork that has been stalled by a lesion, Pol ζ is responsible for adding bases to the strand containing the lesion and then extending replication for a few base pairs before detaching and allowing a high-fidelity enzyme to resume synthesis. During this extension phase, it has the potential to create clustered errors. Experimental work in yeast has confirmed that Pol ζ generates MNMs (Sakamoto et al. 2007; Stone et al. 2012), and the same enzyme has been linked to somatic hypermutation in the MHC (Daly et al. 2012; Saribasak et al. 2012).

Translesion synthesis by Pol ζ is not the only pathway that has the potential to create MNMs. Eukaryotes utilize at least seven different DNA replication enzymes that are considered “error-prone" (Goodman 2002; Waters et al. 2009) and have mutation spectra with low transition/transversion ratios (McDonald et al. 2011). However, we specifically analyzed human MNMs for signatures of Pol ζ activity because a unique dataset was available to make this possible. Specifically, we were able to compare linked adjacent mutations in the 1000 Genomes data to tandem (adjacent) mutations recorded from a yeast strain bred by Stone et al. (2012) to be deficient in nucleotide excision-repair machinery and rely heavily on Pol ζ to bypass lesions that stall replication forks. Stone et al. (2012) recorded a total of 61 spontaneous tandem mutations; these were even more heavily weighted toward transversions than linked SNPs in the 1000 Genomes data, with 52.5% transversion pairs, 37.7% mixed pairs, and only 9.8% transition pairs.

Two particular tandem mutations composed more than 60% of the tandem mutations in the Stone et al. (2012) yeast. One of them, GA → TT, is a transversion pair that made up 31% of the total. The other, GC → AA, is a mixed pair that made up 30% of the total. We found that these were also by far the most common adjacent linked SNPs in the 1000 Genomes data, with GC → AA comprising 16% of the total and GA → TT comprising 11%. No other single mutation type accounts more than 5% of the linked adjacent mutations in the 1000 Genomes data, and no other type accounts for more than 7% of the Stone et al. (2012) tandem mutations (Figure 5.8).

In addition to 61 tandem mutations affecting adjacent base pairs, Stone et al. (2012) recorded 210 complex mutations where two or more substitutions, insertions, and/or deletions occurred at non-adjacent sites within a single 20 bp window. From this dataset, we extracted 84 pairs of simultaneous substitutions at distances of 2–14 bp apart. These pairs had almost the same transition/transversion makeup as the tandem substitutions, being comprised of 53.6% transversions, 36.9% mixed pairs, and 9.5% transitions.

Among the non-adjacent yeast mutation pairs, GA → TT and GC → AA were not particularly common, making up only 4.8% and 1.2% of the total, respectively. However, 44.0% of the derived allele pairs were “AA" or “TT" (compared to 72.1% of adjacent mutation pairs). This percentage is much higher than what we would expect in two mutations that occurred independently. Mutation accumulation studies have shown that 33% of yeast mutations have derived allele A (by A/T symmetry, 33% also have derived allele T) (Lynch et al. 2008). From this, we expect the fraction of AA/TT derived allele pairs to be only $2 \times 0.33^2 = 0.22$. We found that AA and TT were similarly overrepresented among the derived allele pairs in linked human SNPs. In Figure 5.9, we plot the fraction $f_{\mathrm{AA}}(L)$ of derived AA/TT allele pairs as a function of the distance L between perfect LD SNPs, charting its decline from $f_{\mathrm{AA}}(1) = 0.445$ through $f_{\mathrm{AA}}(100) = 0.144$.

5.2.7 Correcting downstream analyses for multinucleotide mutation

As evidenced by Figures 5.1 and 5.6, MNMs can have considerable impact on summary statistics like the site frequency spectrum and the prevalence of linkage disequilibrium. These summary statistics provide clues about the genealogical histories of datasets and can be leveraged to infer demographic history, natural selection, population structure, recom-

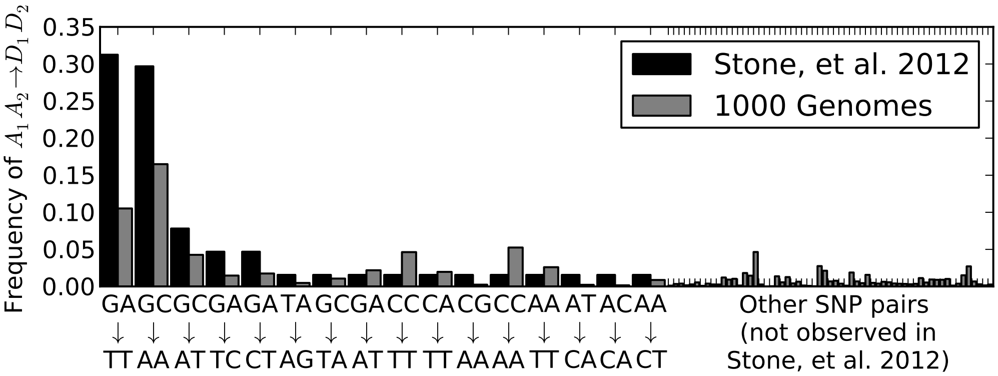


Figure 5.8: **Tandem mutations caused by Pol ζ.** Black bars plot the frequencies of specific tandem mutations observed by Stone, *et al.* in yeast deficient in nucleotide-excision repair machinery. Each mutation type is pooled with its reverse complement because there is no way to know on which DNA strand a mutation occurred. The two mutations GC → AA and GA → TT account for more than 60% of all tandem mutations observed by Stone, *et al.* As shown in grey, these are also the two most common types of mutations occurring at adjacent sites of the 1000 Genomes data in perfect LD.

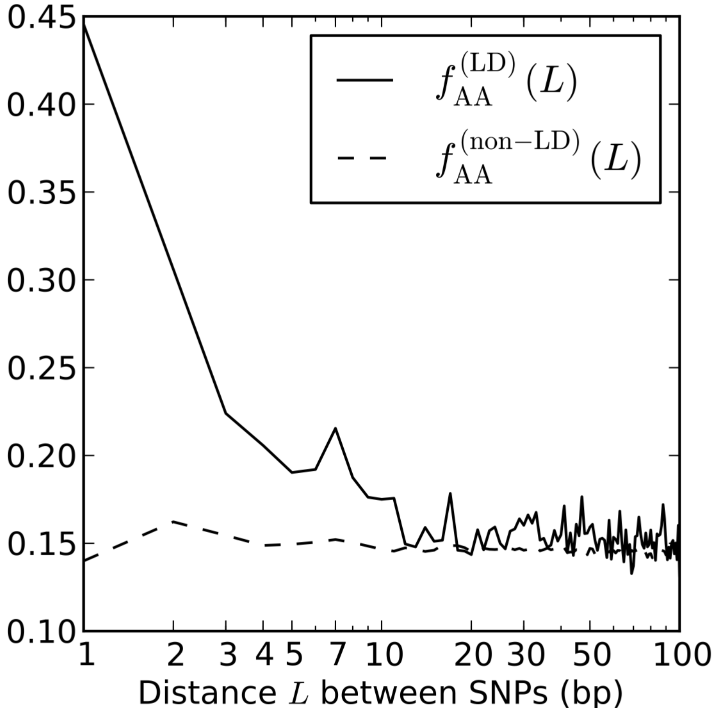


Figure 5.9: **Linked derived AA/TT allele pairs in the 1000 Genomes data.** After observing that a high fraction of yeast MNMs had homogeneous AA/TT derived allele pairs, we tabulated the frequencies $f_{\mathrm{AA}}^{(\mathrm{LD})}(L)$ of AA/TT derived allele pairs among perfect LD SNPs L bp apart in the 1000 Genomes data. For comparison, we also plot $f_{\mathrm{AA}}^{(\text{non-LD})}(L)$, the frequency of AA/TT derived allele pairs among SNPs not in perfect LD. This fraction is consistently lower than $f_{\mathrm{AA}}^{(\mathrm{LD})}(L)$ and does not decrease with the distance between SNPs.

bination rates, and other quantities of interest. However, accurate inference depends on accurately modeling the process that generates data, and most population genetic models omit MNMs.

One strategy for improving the accuracy of downstream analyses without adding much to their complexity is to identify MNMs in a probabilistic way and remove them from the data. For each pair of SNPs occurring in perfect LD, we can estimate the probability that they were caused by an MNM as a function of their inter-SNP distance and transition/transversion status, then use this information to correct summary statistics for the presence of MNMs. To illustrate, we devise a method for correcting the correlation coefficient $r^2(L)$ that is commonly used to measure linkage disequilibrium as a function of genomic distance L (Hill & Robertson 1968). We computed $r^2(L)$ in the 1000 Genomes data as described in the methods and then devised a corrected statistic $r^2_{\text{MNM}}(L)$ that accounts for MNMs and estimates the average correlation between independent mutations. As shown in Figure 5.10, $r^2_{\text{MNM}}(L)$ is significantly less than $r^2(L)$ at short genomic distances.

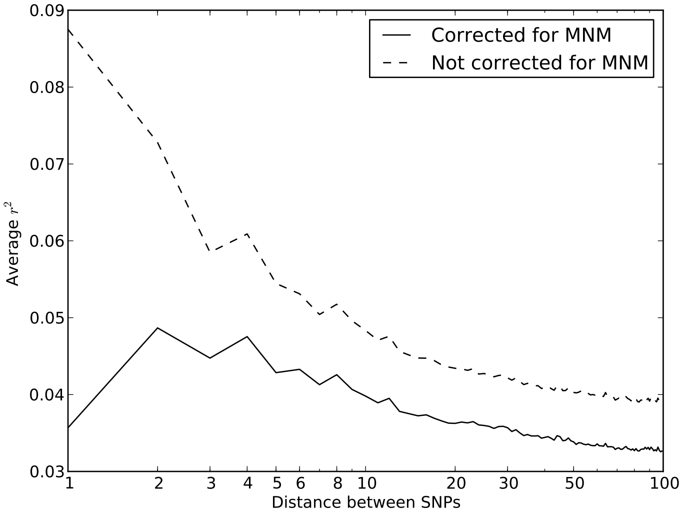


Figure 5.10: **Average r^2 LD correlations between 1000 Genomes SNP pairs.** The correlation coefficient r^2 between allele frequencies at neighboring sites is often used to measure the decay rate of genealogical correlation with genomic distance. However, we have seen that multinucleotide mutation creates excess LD compared to the expectation under independent mutation. We computed the average r^2 across all SNP pairs L bp apart on chromosome 22, then corrected this value for the presence of MNM. r^2_{MNM} is lower at a distance of 1 bp than a distance of 2 bp because of double deaminations at CpG sites that occur on separate lineages.

5.3 Discussion

We have uncovered a strong signature of multinucleotide mutation in 1,092 genomes sequenced by the 1000 Genomes Consortium, with a large excess of close LD SNPs that cannot be explained by demography or mutational hotspots. This is consistent with earlier reports of MNM in smaller human datasets; however, MNMs are enriched relative to independent linked SNPs as more lineages are sampled and mutations are localized to increasingly short genealogical branches.

By looking at the allelic composition of close LD SNPs containing MNMs, we found several signatures that are consistent with error-prone lesion bypass by Polymerase ζ. One signature is an excess of transversions, the second is an excess of the dinucleotide mutations GA $\rightarrow$ TT and GC $\rightarrow$ AA, and the third is a bias toward homogeneous AA/TT derived allele pairs. It remains an open question what percentage of human MNMs are introduced by Pol ζ and how many other DNA damage and repair mechanisms come into play. However, it is interesting that Pol ζ appears to create the same mutation types in the human lineage that it creates in yeast with artificial excision repair deficiencies. We are hopeful that MNM can be understood more completely in the future by comparing perfect LD SNPs to *de novo* mutations from other sources.

An important alternative hypothesis for the observed patterns is DNA sequencing or assembly errors in the 1000 Genomes data, but there are several different lines of evidence that show that our results cannot by explained by such errors. First, we observed similar patterns in data sequenced by Complete Genomics using non-Illumina technology. Secondly, the excess close LD SNPs that are enriched for transversions and AA/TT derived alleles are not only singleton mutations, but occur at a range of higher allele frequencies. Errors could only cause such patterns if they occurred in an identical fashion in multiple individuals, mimicking the frequency distribution expected for mutations. Thirdly, as already noted, the MNMs we infer are enriched for the same types as MNMs that were observed *de novo* in yeast. The patterns we observe are consistent with MNM patterns that have been previously found using Sanger sequencing and other high-fidelity variant detection methods (Drake et al. 2005; Levy et al. 2007; Lynch et al. 2008; Chen et al. 2009).

Most commonly used methods for analyzing DNA sequences assume that mutations occur independently of each other. The fact that this assumption is violated in human data, and perhaps most other eukaryotic data, may compromise the accuracy of population genetic inference. Methods based solely on counting mutations, such as SFS based methods (Gutenkunst et al. 2009) will probably be minimally affected and mostly in their measures of statistical confidence. In contrast, methods that explicitly use the spatial distributions of mutations, in particular the number of mutations in short fragments of DNA (Yang & Rannala 1997; Wang & Hey 2010; Nielsen & Wakeley 2001; Gronau et al. 2011) should be strongly affected. Several recently developed methods analyze genomic data by explicitly modeling the spatial distribution of independent mutations (Hobolth et al. 2007; Li & Durbin 2011; Sheehan et al. 2013; Harris & Nielsen 2013), and these are at risk for bias in regions where SNPs are close together. However, confounding of these methods by MNM can be minimized by analyzing only a few individuals at a time and by disregarding pairs of SNPs less than 100 bp apart, which is often coincidentally done for the sake of computational efficiency (Li & Durbin 2011; Harris & Nielsen 2013). MNMs likely have a stronger effect on methods that look at data from many individuals across short, allegedly non-recombining genomic fragments that are only 1 kb long and contain many SNPs fewer than 100 bp apart (Yang & Rannala 1997; Gronau et al. 2011). However, our results can be used to devise bias-correction strategies, because as illustrated in Figure 5.3, it is straightforward to estimate the probability that a given pair of linked SNPs is an MNM. This also has the potential to improve the accuracy of phylogenetic tree branch length estimation and molecular-clock-based inferences, as well as dN/dS estimation, and their associated measures of statistical confidence. Our results are also relevant to the interpretation of evidence that genetic variation is being maintained by balancing selection–such evidence typically involves short loci with closely spaced linked SNPs (Leffler et al. 2013; Charlesworth 2006; Ségurel et al. 2012).

MNMs have the potential to accelerate evolution by quickly changing several amino acids within a single gene (Schrider et al. 2011). Our results indicate that they also have the potential to increase both sequence homogeneity and A/T content. There is evidence that repetitive sequences experience more indels and point mutations than sequences of higher complexity (McDonald et al. 2011), possibly due to the recruitment of error-prone polymerases, giving MNM extra potential to speed up local sequence evolution by triggering downstream mutations. We are hopeful that more details about this process can be elucidated by studying the spatial and allelic distribution of MNMs. In this way, population sequencing data could provide new information about the biochemistry of DNA replication, e.g. providing a way to measure the activity of Pol ζ over evolutionary time. Pol ζ is tightly regulated in embryonic and adult cells because over- and under-expression can each be harmful; excess error-prone DNA replication increases the genomic mutation rate, but impaired translesion synthesis ability can lead to replication fork stalling, DNA breakage, and translocations that are more harmful than point mutations (Waters & Walker 2006; Waters et al. 2009; Ogawara et al. 2010; Northam et al. 2010; Lange et al. 2011). An important avenue for future work will be to assess whether different eukaryotes incur different levels of MNM because of changing evolutionary pressures being exerted on error-prone DNA replication activity throughout the tree of life.

5.4 Methods

5.4.1 Data summary and accession

We performed all of our analyses on SNP calls that were generated by the 1000 Genomes Project Consortium using joint genotype calling on 2–6x whole genome coverage of 1,092 humans sampled worldwide (1000 Genomes Project 2012). All sequences were mapped to the human reference hg19. To determine ancestral alleles, we downloaded alignments of hg19 to the primate genomes PanTro2 (chimpanzee), ponAbe2 (orangutan), and rheMac3 (rhesus macaque) from the UCSC Genome Browser.

5.4.2 Ascertainment of SNP pairs from the 1000 Genomes Phase I data

Let $S(L)$ be a count of SNPs that are polymorphic in a pair of haplotypes and lie L bp apart with their derived alleles on the same haplotype. Similarly, let $D(L)$ be a count of SNPs with derived alleles that lie on opposite haplotypes. To measure $S(L)$ and $D(L)$ precisely from the 1000 Genomes data, we used a stringent procedure for ancestral identification, utilizing only sites that had the same allele present in chimp, orangutan, and rhesus macaque. For each pair p of SNPs L bp apart satisfying this criterion and passing the 4-gamete test (to avoid confounding effects of recombination and sequencing error), we counted the number of haplotypes $N_{\mathrm{AA}}(p)$ carrying the ancestral allele at both sites, the number $N_{\mathrm{AD}}(p)$ carrying the ancestral allele at only the first site, the number $N_{\mathrm{DA}}(p)$ carrying the ancestral allele at only the second site, and the number $N_{\mathrm{DD}}(p)$ with both derived alleles. Singletons are excluded because they cannot be phased. Combining this information across the set $P(L)$ of SNP pairs L bp apart, we obtain counts

$$S(L) = \sum_{p \in P(L)} N_{\mathrm{AA}}(p) \times N_{\mathrm{DD}}(p)$$

and

$$D(L) = \sum_{p \in P(L)} N_{\mathrm{AD}}(p) \times N_{\mathrm{DA}}(p)$$

as desired.

The quantity $S(L) - D(L)$ has been used as an estimate of the number of MNMs lying L bp apart. Since two simultaneous mutations should always lie in perfect LD, $S(L) - D(L)$ should in theory always be smaller than the following count of perfect-LD same lineage pairs:

$$S_{\mathrm{LD}}(L) = \sum_{p \in P(L)} N_{\mathrm{AA}}(p) \times N_{\mathrm{DD}}(p) \times \mathbf{1}(N_{\mathrm{AD}} = N_{\mathrm{DA}} = 0)$$

To count perfect LD mutation pairs in a way that is more robust to genotype and phasing error, we instead compute $S_{\mathrm{LD}}(L)$ as follows:

$$\begin{aligned} S_{\mathrm{LD}}(L) \quad = \quad & \sum_{p \in P(L)} N_{\mathrm{AA}}(p) \times N_{\mathrm{DD}}(p) \\ & \times \mathbf{1}\left(\frac{N_{\mathrm{AD}} + N_{\mathrm{DA}}}{1092} < \min\left(0.02, \frac{2N_{\mathrm{AA}} + N_{\mathrm{AD}} + N_{\mathrm{DA}}}{2 \times 2184}, \frac{N_{\mathrm{AD}} + N_{\mathrm{DA}} + 2N_{\mathrm{DD}}}{2 \times 2184}\right)\right). \end{aligned}$$

This criterion is designed such that genotyping/phasing error up to 2% will not disrupt perfect LD, but such that very low or high frequency alleles will not be considered in perfect LD unless at least half of the minor alleles appear in the same lineages.

We use a slightly different procedure to obtain the counts $N_{\text{ts}}^{\text{LD}}(L)$, $N_{\text{m}}^{\text{LD}}(L)$, and $N_{\text{tv}}^{\text{LD}}(L)$ that do not need to be compared to $S(L) - D(L)$. After dividing $P(L)$ into transition pairs, mixed pairs, and transversion pairs to obtain sets $P_{\text{ts}}(L)$, $P_{\text{m}}(L)$, and $P_{\text{tv}}(L)$, we simply count the number of pairs with derived alleles occur in the exact same set of lineages:

$$N_t^{\text{LD}}(L) = \sum_{p \in P_t(L)} \mathbf{1}(N_{\text{AD}}(p) = N_{\text{DA}}(p) = 0)$$

for each $t \in \{\text{ts}, \text{m}, \text{tv}\}$. Nearby singletons are considered to be in perfect LD if the derived alleles occur in the same diploid individual. It is this counting procedure that we use to obtain the site frequency spectra of perfect LD SNPs shown in Figure 5.6.

5.4.3 Simulating SNP pairs in LD under the coalescent

The simulated data used to generated Figure (5.1) was produced using Hudson's `ms` (Hudson 2002). We simulated 2,184 human haplotypes (1,092 African and 1,092 European) under the demographic model published in (Harris & Nielsen 2013) that was previously inferred from tracts of identity by state in the 1000 Genomes trios. Because we were only interested in SNP pairs separated by 100 bp or less, we simulated a total of 5.6×10^5 independent "chromsomes" of length 10 kb using the mutation rate 2.5×10^{-8} bp^{-1}gen^{-1} and the recombination rate 1.0×10^{-8} bp^{-1}gen^{-1}.

5.4.4 Estimating the contribution of MNM to new point mutations

In the 1000 Genomes data, we counted $N_{\text{SNP}} = 17,140,039$ non-singleton SNPs that met our criterion for ancestral identifiability. For each pair type t, we also counted the number $N_t^{\text{relaxed-LD}}(L)$ of t-type SNP pairs L bp apart that met the relaxed definition of perfect LD given in equation (5.3). We estimate the fraction $f_{\text{MNM}} = 0.019$ produced by MNM using the following equation:

$$f_{\text{MNM}} = \frac{2}{N_{\text{SNP}}} \sum_{t \in \{\text{ts},\text{m},\text{tv}\}} \sum_{L=1}^{100} N_t^{\text{relaxed-LD}}(L) \times m_t(L)$$

This fraction is a lower bound because it discounts singletons and MNMs spanning more than 100 bp.

5.4.5 Calculating r^2 with a correction for multinucleotide mutation

Given two SNPs s_A, s_B with major alleles A, B and minor alleles a, b, let p_{AB}, p_{Ab}, p_{aB}, and p_{ab} be population frequencies of each of the four associated haplotypes. Let p_A, p_a, p_B, and p_b be the allele frequencies at individual loci. One measure of linkage disequilibrium between the loci is the correlation coefficient

$$r^2(s_A, s_B) = \frac{|p_{AB}p_{ab} - p_{aB}p_{Ab}|}{\sqrt{p_A p_a p_B p_b}}.$$

LD decays as a function of the genetic distance between loci. It is often useful to summarize the rate of this decay by computing the average value of $r^2(s, s')$ over all SNP pairs (s, s') that occur L bp apart. Letting $S(L)$ denote this set of SNP pairs, we define

$$r^2(L) = \frac{1}{|S(L)|} \sum_{(s_1,s_2)\in S(L)} r^2(s_1, s_2).$$

To avoid averaging together the effects of MNM and linked independent mutation, it would be ideal to replace $S(L)$ with the of SNPs LD bp apart that were produced by independent pairs of mutations.

Although it is not possible to classify a SNP pair in perfect LD as an MNM unambiguously, we can correct for MNM by estimating the probability that each observed SNP s was generated as part of a pair of simultaneous mutations. This probability, $\mathbb{P}_{\text{MNM}}(s)$, is calculated as a function of the nearest SNP s_{LD} occurring in perfect LD with s. If s is not in perfect LD with any other SNP within 1000 bp, we assume that s was generated by an ordinary point mutation and let $\mathbb{P}_{\text{MNM}}(s) = 0$. Otherwise, letting $A(s, s_{\text{LD}})$ denote the allelic state of the pair (s, s_{LD}) (either transitions (ts), transversions (tv), or mixed (m)) and L denote the distance between s and s_{LD}, we estimate that $\mathbb{P}_{\text{MNM}}(s) = m_{A(s,s_{\text{LD}})}(L)$. Note that when s_1 and s_2 are in perfect LD and mutually closer to one another than to any other SNP in perfect LD,

$$\frac{1}{2}(\mathbb{P}_{\text{MNM}}(s_1) + \mathbb{P}_{\text{MNM}}(s_2)) = \mathbb{P}_{\text{MNM}}(s_1) = \mathbb{P}_{\text{MNM}}(s_2).$$

After estimating $\mathbb{P}_{\text{MNM}}(s)$ for each SNP s that occurs in $S(L)$, we use these values to compute a weighted average $r^2_{\text{MNM}}(L)$ that downweights each SNP by the probability that it is part of a complex mutation pair:

$$r^2_{\text{MNM}}(L) = \frac{\sum_{(s_1,s_2)\in S(L)} r^2(s_1, s_2)(1 - (\mathbb{P}_{\text{MNM}}(s_1) + \mathbb{P}_{\text{MNM}}(s_2))/2)}{\sum_{(s_1,s_2)\in S(L)} 1 - (\mathbb{P}_{\text{MNM}}(s_1) + \mathbb{P}_{\text{MNM}}(s_2))/2}.$$

Chapter 6

Evidence for recent, population-specific evolution of the human mutation rate

Originally published as K. Harris. (2015) “Evidence for recent, population-specific evolution of the human mutation rate." *Proceedings of the National Academy of Sciences USA* 112 (11) 3439–3444.

One underlying assumption of the coalescent is that point mutations occur at a constant rate per site per generation over time. In contradiction of this assumption, I present evidence that the rate of a particular mutation type has recently increased in the European population, rising in frequency by 50% during the 40,000–80,000 years since Europeans began diverging from Asians. A comparison of single nucleotide polymorphisms (SNPs) private to Africa, Asia, and Europe in the 1000 Genomes data reveals that private European variation is enriched for the transition 5’-TCC-3’→5’-TTC-3’. This transition is known to be the most common somatic mutation present in melanoma skin cancers, as well as the mutation most frequently induced *in vitro* by UV, raising the possibility that mutation spectrum changes might have accompanied local adaptation of human DNA damage response pathways. Regardless of its causality, this change indicates that DNA replication fidelity has not remained stable even since the origin of modern humans and might have changed numerous times during our recent evolutionary history.

6.1 Introduction

The “molecular clock" assumption posits that mutations accumulate at a constant rate over time. This assumption underlies most methods for inferring demographic history, including the methods discussed in Chapters 2, 3, and 4. It is generally assumed that the molecular clock is valid over timescales as short as the lifespan of the human species. However, a few lines of evidence have challenged its validity over the timescale of the human-chimpanzee divergence.

As early as the 1960s, Goodman (1961) and Li & Tanimura (1987) inferred that humans appeared to be accumulating mutations more slowly than chimpanzees, who in turn appeared to be mutating more slowly than gorillas and monkeys. They hypothesized that a “hominoid slowdown" in the mutation rate had occurred, possibly because of changes in generation time. Additional evidence for a hominoid slowdown came to light recently as a byproduct of efforts to measure the human mutation rate. It recently became possible to sequence parent-offspring trios and directly count the rate of new mutations per site per generation; however, the resulting estimates differ more than 2-fold from earlier estimates inferred indirectly from the genetic divergence between humans and chimpanzees (Kong et al. 2012; 1000 Genomes Project 2010). Scally & Durbin (2012) has suggested that the hominoid slowdown might be responsible for this discrepancy, but Ségurel et al. (2014) and others have noted that sequencing errors might be enough to explain the difference without invoking mutation rate change over relatively short timescales.

Trio-based estimates of human and chimp mutation rates have so far both fallen in the range of $1.0 - 1.25 \times 10^{-8}$ mutations per site per generation (Kong et al. 2012; 1000 Genomes Project 2010; Venn et al. 2014). However, the two species appear to differ in the distribution of *de novo* mutations between the male and female germlines and among different mutation types (e.g. a higher proportion of chimp mutations are CpG transitions) (Venn et al. 2014). These patterns suggest that there has been some degree of mutation rate evolution since the two species diverged.

To my knowledge, previous studies have presented no evidence of mutation rate evolution on a timescale as recent as the human migration out of Africa. Most human trios that have been sequenced are European in origin, meaning that there exist few measurements of *de novo* mutation patterns on diverse genetic backgrounds. However, there is some reason to suspect that mutation rates might have changed due to recent regional adaptations affecting DNA repair. SNPs that affect gene expression in DNA damage response pathways show evidence of recent diversifying selection, exhibiting geographic frequency gradients that appear to be correlated with environmental UV exposure (Fraser 2013). I sought to test whether mutation rates vary between populations using rare segregating SNPs that arose as new mutations relatively recently (1000 Genomes Project 2010, 2012), examining the 1000 Genomes data for mutation spectrum asymmetries that could be informative about human mutation rate evolution.

6.2 Results

6.2.1 Mutation spectra of continent-private variation

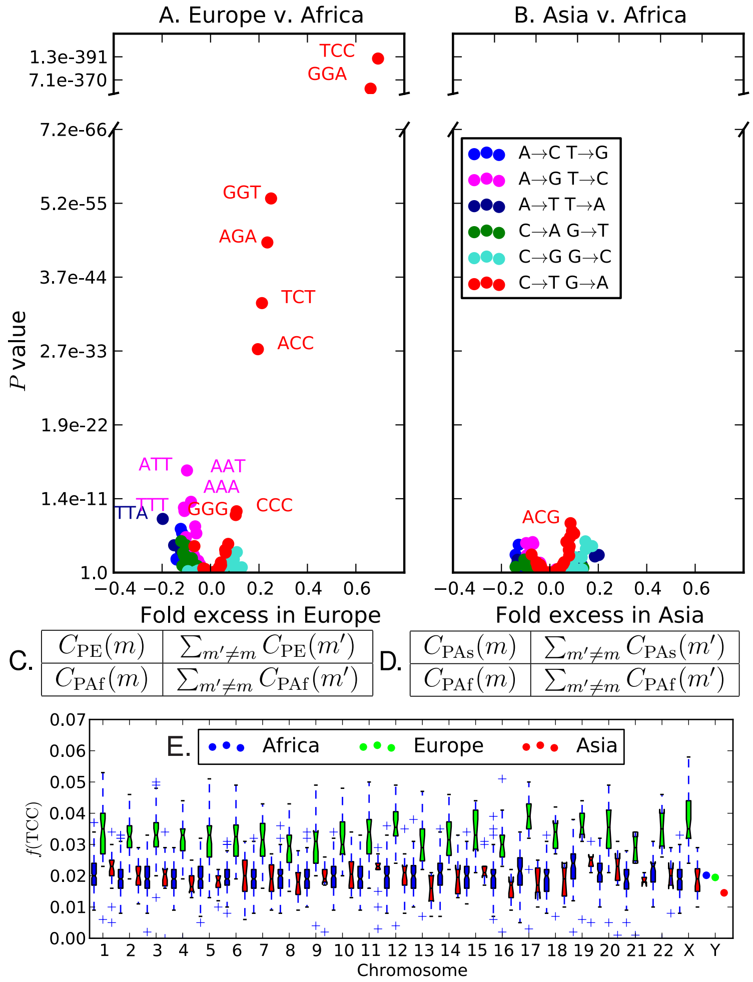
To test for differences in the spectrum of mutagenesis between populations, I compiled sets of population-private variants from the 1000 Genomes Phase I panel of 1,092 human genome sequences (1000 Genomes Project 2012). Excluding singletons and SNPs with imputation quality lower than RSQ = 0.95, which might be misleadingly classified as population-private due to imputation error, there remain 462,876 private European SNPs (PE) that are variable in Europe but fixed ancestral in all non-admixed Asian and African populations, as well as 265,988 private Asian SNPs (PAs) that are variable in Asia but fixed ancestral in Africa and Europe. These SNPs should be enriched for young mutations that arose after humans had already left Africa and begun adapting to temperate latitudes. I compared PE and PAs to the set of 3,357,498 private African SNPs (PAf) that are variable in the Yorubans (YRI) and/or Luhya (LWK) but fixed ancestral in Europe and Asia. One notable feature of PE is the percentage of SNPs that are C→T transitions, which is higher (41.01%) than the corresponding percentages in PAs (38.99%) and PAf (38.29%).

Excess C→T transitions are characteristic of several different mutagenic processes including UV damage and cytosine deamination (Alexandrov et al. 2013). To some extent, these processes can be distinguished by partitioning SNPs into 192 different context-dependent classes, looking at the reference base pairs immediately upstream and downstream of the variable site (Hwang & Green 2004). For each mutation type $m = B_{5'}B_AB_{3'} \rightarrow B_{5'}B_DB_{3'}$ and each private SNP set P, I obtained the count $C_P(m)$ of type-m mutations in set P and used a χ^2 test to compare $C_{\mathrm{PE}}(m)$ and $C_{\mathrm{PAs}}(m)$ to $C_{\mathrm{PAf}}(m)$.

As shown in Figure 6.2.1A, the strongest candidate for mutation rate change is the transition 5'-TCC-3'→5'-TTC-3' (hereafter abbreviated as TCC→T). Combined with its reverse strand complement 5'-GGA-3'→5'-GAA-3', TCC→T has frequency 3.32% in PE compared with 1.98% in PAf and 2.04% in PAs. Several other C→T transitions are also moderately more abundant in PE than PAf, in most cases flanked by either a 5' T or a 3' C.

The TCC→T frequency difference holds genome-wide, evident on every chromosome except for chromosome Y, which has too little population-private variation to yield accurate measurements of context-dependent SNP frequencies (Figure 6.2.1E). The most parsimonious explanation is that Europeans experienced a genetic change increasing the rate of TCC→T mutations. C→T transitions may not be the only mutations that experienced recent rate change; for example, TTA→TAA mutations appear to be less abundant in Europe than in Africa.

If mutation type m occurs at a higher rate in Europe than in Asia, a European haplotype should contain excess type-m derived alleles compared to an Asian haplotype. This prediction is tested in Section E.1 of the Supporting Information. The results suggest that many mutation types occur at slightly higher rates in Europe compared to Asia, with C→T transitions, particularly TCC→T, showing the strongest signal of rate differentiation. This asymmetry cannot be explained by a demographic event such as a population bottleneck; however, it should be interpreted with caution because many bioinformatic biases have the potential to confound this test. Prüfer, et al. used a similar technique to quantify divergence between archaic and modern genomes and found that branch length differences between sequencing batches often exceeded branch length differences between populations (Prüfer et al. 2014). In addition, because the 1000 Genomes Phase I dataset is heavily imputed and contains more European genomes than Asian genomes, rare European variants might be ascertained more completely than rare Asian variants. This could produce a false overall excess of European derived alleles, but seems unlikely to elevate the discovery rate of



$C_{\mathrm{PE}}(m)$	$\sum_{m' \neq m} C_{\mathrm{PE}}(m')$
$C_{\mathrm{PAf}}(m)$	$\sum_{m' \neq m} C_{\mathrm{PAf}}(m')$

$C_{\mathrm{PAs}}(m)$	$\sum_{m' \neq m} C_{\mathrm{PAs}}(m')$
$C_{\mathrm{PAf}}(m)$	$\sum_{m' \neq m} C_{\mathrm{PAf}}(m')$

Figure 6.1: **Overrepresentation of 5'-TCC-3'→5'-TTC-3' within Europe.** Panels A,B: The x coordinate of each point in gives the fold frequency difference $(f_{\mathrm{PE}}(m) - f_{\mathrm{PAf}}(m))/f_{\mathrm{PAf}}(m)$ (resp. $(f_{\mathrm{PAs}}(m) - f_{\mathrm{PAf}}(m))/f_{\mathrm{PAf}}(m)$), while the y coordinate gives the Pearson's χ^2 p-value of its significance. Outlier points are labeled with the ancestral state of the mutant nucleotide flanked by two neighboring bases, and the color of the point specifies the ancestral and derived alleles of the mutant site. Panels C and D show the χ^2 contingency tables used to compute the respective p values in Panels A and B. Panel E shows the distribution of $f(\mathrm{TCC})$ across bins of 1000 consecutive population-private SNPs. Only chromosome-wide frequencies are shown for Chromosome Y because of its low SNP count.

TCC→T in Europe relative to other mutation types.

6.2.2 Robustness to sources of bioinformatic error

Figures 6.2.1, E.2, and E.3 suggest that the human mutation rate is remarkably labile, with significant change having occurred since the relatively recent European/Asian divergence. In this section, I summarize evidence that this conclusion is not founded on bioinformatic artifacts. I focus on confirming the veracity of the TCC→T excess in Europe, but do not discount the possibility that other mutation types might have experienced smaller rate changes.

To rule out the possibility that TCC→T excess in Europe is a bioinformatic artifact specific to the 1000 Genomes data, I reproduced Figure 6.2.1A,B in a set of human genomes sequenced at high coverage using Complete Genomics technology (Supporting Information Section E.3) (Drmanac et al. 2010). I also partitioned the 1000 Genomes data into bins based on GC content and imputation accuracy, finding that the TCC→T excess in Europe was easily discernible within each bin (Supporting Information Sections E.5 and E.7). Three other C→T transitions (TCT→TTT, ACC→ATC, and CCC→CTC) are also more abundant in Europe than Africa across a broad range of GC contents and base qualities. In contrast, genomic regions that differ in GC content and/or quality show little consistency as to which mutation types show the most frequency differentiation between Africa and Asia.

As mentioned previously, singleton variants (minor allele count = 1) were excluded from all analyses. When singletons are included, they create spurious between-population differences that are not reproducible with non-singleton SNPs (Supporting Information Section E.6). This is true of both the low coverage 1000 Genomes dataset and the smaller, higher coverage Complete Genomics dataset, suggesting that singletons are error-prone even in high coverage genomes.

6.2.3 Antiquity of the European mutation rate change

The 1000 Genomes Phase I dataset contains samples from five European sub-populations: Italians (TSI), Spanish (IBS), Utah residents of European descent (CEU), British (GBR), and Finnish (FIN). All of these populations have elevated TCC→T frequencies, suggesting that the European mutation rate changed before subpopulations diversified across the continent. To assess this, I let $\mathrm{P}_{\mathrm{total}}$ denote the combined set of private variants from PE, PAs, and PAf, and for each haplotype h let $\mathrm{P}_{\mathrm{total}}(h)$ denote the subset of $\mathrm{P}_{\mathrm{total}}$ whose derived alleles are found on haplotype h. $f_h(\mathrm{TCC})$ then denotes the frequency of TCC→T within $\mathrm{P}_{\mathrm{total}}(h)$. For each 1000 Genomes population P, Figure 6.2.3 shows the distribution of $f_h(\mathrm{TCC})$ across all haplotypes h sampled from P, and it can be seen that the distribution of $f(\mathrm{TCC})$ values found in Europe does not overlap with the distributions from Asia and Africa. In contrast, the four admixed populations ASW (African Americans), MXL (Mexicans), PUR (Puerto Ricans), and CLM (Colombians) display broader ranges of $f(\mathrm{TCC})$ with extremes overlapping both the European and non-European distributions. The African American $f(\mathrm{TCC})$ values are only slightly higher on average than the non-admixed African values, but a few African American individuals have much higher $f(\mathrm{TCC})$ values in the middle of the admixed American range, presumably because they have more European ancestry than the other African Americans who were sampled.

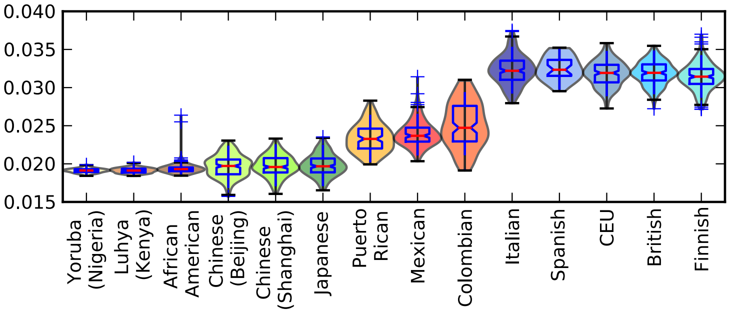


Figure 6.2: **Variation of f(TCC) within and between populations.** This plot shows the distribution of f(TCC) within each 1000 Genomes population, i.e. the proportion of all derived variants from PA, PE, and PAf present in a particular genome that are TCC→T mutations. There is a clear division between the low f(TCC) values of African and Asian genomes and the high f(TCC) values of European genomes. The slightly admixed African Americans and more strongly admixed Latin American populations have intermediate f(TCC) values reflecting partial European ancestry.

Within Europe, Figure 6.2.3 shows a slight f(TCC) gradient running from North to South; the median f(TCC) is lowest in the Finns and highest in the Spanish and Italians. In this way, TCC→TTC frequency appears to correlate negatively with recent Asian co-ancestry (Supporting Information Section E.2).

To roughly estimate the time when the TCC→T rate increased, I downloaded allele age estimates that were generated from the Complete Genomics data using the program ARG-weaver (`http://compgen.bscb.cornell.edu/ARGweaver/ CG_results/`) (Rasmussen et al. 2014). Based on these estimates, TCC→T rate acceleration appears to have occurred between 25,000 and 60,000 years ago, not long after Europeans diverged from Asians (Supporting Information Section E.4). In the 1000 Genomes, data, TCC→T frequency differentiation is greatest for private alleles of frequency less than 0.02 (Supplementary Figure E.6B).

It is hard to tell from current data whether skin lightening predated TCC→T acceleration in Europe. A 7,000-year-old Early European farmer was found to be homozygous for the skin-lightening SLC24A5 allele (Lazaridis et al. 2014), suggesting that light skin was relatively prevalent by 7,000 years ago and could have originated much earlier. An attempt to date the origin of this allele yielded a 95% confidence interval of 6,000 to 38,000 years ago (Beleza et al. 2013), which overlaps with the time interval when the TCC→T rate appears to have accelerated.

6.2.4 Reversal of TCC→T transcription strand bias in Europe

Transcribed genomic regions are subject to transcription-coupled repair (TCR) of damaged nucleotides that occur on the sense DNA strand. This can lead to patterns of strand bias that contain information about underlying mutational mechanisms. For example, CpG transitions generally result from deamination damage to the cytosine rather than the guanine, and damaged Cs that occur on the transcribed strand are repaired more often than damaged Cs occurring on the nontranscribed strand. As a result, CpG transitions in genic regions are usually oriented with the C→T change on the nontranscribed strand (Skandalis et al. 1994).

To assess the strand bias of genic TCC→T mutations and look for strand bias differences between populations, I counted the occurrences of each A/C ancestral mutation m from each private SNP set P on transcribed gene strands versus nontranscribed gene strands, denoting these counts $\mathbf{T}(P,m)$ and $\mathbf{N}(P,m)$, respectively. Strand biases $\mathbf{S}(P,m) = \mathbf{N}(P,m)/\mathbf{T}(P,m)$ were compared between populations using a χ^2 test (Figure 6.2.4). Private Asian and African TCC→T SNPs were found exhibit the strand bias that is typical of A/C→G/T mutations (Green et al. 2003), with the C→T change usually affecting the antisense strand and the G→A change usually affecting the sense strand. In contrast, private European TCC→T SNPs exhibit no discernible strand bias; the C→T change affects the sense strand about 50% of the time (Figure 6.2.4E). TCC→T is the only mutation type that exhibits a significant strand bias difference between populations at the level $p < 0.01$.

Given that the TCC→T mutation rate is the same in genic and intergenic regions because TCR is ineffective at preventing TCC→T mutations in Europeans, we should expect the frequency $f(\text{TCC})$ to be slightly higher among private genic SNPs than among private intergenic SNPs. This is because the frequencies of all mutation types sum to 1; mutation types that are efficiently prevented by TCR should have lower frequencies in genic regions than in intergenic regions, and mutation types that are not very susceptible to TCR must have higher genic frequencies to compensate. As predicted by this logic, $f(\text{TCC})$ is higher among private genic European SNPs than among private intergenic European SNPs (Figure 6.2.4F). In contrast, when PAs and PAf are partitioned into genic and intergenic SNPs, the genic SNP sets have lower TCC→T frequencies, suggesting that TCR of this mutation type is relatively efficient in non-Europeans. This TCR differential alone could modestly elevate the European TCC→T mutation rate. However, it is not likely to be the sole cause of the observed TCC→T rate acceleration because this acceleration is evident in both genic and intergenic regions.

6.3 Discussion

It is beyond the scope of this chapter to pinpoint why the rate of TCC→T increased in Europe. However, some promising clues can be found in the literature on ultraviolet-induced mutagenesis. In the mid-1990s, Drobetsky & Sage (1993) and Marionnet et al. (1995) each observed that TCC→T dominated the mutational spectra of single genes isolated from

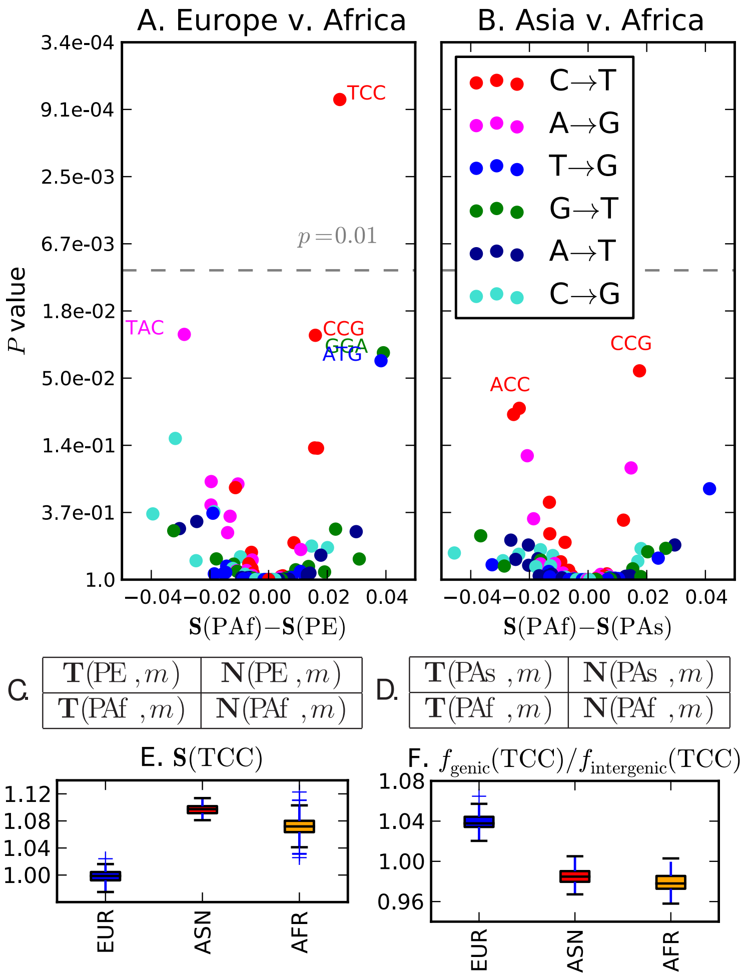


Figure 6.3: **Differences in transcriptional strand bias.** Each point in panels A and B represents a mutation type with an A or C ancestral allele. The x coordinate of each point in panel A is the PAf strand bias minus the PE strand bias; similarly, the x coordinates in panel B describe the PAf strand bias minus the PAs strand bias. The y coordinate of each point is the χ^2 p-value of the strand bias difference. At the $p = 0.01$ significance level (grey dashed line), only TCC→T has a significant strand bias difference between Europe and Africa, while no mutation type significantly differs in strand bias between Asia and Africa. Panel C shows the variance of strand bias in each population across 100 bootstrap replicates. Similarly, Panel D shows the distribution across bootstrap replicates of the ratio between genic $f(\mathrm{TCC})$ and intergenic $f(\mathrm{TCC})$.

UV-irradiated cell cultures. Much more recently, Alexandrov et al. (2013) systematically inferred "mutational signatures" from 7,042 different cancers and found that melanoma has a unique mutational signature not present in any other cancer type they studied. Melanoma somatic mutations consist almost entirely of C→T transitions, 28% of which are TCC→T mutations (Alexandrov et al. 2013; Pleasance et al. 2010). The mutation types CCC→CTC and TCT→TTT, two other candidates for rate acceleration in Europe, are also prominent in the spectrum of melanoma (Supporting Information Section E.8). Incidentally, melanoma is not only associated with UV light exposure, but also with European ancestry, occurring at very low rates in Africans, African Americans, and even light-skinned Asians (Crombie 1979; Hu et al. 2005; Bakos et al. 2009). A study of the California Cancer Registry found that the annual age-adjusted incidence of melanoma cases per 100,000 people was 0.8-0.9 for Asians, 0.7-1.0 for African Americans, and 11.3–17.2 for Caucasians (Cress & Holly 1997). Melanoma incidence in admixed Hispanics is strongly correlated with European ancestry (Cress & Holly 1997; Hu et al. 2005; Bakos et al. 2009).

The association of TCC→T mutations with UV exposure is not well understood, but two factors appear to be in play: 1) the propensity of UV to cross-link the TC into a base dimer lesion and 2) poorer repair efficacy at TCC than at other motifs where UV lesions can form (Brash et al. 1987; Drobetsky et al. 1987). Drobetsky & Sage (1993) compared the incidence of UV lesions to the incidence of mutations in irradiated cells and found that TCC motifs were not hotspots for lesion formation, but instead were disproportionately likely to have lesions develop into mutations rather than undergoing error-free repair.

Despite the strong evidence that UV causes TCC→T mutations, the question remains how UV could affect germline cells that are generally shielded from solar radiation. Although the testes contain germline tissue that lies close to the skin with minimal shielding, to my knowledge it has not been tested whether UV penetrates this tissue effectively enough to induce spermatic mutations. Another possibility is that UV might indirectly cause germline mutations by degrading folate, a DNA synthesis cofactor that is transmitted through the bloodstream and required during cell division (Branda & Eaton 1978; Jablonski & Chaplin 2000, 2010; Off et al. 2005). Folate deficiency is known to cause DNA damage including uracil misincorporation and double-strand breaks, leading in some cases to birth defects and reduced male fertility (Blount et al. 1997; Wallock et al. 2001; Stover 2009). It is therefore possible that folate depletion could cause some of the mutations observed in UV-irradiated cells, and that these same mutations might appear in the germline of a light-skinned individual rendered folate-deficient by sun exposure. It has also been hypothesized that, in a variety of species, differences in metabolic rate can drive latitudinal gradients in the rate of molecular evolution (Gillooly et al. 2005; Allen et al. 2006; Wright et al. 2006).

Although the data presented here do not reveal a clear mechanism, they leave little doubt that the European population experienced a recent mutation rate increase. TCC→T and a few other C→T transitions exhibit the clearest evidence of European rate acceleration, but other mutation types might have experienced smaller rate changes within Europeans or other human populations. Pinpointing finer-scale mutation rate changes will be an important avenue for future work.

Even if the overall European mutation rate increase was small, it adds to a growing body of evidence that molecular clock assumptions break down on a faster timescale than generally assumed during population genetic analysis. It was once assumed that the human lineage's mutation rate had changed little since we shared a common ancestor with chimpanzees, but this assumption is losing credibility due to the conflict between direct mutation rate estimates and molecular-clock-based estimates (Ségurel et al. 2014; Scally & Durbin 2012). Although this conflict might have arisen from a gradual decrease in the rate of germline mitoses per year as our ancestors evolved longer generation times (Goodman 1961; Li & Tanimura 1987), the results of this paper indicate that another force may have come into play: change in the mutation rate per mitosis. If the mutagenic spectrum was able to change during the last 60,000 years of human history, it might have changed numerous times during great ape evolution and beforehand. Given such a general challenge to the molecular clock assumption, it may be wise to infer demographic history from mutations such as CpG transitions that accumulate in a more clocklike way than other mutations (Ségurel et al. 2014; Hwang & Green 2004). At the same time, less clocklike mutations may provide valuable insights into the changing biology of genome integrity.

6.4 Methods

Publicly available VCF files containing the 1000 Genomes Phase I data were downloaded from www.1000genomes.org/data. Ancestral states were inferred using the UCSC alignment of the chimp PanTro4 to the human reference genome hg19. These data were then subsampled to obtain four sets of SNPs: PE (derived allele private to Europe), PAs (derived allele private to Asia), PAf (derived allele private to Africa), and PAsE (fixed ancestral in Africa but variable in both Asia and Europe).

6.4.1 Construction of private SNP sets PE, PAs, PAf, and PAsE

The definitions of PE, PAs, and PAf differ slightly from the definitions of continent-private SNPs in the manuscript announcing the release of the 1000 Genomes Phase I data (1000 Genomes Project 2012). In that paper, a SNP is considered private to Africa if it is variable in at least one of the populations LWK (Luhya from Kenya), YRI (Yoruba from Nigeria), and ASW (African Americans from the Southwestern USA). In contrast, I consider a SNP to be private to Africa if it is variable in either LWK or YRI and is not variable in any of the following samples: CHB (Chinese from Beijing), CHS (Chinese from Shanghai), JPT (Japanese from Tokyo), CEU (Individuals of Central European descent from Utah), GBR (Great Britain), IBS (Spanish from the Iberian Peninsula), TSI (Italians from Tuscany), and FIN (Finnish). A private African SNP might or might not be variable in any of the admixed samples ASW, MXL (Mexicans from Los Angeles), CLM (Colombians from Medellin), and PUR (Puerto Ricans). Similarly, a private European SNP in PE is variable in one or more of the CEU, GBR, IBS, TSI, and FIN, is not variable in any of YRI, LWK, CHB, CHS, or JPT, and might or might not be variable in ASW, MXL, CLM, and PUR. The private Asian SNPs in PAs are variable in one or more of CHB, CHS, or JPT, are not variable in any of YRI, LWK, CEU, GBR, IBS, TSI, and FIN, and might or might not be variable in ASW, MXL, CLM, and PUR. These definitions are meant to select for mutations that have been confined to a single continent for most of their history except for possible recent transmission to the Americas. The shared European-Asian SNPs in PAsE are variable in one or more of CHB, CHS, or JPT plus one or more of CEU, GBR, IBS, TSI, and FIN and are not variable in YRI or LWK. Singletons are excluded to minimize the impact of possible sequencing error, and variants with imputation quality lower than RSQ $= 0.95$ are excluded to minimize erroneous designation of shared SNPs as population-private.

6.4.2 Statistical analysis of frequency differences

Given two SNP sets P_1 and P_2 and one SNP type m, a Pearson's χ^2 test was used to measure the significance of the difference between the frequency of m in P_1 and the frequency of m in P_2.

Let $C_{P_i}(m)$ denote the number of type-m SNPs in set P_i, and let $T(P) = \sum_{m \in M} C_P(m)$ denote the total number of SNPs in P. The expected values of $C_{P_1}(m)$ and $C_{P_2}(m)$, assuming no frequency differences between P_1 and P_2, are calculated as follows based on the 4×4 contingency tables in Figure 6.2.1C,D:

$$\mathbb{E}(C_{P_i}(m)) = \frac{T(P_i)(C_{P_i}(m) + C_{P_{3-i}}(m))}{T(P_1) + T(P_2)}$$

The following χ^2 value with one degree of freedom measures the significance of the difference between $f_m(P_1)$ and $f_m(P_2)$:

$$\begin{aligned}\chi^2 \quad = \quad & \sum_{i=1}^{2} \frac{(C_{P_i}(m) - \mathbb{E}(C_{P_i}(m)))^2}{\mathbb{E}(C_{P_i}(m))} \\ & + \frac{(T(P_i) - C_{P_i}(m) - \mathbb{E}(T(P_i) - C_{P_i}(m)))^2}{\mathbb{E}(T(P_i) - C_{P_i}(m))}\end{aligned}$$

6.4.3 Nonparametric bootstrapping within chromosomes

To assess the variance of $f(\text{TCC})$ within each of the autosomes and the X chromosome, each private SNP set PE, PAs, and PAf was partitioned into non-overlapping bins of 1,000 consecutive SNPs. The frequency $f(\text{TCC})$ of the mutation TCC→T was computed for each bin and the distribution of these estimates is shown in Figure 6.2.1C. No partitioning into separate bins was performed for chromosome Y because the entire chromosome has only 1,130 private European SNPs, 1,857 private Asian SNPs and 3,852 private African SNPs. Instead the global frequency of TCC→T was computed for each SNP set restricted to the Y chromosome.

6.4.4 Quantifying strand bias

Gene locations and transcription directions for hg19 were downloaded from the UCSC Genome browser. For the purpose of this analysis, each SNP located between the start codon and stop codon of an annotated gene is considered to be a genic SNP. All SNPs not located within introns or exons are considered to be intergenic SNPs.

Within each private SNP set P, each mutation type m with an A or C ancestral allele was counted separately on transcribed and non-transcribed genic strands to obtain counts $\mathbf{T}(\mathrm{P}, m)$ and $\mathbf{N}(\mathrm{P}, m)$. (Each mutation with a G/T ancestral allele on the transcribed strand is equivalent to a complementary A/C ancestral mutation on the non-transcribed strand.) The strand bias of mutation m is then defined to be $\mathbf{S}(\mathrm{P}, m) = \mathbf{N}(\mathrm{P}, m)/\mathbf{T}(\mathrm{P}, m)$. The significances of the strand bias differences $\mathbf{S}(\mathrm{PAf}, m) - \mathbf{S}(\mathrm{PE}, m)$ and $\mathbf{S}(\mathrm{PAf}, m) - \mathbf{S}(\mathrm{PAs}, m)$ were assessed using a χ^2 test with 1 degree of freedom. Assuming no difference in strand bias between P_1 and P_2, the expected numbers of transcribed-strand and nontranscribed-strand mutations are the following:

$$\mathbb{E}(\mathbf{T}(P_i, m)) = \frac{(\mathbf{T}(P_i, m) + \mathbf{T}(P_{3-i}, m))(\mathbf{T}(P_i, m) + \mathbf{N}(P_i, m))}{\mathbf{T}(P_i, m) + \mathbf{T}(P_{3-i}, m) + \mathbf{N}(P_i, m) + \mathbf{N}(P_{3-i}, m)}$$

$$\mathbb{E}(\mathbf{N}(P_i, m)) = \frac{(\mathbf{N}(P_i, m) + \mathbf{N}(P_{3-i}, m))(\mathbf{T}(P_i, m) + \mathbf{N}(P_i, m))}{\mathbf{T}(P_i, m) + \mathbf{T}(P_{3-i}, m) + \mathbf{N}(P_i, m) + \mathbf{N}(P_{3-i}, m)}$$

The χ^2 value measuring the significance of the difference between $\mathbf{N}(P_1, m)/\mathbf{T}(P_1, m)$ and $\mathbf{N}(P_2, m)/\mathbf{T}(P_2, m)$ is computed as follows:

$$\chi^2 = \sum_{i=1}^{2} \frac{(\mathbf{T}(P_i, m) - \mathbb{E}(\mathbf{T}(P_i, m)))^2}{\mathbb{E}(\mathbf{T}(P_i, m))} + \frac{(\mathbf{N}(P_i, m) - \mathbb{E}(\mathbf{N}(P_i, m)))^2}{\mathbb{E}(\mathbf{N}(P_i, m))}$$

Non-parametric bootstrapping was used to estimate the variance of TCC→T strand bias within each population. The transcribed portion of the genome was partitioned into 100 bins containing approximately equal numbers of SNPs, and 100 replicates were generated each by sampling 100 bins with replacement. For each replicate, the frequency of TCC→T was calculated on the transcribed and non-transcribed strands. These two frequencies were added together to obtain the cumulative TCC→T frequency within genic regions. The distribution of $\mathbf{S}(\mathrm{TCC} \to \mathrm{T})$ across replicates is shown for each population in Figure 6.2.4C.

Bootstrapping was similarly applied to intergenic SNPs by partitioning the non-genic portion of the genome into 100 bins with similar SNP counts. 100 bootstrap replicates were generated by sampling 100 bins with replacement, and the intergenic TCC→T frequency was computed for each replicate.

The distribution of ratios in Figure 6.2.4D was generated by pairing up each genic bootstrap replicate with a unique intergenic bootstrap replicate and calculating the ratio of genic $f(\mathrm{TCC})$ to intergenic $f(\mathrm{TCC})$, thereby obtaining 100 estimates of the ratio $f_{\mathrm{genic}}(\mathrm{TCC})/f_{\mathrm{intergenic}}(\mathrm{TCC})$.

Chapter 7

Conclusion

This dissertation has explored the landscape of fine-scale distinctions between complex evolutionary models that are rooted in coalescent theory. Large-scale genomic datasets that have become available over the past few years provide power to distinguish evolutionary hypotheses that differ subtly from each other; however, harnessing the power of genomic datasets in this way has presented many computational and statistical challenges. One obvious challenge is that genomic datasets are large, making them cumbersome to store and manipulate. Another challenge is inherent to the complexity of the coalescent: it is usually impossible to exactly and efficiently calculate the probability of observing a particular dataset given a particular model.

To calculate an approximate likelihood of a genomic dataset given a coalescent model, it is necessary to simplify the data, approximate the model, or both. The more aggressively a dataset is summarized, the more complexity can be retained within the model; however, summarizing the data too aggressively can destroy information that is needed to perform finescale model choice. Together, Chapters 2 and 3 can be read as a case study illustrating this tradeoff.

Chapter 2 introduces a novel summary statistic, the IBS tract length spectrum, that retains more information about demographic history than the commonly used site frequency spectrum (SFS). I show that IBS tracts have power to distinguish among histories that are more difficult to distinguish using only the SFS. A caveat is that the SFS contains some information about the sharing of variants across multiple individuals that is not captured by the IBS tract spectrum (which summarizes linkage disequilibrium information between a pair of genome sequences). As a consequence, some human demographic histories inferred from the SFS are not compatible with IBS tract information, but histories inferred from IBS tracts are conversely not always compatible with the SFS.

The method diCal, partially described in Chapters 3 and 4, compresses genomic data to a lesser extent than the IBS tract inference method does. By using a hidden Markov model to parse an alignment of multiple DNA sequences, diCal essentially integrates IBS tract information and SFS information together. The downside of retaining this greater degree of data complexity is that less complexity can be incorporated into the evolutionary models being evaluated. We can use diCal to infer effective population size changes, but not in conjunction with population divergence or migration. Expanded versions of diCal accommodate population subdivision and migration (Steinrücken et al. 2012; Steinrücken et al. 2013), but at significantly higher computational cost compared to IBS tract inference.

Algorithmic improvements can sometimes reduce computational complexity without requiring simplifications to either the data or the model. This is true of diCal 2.0 as described in Chapter 4 compared to diCal 1.0 described in Chapter 3. The method is made significantly faster by departing from standard Hidden Markov Model decoding methods and tailoring a new method to the specific structure of the Markovian coalescent.

The immediate goal of the method development in Chapters 2 through 4 is to infer detailed population histories from whole-genome sequences. However, both diCal and the IBS tract method have additional potential applications. diCal provides local estimates of times to common ancestry that, with some effort, can be pieced together into gene phylogenies (J. Kamm, S. Sheehan, and Y. Song; work in preparation). This information could in turn be used to resolve diploid sequences into phased haplotypes, as noted by Paul et al. (2011).

Although the IBS tract method does not lend itself to TMRCA decoding or phasing, it has other collateral advantages. One advantage is transparency: it is easy to visually inspect predicted and observed IBS tracts to gain intuition about how well each model fits the data. This made it straightforward to see that the distribution of IBS tracts shorter than 100 bp deviated significantly from the predictions of any model explored in the first three chapters. In section 2.4.3, I hypothesized that this discrepancy could be explained by cryptic variation in the mutation rate across the genome. However, when I went on to examine this pattern in more detail, I found that multinucleotide mutations (MNMs) provided a better explanation than simple mutation rate variation. Chapter 5 thoroughly describes the evidence for this claim, pinpointing that Polymerase ζ is likely responsible for the most common class of human SNP clusters, dinucleotide GC→TT mutations. My description of Polymerase ζ activity in the human germline has spurred further research linking Pol ζ activity to Costello Syndrome, a serious congenital disease caused by *de novo* mutation of the HRAS gene (Seplyarskiy et al. 2014).

The Pol ζ mutational signature drew my awareness to the fact that different mutational classes can have different genomic distributions. In contrast, coalescent models for demographic inference generally lump all point mutations together into a single category is are distributed at a constant rate per site per generation. Further investigation revealed that certain types of mutations have different relative frequencies in different human populations, with Europeans appearing to have higher C→T transition rates than Africans or Asians. Chapter 6 presents evidence that the context-dependent transition TCC→TTC has undergone a rapid rate acceleration in the European population, perhaps rapid enough to be the result of locally adaptive changes in DNA replication or repair.

Despite the substantial number of topics that are explored throughout these five chapters, this dissertation has only scratched the surface of the power and shortcomings of complex evolutionary models. Along with others, I have shown that the coalescent with recombination can furnish good descriptions of whole-genome data, and also that high demographic complexity must be incorporated to describe genomic data well. Just as importantly, I have explored some limitations of the models that are used for demographic inference, highlighting features of genomic data that cannot be explained by models where mutation rates are uniform in time and space. This is not meant to discredit the utility of uniform-mutation-rate models, which may remain the sensible choice for demographic inference given the tradeoff between complexity and efficiency. Rather, I highlight the fact that biological discoveries can emerge from the space between model and data. Each time we formulate a better evolutionary model, we essentially peel back layers of genomic data complexity and reveal patterns that were once hidden underneath.

Appendix A

Supporting Information for Chapter 2: Inferring demographic history from a spectrum of shared haplotype lengths

A.1 Derivation of the IBS tract length formula

A.1.1 One intra-tract recombination

Consider an alignment between sequences from two populations of constant size N that diverged at time τ_s, measured in units of $2N$ generations before the present. In the main text, we computed the frequency $H^{(0)}_{\tau_s}(L)$ of L-base IBS tracts with no historical recombinations. We now proceed to compute the frequency $H^{(1)}_{\tau_s}(L)$ of L-base IBS tracts with a history that contains exactly one recombination event. We must marginalize over two coalescence times t_0, t to the left and right of the historical recombination site, respectively, as well as the time of recombination $t_{(r)}$ and the location $L_{(r)}$ of the recombination site:

$$\begin{aligned} H^{(1)}_{\tau_s}(L) &= \sum_{L_{(r)}=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=0}^{\min(t_0,t)} e^{-t_0} e^{-t_0 L_{(r)}(\rho+\theta)}(1-e^{-t_0\theta}) \cdot \rho e^{-\rho t_{(r)}} \\ &\quad \cdot e^{-(t-t_{(r)})} \cdot e^{-t(L-L_{(r)})(\rho+\theta)} dt_{(r)} dt dt_0 \\ &\approx \int_{L_r=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=0}^{\min(t_0,t)} e^{-t_0} e^{-t_0 L_r(\rho+\theta)}(1-e^{-t_0\theta}) \cdot \rho e^{-\rho t_{(r)}} \\ &\quad \cdot e^{-(t-t_{(r)})} \cdot e^{-t(L-L_r)(\rho+\theta)} dt_{(r)} dt dt_0 dL_r \end{aligned}$$

In all formulae, the population size N appears as an implicit factor in $\theta = 4N\mu$ and $\rho = 4Nr$.

Marginalizing over the ordering of t_0 and t to evaluate this integral, we find that

$$H^{(1)}_{\tau_s}(L) = e^{-\tau_s L(\rho+\theta)} \log\left(\frac{(1+(L-1)(\rho+\theta)-\rho)(1+(L-1)(\rho+\theta))}{(1+\theta)(1+\rho+\theta)}\right)$$

$$
\begin{aligned}
&\cdot\frac{1-e^{-\tau_s\rho}}{(\rho+\theta)(2+L(\rho+\theta)-\rho)}\\
&-e^{-\tau_s(L(\rho+\theta)+\theta)}\log\left(\frac{(1+(L-1)(\rho+\theta))^2}{(1+\theta)(1+\rho+2\theta)}\right)\\
&\cdot\frac{(1-e^{-\tau_s\rho})}{(\rho+\theta)(2+L(\rho+\theta)-\rho+\theta)}\\
&+\frac{\rho e^{-\tau_s L(\rho+\theta)}}{(1-\rho)(\rho+\theta)}\left(\log\left(\frac{1+(L-1)(\rho+\theta)}{1+\rho+\theta}\right)\left(\frac{1}{1+L(\rho+\theta)}-\frac{1}{2+L(\rho+\theta)-\rho}\right)\right.\\
&-\log\left(\frac{1+L(\rho+\theta)-\rho}{1+\rho+2\theta}\right)\left(\frac{e^{-\tau_s\theta}}{1+L(\rho+\theta)+\theta}-\frac{e^{-\tau_s\theta}}{2+L(\rho+\theta)+\theta-\rho}\right)\\
&+\log\left(\frac{1+(L-1)(\rho+\theta)-\rho}{1+\theta}\right)\left(\frac{1}{1+L(\rho+\theta)}-\frac{e^{-\tau_s\theta}}{1+L(\rho+\theta)+\theta}\right.\\
&\left.\left.-\frac{1}{2+L(\rho+\theta)-\rho}+\frac{e^{-\tau_s\theta}}{2+L(\rho+\theta)+\theta-\rho}\right)\right)
\end{aligned}
$$

A.1.2 Two intra-tract recombinations

The accuracy of the approximation $H_{\tau_s}(L)\approx H_{\tau_s}^{(0)}(L)+H_{\tau_s}^{(1)}(L)$ decreases as the mutation rate decreases below the recombination rate, making it desirable to compute the more accurate formula $H_{\tau_s}(L)\approx H_{\tau_s}^{(0)}(L)+H_{\tau_s}^{(1)}(L)+H_{\tau_s}^{(2)}(L)$. Conceptually, computing $H_{\tau_s}^{(2)}(L)$ is no different from computing $H_{\tau_s}^{(1)}(L)$; it simply requires integrating over three coalescence times t_0,t_1,t_2, two recombination times $t_{(r,1)}<\min(t_0,t_1), t_{(r,2)}<\min(t_1,t_2)$, and the locations $L_1<L_2$ of two distinct recombination sites:

$$
\begin{aligned}
H_{\tau_s}^{(2)}(L) \;=\; &\int_{L_1=1}^{L-2}\int_{L_2=1}^{L-L_1-1}\int_{t_0=\tau_s}^{\infty}\int_{t_1=\tau_s}^{\infty}\int_{t_2=\tau_s}^{\infty}\int_{t_{(r,1)}=0}^{\min(t_0,t_1)}\int_{t_{(r,2)}=0}^{\min(t_1,t_2)} e^{-t_0}e^{-t_0L_1(\rho+\theta)}\\
&\cdot(1-e^{-t_0\theta})\cdot\rho e^{-\rho t_{(r,1)}}\cdot e^{-(t_1-t_{(r,1)})}\cdot e^{-t_1L_1(\rho+\theta)}\cdot\rho e^{-\rho t_{(r,2)}}\cdot e^{-(t_2-t_{(r,2)})}\\
&\cdot e^{-t_2(L-L_1-L_2)(\rho+\theta)}\; dt_{(r,2)}dt_{(r,1)}dt_2dt_1dt_0dL_2dL_1
\end{aligned}
$$

The result is that

$$
H_{\tau_s}^{(2)}(L)=\left(\frac{\rho}{\rho+\theta}\right)^2\left(e^{-\tau_sL(\rho+\theta)}(2p_1-p_2-p_3)-e^{-\tau_s(L(\rho+\theta)+\theta)}(2p_4-p_5-p_6)\right),
$$

where

$$
\begin{aligned}
p_1 \;=\; &\frac{1}{(1-\rho)(1+L(\rho+\theta))(2-\rho+L(\rho+\theta))}\left(\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\log(1+(L-1)(\rho+\theta))\right.\\
&-\log(1+\theta)\log\left(\frac{1-\rho+(L-1)(\rho+\theta)}{1+2\rho+\theta}\right)
\end{aligned}
$$

$$
\begin{aligned}
&-\mathrm{Li}\left(\frac{\rho+(L-2)(\rho+\theta)}{1+(L-1)(\rho+\theta)}\right)+\mathrm{Li}\left(\frac{2\rho+\theta}{1+(L-1)(\rho+\theta)}\right)\\
&-\log(1+\theta)\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\Bigg)\\
&+\frac{1}{(1-\rho)(3-2\rho+L(\rho+\theta))(2-\rho+L(\rho+\theta))}\,(\\
&-\log(2+\theta+(L-2)(\rho+\theta))\log\left(\frac{(1+(L-2)(\rho+\theta)}{1+\rho+\theta}\right)\\
&+\mathrm{Li}\left(\frac{1+(L-2)(\rho+\theta)}{2+\theta+(L-2)(\rho+\theta)}\right)-\mathrm{Li}\left(\frac{1+\rho+\theta}{2+\theta+(L-2)(\rho+\theta)}\right)\\
&-\log(1+\theta)\log\frac{1+\rho+2\theta}{\rho+\theta}+\log(1+\theta+(L-3)(\rho+\theta))\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{\rho+\theta}\right)\\
&-\mathrm{Li}\left(-\frac{1+\theta+(L-3)(\rho+\theta)}{1+\theta}\right)\\
&-\log(1+\theta+(L-3)(\rho+\theta))\log\left(\frac{2-2\rho+(L-1)(\rho+\theta)}{1+\theta}\right)\\
&+\mathrm{Li}(-1)+\log(1+t)\log(2)+\log(1+\theta)\log\left(\frac{2-2\rho+(L-1)(\rho+\theta)}{2+2\theta}\right)\\
&-\mathrm{Li}\left(-\frac{1+\theta}{\rho+\theta}\right)+\mathrm{Li}\left(-\frac{1+\theta+(L-3)(\rho+\theta)}{\rho+\theta}\right)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
p_2 \;=\;& \frac{1}{(1-\rho)(1+L(\rho+\theta))(2-\rho+L(\rho+\theta))}\Bigg(\log(1+\theta)\log\left(\frac{1+(L-2)(\rho+\theta)}{1+\rho+\theta}\right)\\
&-\log(2+\theta+(L-2)(\rho+\theta))\log\left(\frac{1+(L-2)(\rho+\theta)}{1+\rho+\theta}\right)+\mathrm{Li}\left(\frac{1+(L-2)(\rho+\theta)}{2+\theta+(L-2)(\rho+\theta)}\right)\\
&-\mathrm{Li}\left(\frac{1+\rho+\theta}{2+\theta+(L-2)(\rho+\theta)}\right)+\log(1+(L-1)(\rho+\theta))\log\left(\frac{r+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\\
&+\mathrm{Li}\left(\frac{2\rho+\theta}{1+(L-1)(\rho+\theta)}\right)-\mathrm{Li}\left(\frac{\rho+(L-2)(\rho+\theta)}{1+(L-1)(\rho+\theta)}\right)\\
&-\log(1+\theta)\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
p_3 \;=\;& \frac{1}{(1-\rho)(1+L(\rho+\theta))(2-\rho+L(\rho+\theta))}\Bigg(\log(1+(L-1)(\rho+\theta))\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)\\
&-\log(1+\theta)\log\left(\frac{\rho+(L-2)(\rho+\theta)}{2\rho+\theta}\right)+\mathrm{Li}\left(-\frac{\rho+(L-2)(\rho+\theta)}{1+\theta}\right)-\mathrm{Li}\left(-\frac{2\rho+\theta}{1+\theta}\right)\\
&-\log(1+(L-1)(\rho+\theta))\log\left(\frac{1+(L-2)(\rho+\theta)}{1+\rho+\theta}\right)+\log(\rho+\theta)\log\left(\frac{1+(L-2)(\rho+\theta)}{1+\rho+\theta}\right)\\
&-\mathrm{Li}\left(-\frac{1+(L-2)(\rho+\theta)}{\rho+\theta}\right)+\mathrm{Li}\left(-\frac{1+r+t}{r+t}\right)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
p_4 \;=\;& \frac{1}{(1-\rho)(1+\theta+L(\rho+\theta))(2-\rho+\theta+L(\rho+\theta))}\biggl(\log\left(\frac{L-1}{2}\right)\log(1+\theta+(L-1)(\rho+\theta)) \\
& -\log(1+\theta)\log\left(\frac{1+\rho+2\theta}{\rho+\theta}\right)+\log(1+\theta+(L-3)(\rho+\theta))\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{\rho+\theta}\right) \\
& -\mathrm{Li}\left(-\frac{1+\theta}{\rho+\theta}\right)+\mathrm{Li}\left(-\frac{1+\theta+(L-3)(\rho+\theta)}{\rho+\theta}\right)-\log(1+\theta)\log\left(\frac{1-\rho+(L-1)(\rho+\theta)}{1+2\rho+\theta}\right) \\
& -\mathrm{Li}\left(\frac{(L-1)(\rho+\theta)}{1+\theta+(L-1)(\rho+\theta)}\right)+\mathrm{Li}\left(\frac{2(\rho+\theta)}{1+\theta+(L-1)(\rho+\theta)}\right)-\log(1+\theta)\log\left(\frac{L-1}{2}\right)\biggr) \\
& +\frac{1}{(1-\rho)(3-2\rho+\theta+L(\rho+\theta))(2-\rho+\theta+L(\rho+\theta))}\biggl(\mathrm{Li}\left(\frac{1+\theta+(L-2)(\rho+\theta)}{2+2\theta+(L-2)(\rho+\theta)}\right) \\
& -\log(2+2\theta+(L-2)(\rho+\theta))\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{1+2\theta+\rho}\right) \\
& -\mathrm{Li}\left(\frac{1+2\theta+\rho}{2+2\theta+(L-2)(\rho+\theta)}\right)+\log(1+\theta)\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{1+2\theta+\rho}\right) \\
& -\mathrm{Li}\left(-\frac{1+\theta+(L-3)(\rho+\theta)}{1+\theta}\right) \\
& -\log(1+\theta+(L-3)(\rho+\theta))\log\left(\frac{2-2\rho+(L-1)(\rho+\theta)}{1+\theta}\right) \\
& +\mathrm{Li}(-1)+\log(1+\theta)\log(2)+\log(1+\theta)\log\left(\frac{2-2\rho+(L-1)(\rho+\theta)}{2+2\theta}\right)\biggr)
\end{aligned}
$$

$$
\begin{aligned}
p_5 \;=\;& \frac{1}{(1-\rho)(2-\rho+\theta+L(\rho+\theta))(1+\theta+L(\rho+\theta))}\biggl(\mathrm{Li}\left(\frac{1+\theta+(L-2)(\rho+\theta)}{2+2\theta+(L-2)(\rho+\theta)}\right) \\
& -\log(2+2\theta+(L-2)(\rho+\theta))\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{1+2\theta+\rho}\right) \\
& -\mathrm{Li}\left(\frac{1+2\theta+\rho}{2+2\theta+(L-2)(\rho+\theta)}\right)+\log(1+\theta)\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{1+2\theta+\rho}\right) \\
& +\log(1+\theta+(L-1)(\rho+\theta))\log\left(\frac{L-1}{2}\right) \\
& +\mathrm{Li}\left(\frac{2(\rho+\theta)}{1+\theta+(L-1)(\rho+\theta)}\right)-\mathrm{Li}\left(\frac{(L-1)(\rho+\theta)}{1+\theta+(L-1)(\rho+\theta)}\right)-\log(1+\theta)\log\left(\frac{L-1}{2}\right)\biggr)
\end{aligned}
$$

$$
\begin{aligned}
p_6 \;=\;& \frac{1}{(1-\rho)(2-\rho+\theta+L(\rho+\theta))(1+\theta+L(\rho+\theta))}\biggl(\log(1+\theta+(L-1)(\rho+\theta))\log\left(\frac{L-1}{2}\right) \\
& -\log(1+\theta)\log\left(\frac{L-1}{2}\right)+\mathrm{Li}\left(-\frac{(L-1)(\rho+\theta)}{1+\theta}\right)-\mathrm{Li}\left(-\frac{2(\rho+\theta)}{1+\theta}\right) \\
& -\log(1+\theta+(L-1)(\rho+\theta))\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{1+2\theta+\rho}\right)+\mathrm{Li}\left(-\frac{1+\rho+2\theta}{\rho+\theta}\right) \\
& +\log(\rho+\theta)\log\left(\frac{1+\theta+(L-2)(\rho+\theta)}{1+2\theta+\rho}\right)-\mathrm{Li}\left(-\frac{1+\theta+(L-2)(\rho+\theta)}{\rho+\theta}\right)\biggr)
\end{aligned}
$$

To approximate the polylog function with elementary functions, we make use of the fact that

$$\lim_{z\to\infty} \frac{\mathrm{Li}(-z)}{-\pi^2/6 - \log(z)^2/2} = 1 \tag{A.16}$$

and

$$\mathrm{Li}(-1+\epsilon) = -\frac{\pi^2}{12} + \epsilon\log(2) + O(\epsilon^2). \tag{A.17}$$

Specifically, we use (A.16) to approximate $\mathrm{Li}(z)$ for $z < -4$ and (A.17) to approximate $\mathrm{Li}(z)$ for $z > -2$. For the intermediate regime, we use the unique cubic polynomial that joins up with the left and right halves to form a function with continuous first derivative.

A.1.3 Mixed admixture status

We move on to consider a history parameterized by four parameters: a population size N and a split time τ_s, a later admixture time τ_a and an admixture fraction f that denotes the percentage of individuals in the recipient population that recently migrated over from the donor population. Given this complex history, we will calculate the frequency $H^{(1)}_{\tau_a,\tau_s,f}(L)$.

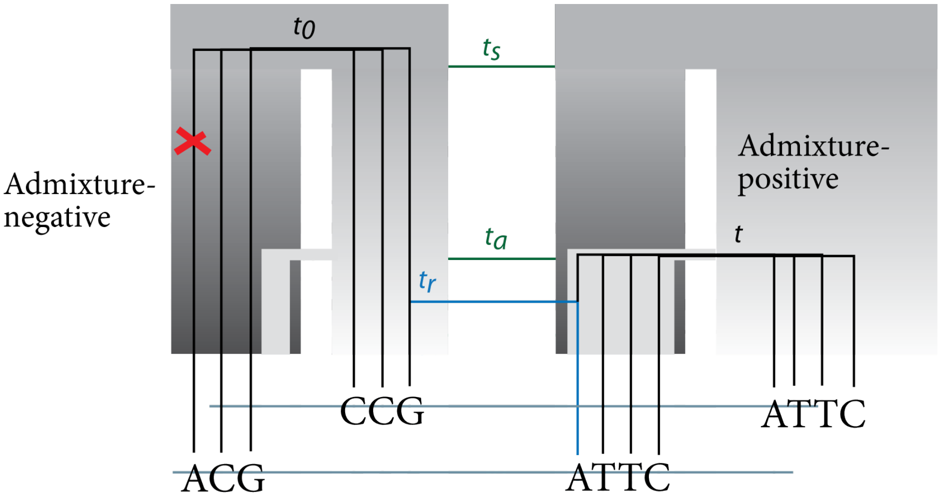


Figure A.1: This picture represents an IBS tract of mixed admixture status. The left-hand side is admixture-negative, constrained to coalesce before the divergence time τ_s, while the right-hand side is admixture positive, constrained only to coalesce before the admixture time τ_a.

As before, we must marginalize over two coalescence times t_0 and t, as well as a time of recombination $t_{(r)} < \min(t_0, t)$. In addition, we must consider the "admixture status" of each tract half: whether one of the sequences was involved in the historical migration or whether they were constrained to coalesce before the split time. We will say that a locus is admixture-positive if one of the sequences was involved in the migration and admixture-negative otherwise (see Figure A.1). This allows us to introduce the conditional coalescence density function $\zeta(t|(t_{(r)}, a))$, which denotes the coalescence time density function given that at time $t_{(r)}$, the base pair was uncoalesced with admixture status a. When $t_{(r)} < \tau_a$, the admixture status at the time of recombination is undetermined, which will be denoted '0'. Conversely, the admixture status cannot be undetermined at more ancient times of recombination $t_{(r)} > \tau_a$. For the one-pulse, constant size history considered here, $\zeta(t|(t_{(r)}, a))$ is the following:

$$\begin{aligned}
\zeta(t|(t_{(r)}, 0)) &= \zeta(t) = fe^{-(t-\tau_a)}\mathbf{1}(t \geq \tau_a, t_{(r)} < \tau_a) + (1-f)e^{-(t-\tau_s)}\mathbf{1}(t \geq \tau_s, t_{(r)} < \tau_a) \\
\zeta(t|(t_{(r)}, +)) &= e^{-(t-t_{(r)})}\mathbf{1}(\tau_a \leq t_{(r)} < t) \\
\zeta(t|(t_{(r)}, -) &= e^{-(t-\max(t_{(r)},\tau_s))}\mathbf{1}(\tau_a \leq t_{(r)} < t)
\end{aligned}$$

If we let a denote the admixture status at the time of recombination, then

$$\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L) &= \sum_{L_r=1}^{L-1} \sum_{a\in\{+,-,0\}} \int_{t_0=0}^{\infty}\int_{t=0}^{\infty}\int_{t_{(r)}=0}^{\min(t_0,t)} \zeta(t_0)e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta})\cdot\rho e^{-\rho t_{(r)}} \\
&\quad \cdot\mathbb{P}(a|t_0,t_{(r)})\cdot\zeta(t|(t_{(r)},a))e^{-t(L-L_r)(\rho+\theta)+t\rho}dt_{(r)}dtdt_0 \\
&\approx \sum_{a\in\{+,-,0\}} \int_{L_r=1}^{L-1}\int_{t_0=0}^{\infty}\int_{t=0}^{\infty}\int_{t_{(r)}=0}^{\min(t_0,t)} \zeta(t_0)e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta})\cdot\rho e^{-\rho t_{(r)}} \\
&\quad \cdot\mathbb{P}(a|t_0,t_{(r)})\cdot\zeta(t|(t_{(r)},a))e^{-t(L-L_r)(\rho+\theta)+t\rho}dL_rdt_{(r)}dtdt_0.
\end{aligned}$$

Therefore, our goal is to evaluate the expression

$$\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L) &= \sum_{a\in\{+,-,0\}} \int_{L_r=1}^{L-1}\int_{t_0=0}^{\infty}\int_{t=0}^{\infty}\int_{t_{(r)}=0}^{\min(t_0,t)} \zeta(t_0)e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta})\cdot\rho e^{-\rho t_{(r)}} \\
&\quad \cdot\mathbb{P}(a|t_0,t_{(r)})\cdot\zeta(t|(t_{(r)},a))e^{-t(L-L_r)(\rho+\theta)+t\rho}dL_rdt_{(r)}dtdt_0
\end{aligned}$$

in closed form. Marginalizing over the admixture status at the time of recombination yields that

$$\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L) &= \int_{L_r=1}^{L-1}\int_{t_0=0}^{\infty}\int_{t=0}^{\infty}\int_{t_{(r)}=0}^{\tau_a} (fe^{-(t_0-\tau_a)}\mathbf{1}(t_0\geq\tau_a) + (1-f)e^{-(t_0-\tau_s)}\mathbf{1}(t_0\geq\tau_s)) \\
&\quad \cdot e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta})\cdot\rho e^{-\rho t_{(r)}}\cdot(fe^{-(t-\tau_a)}\mathbf{1}(t\geq\tau_a) + (1-f)e^{-(t-\tau_s)}\mathbf{1}(t\geq\tau_s)) \\
&\quad \cdot e^{-t(L-L_r)(\rho+\theta)+t\rho}dL_rdt_{(r)}dtdt_0 \\
&\quad + \int_{L_r=1}^{L-1}\int_{t_0=\tau_a}^{\infty}\int_{t=\tau_a}^{\infty}\int_{t_{(r)}=\tau_a}^{\min(t_0,t)} fe^{-(t_0-\tau_a)}e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta})\cdot\rho e^{-\rho t_{(r)}}
\end{aligned}$$

$$
\begin{aligned}
&\cdot e^{-(t-t_{(r)})} e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0 \\
&+ \int_{L_r=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=\tau_a}^{\min(t_0,t)} (1-f) e^{-(t_0-\tau_a)} e^{-t_0 L_r(\rho+\theta)} (1-e^{-t_0\theta}) \cdot \rho e^{-\rho t_{(r)}} \\
&\cdot e^{-(t-\max(\tau_s,t_{(r)}))} e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0
\end{aligned}
$$

The easiest integrals to compute are the ones where the whole segment has the same admixture status. These scenarios split into two classes: one where the recombination is more recent than τ_a (class R) and one where the one where the recombination is more ancient than τ_a (class A). When admixture status is negative it also matters whether recombination happened before or after τ_s, which divides class A into two subclasses A_i (intermediate recombination times, $\tau_a \leq t_{(r)} < \tau_s$) and A_a (ancient recombination times, $t_{(r)} \geq \tau_s$). The scenarios with mixed admixture status involve exponential integrals, which cannot be computed in closed form and must be approximated further. In total, there are seven integrals we must do:

$$
\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L) \quad = \quad & H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,+) + H^{(1)}_{\tau_a,\tau_s,f}(L,+,A,+) + H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,-) \\
& + H^{(1)}_{\tau_a,\tau_s,f}(L,-,A_i,-) + H^{(1)}_{\tau_a,\tau_s,f}(L,-,A_a,-) + H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,-) \\
& + H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,+),
\end{aligned}
$$

where

$$
\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,+) \quad :=& \quad \int_{L_r=1}^{L-1} \int_{t_0=\tau_a}^{\infty} \int_{t=\tau_a}^{\infty} \int_{t_{(r)}=0}^{\tau_a} f e^{-(t_0-\tau_a)} \cdot e^{-t_0 L_r(\rho+\theta)} (1-e^{-t_0\theta}) \\
& \cdot \rho e^{-\rho t_{(r)}} \cdot f e^{-(t-\tau_a)} \cdot e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0 \\
H^{(1)}_{\tau_a,\tau_s,f}(L,+,A,+) \quad :=& \quad \int_{L_r=1}^{L-1} \int_{t_0=\tau_a}^{\infty} \int_{t=\tau_a}^{\infty} \int_{t_{(r)}=\tau_a}^{\min(t_0,t)} f e^{-(t_0-\tau_a)} \cdot e^{-t_0 L_r(\rho+\theta)} (1-e^{-t_0\theta}) \\
& \cdot \rho e^{-\rho t_{(r)}} \cdot e^{-(t-t_{(r)})} \cdot e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0 \\
H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,-) \quad :=& \quad \int_{L_r=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=0}^{\tau_a} (1-f) e^{-(t_0-\tau_s)} \cdot e^{-t_0 L_r(\rho+\theta)} (1-e^{-t_0\theta}) \\
& \cdot \rho e^{-\rho t_{(r)}} \cdot (1-f) e^{-(t-\tau_s)} \cdot e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0 \\
H^{(1)}_{\tau_a,\tau_s,f}(L,-,A_i,-) \quad :=& \quad \int_{L_r=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=\tau_a}^{\tau_s} (1-f) e^{-(t_0-\tau_s)} \cdot e^{-t_0 L_r(\rho+\theta)} (1-e^{-t_0\theta}) \\
& \cdot \rho e^{-\rho t_{(r)}} \cdot e^{-(t-\tau_s)} \cdot e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0 \\
H^{(1)}_{\tau_a,\tau_s,f}(L,-,A_a,-) \quad :=& \quad \int_{L_r=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=\tau_s}^{\min(t_0,t)} (1-f) e^{-(t_0-\tau_s)} \cdot e^{-t_0 L_r(\rho+\theta)} (1-e^{-t_0\theta}) \\
& \cdot \rho e^{-\rho t_{(r)}} \cdot e^{-(t-\tau_s)} \cdot e^{-t(L-L_r)(\rho+\theta)+t\rho} dL_r dt_{(r)} dt dt_0
\end{aligned}
$$

$$
\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,-) \;:=\; & \int_{L_r=1}^{L-1}\int_{t_0=\tau_a}^{\infty}\int_{t=\tau_s}^{\infty}\int_{t_{(r)}=0}^{\tau_a} fe^{-(t_0-\tau_a)}\cdot e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta}) \\
& \cdot\rho e^{-\rho t_{(r)}}\cdot(1-f)e^{-(t-\tau_s)}\cdot e^{-t(L-L_r)(\rho+\theta)+t\rho}dL_r dt_{(r)}dtdt_0 \\
H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,+) \;:=\; & \int_{L_r=1}^{L-1}\int_{t_0=\tau_s}^{\infty}\int_{t=\tau_a}^{\infty}\int_{t_{(r)}=0}^{\tau_a} (1-f)e^{-(t_0-\tau_s)}\cdot e^{-t_0L_r(\rho+\theta)}(1-e^{-t_0\theta}) \\
& \cdot\rho e^{-\rho t_{(r)}}\cdot fe^{-(t-\tau_a)}\cdot e^{-t(L-L_r)(\rho+\theta)+t\rho}dL_r dt_{(r)}dtdt_0
\end{aligned}
$$

The first five of these integrals can be evaluated in closed form with no further approximation. After some algebra, we find that

$$
\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,+) \;=\; & f^2e^{-\tau_aL(\rho+\theta)}\log\Bigg(\frac{(1+(L-1)(\rho+\theta)-\rho)(1+(L-1)(\rho+\theta)}{(1+\theta)(1+\rho+\theta)}\Bigg) \\
& \cdot\frac{1-e^{-\tau_a\rho}}{(\rho+\theta)(2+L(\rho+\theta)-\rho)} \\
& -f^2e^{-\tau_a(L(\rho+\theta)+\theta)}\log\Bigg(\frac{(1+(L-1)(\rho+\theta))^2}{(1+\theta)(1+\rho+2\theta)}\Bigg) \\
& \cdot\frac{(1-e^{-\tau_a\rho})}{(\rho+\theta)(2+L(\rho+\theta)-\rho+\theta)} \\
H^{(1)}_{\tau_a,\tau_s,f}(L,+,A,+) \;=\; & \frac{f\rho e^{-\tau_aL(\rho+\theta)}}{(1-\rho)(\rho+\theta)}(\\
& \log\Bigg(\frac{1+(L-1)(\rho+\theta)}{1+\rho+\theta}\Bigg)\Bigg(\frac{1}{1+L(\rho+\theta)}-\frac{1}{2+L(\rho+\theta)-\rho}\Bigg) \\
& -\log\Bigg(\frac{1+L(\rho+\theta)-\rho}{1+\rho+2\theta}\Bigg)\Bigg(\frac{e^{-\tau_a\theta}}{1+L(\rho+\theta)+\theta}-\frac{e^{-\tau_a\theta}}{2+L(\rho+\theta)+\theta-\rho}\Bigg) \\
& +\log\Bigg(\frac{1+(L-1)(\rho+\theta)-\rho}{1+\theta}\Bigg)\Bigg(\frac{1}{1+L(\rho+\theta)}-\frac{e^{-\tau_a\theta}}{1+L(\rho+\theta)+\theta} \\
& -\frac{1}{2+L(\rho+\theta)-\rho}+\frac{e^{-\tau_a\theta}}{2+L(\rho+\theta)+\theta-\rho}\Bigg)\Bigg) \\
H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,-) \;=\; & (1-f)^2e^{-\tau_sL(\rho+\theta)}\log\Bigg(\frac{(1+(L-1)(\rho+\theta)-\rho)(1+(L-1)(\rho+\theta)}{(1+\theta)(1+\rho+\theta)}\Bigg) \\
& \cdot\frac{1-e^{-\tau_a\rho}}{(\rho+\theta)(2+L(\rho+\theta)-\rho)} \\
& -(1-f)^2e^{-\tau_s(L(\rho+\theta)+\theta)}\cdot \\
& \log\Bigg(\frac{((1+(L-1)(\rho+\theta)-\rho)(1+(L-1)(\rho+\theta))}{(1+\theta)(1+\rho+2\theta)}\Bigg) \\
& \cdot\frac{(1-e^{-\tau_a\rho})}{(\rho+\theta)(2+L(\rho+\theta)-\rho+\theta)} \\
H^{(1)}_{\tau_a,\tau_s,f}(L,-,A_i,-) \;=\; & (1-f)e^{-\tau_sL(\rho+\theta)}\log\Bigg(\frac{(1+(L-1)(\rho+\theta)-\rho)(1+(L-1)(\rho+\theta)}{(1+\theta)(1+\rho+\theta)}\Bigg) \\
& \cdot\frac{e^{-\tau_a\rho}-e^{-\tau_s\rho}}{(\rho+\theta)(2+L(\rho+\theta)-\rho)}
\end{aligned}
$$

$$\begin{aligned}
&-(1-f)e^{-\tau_a(L(\rho+\theta)+\theta)}\\
&\cdot\log\left(\frac{((1+(L-1)(\rho+\theta)-\rho)(1+(L-1)(\rho+\theta))}{(1+\theta)(1+\rho+2\theta)}\right)\\
&\cdot\frac{e^{-\tau_a\rho}-e^{-\tau_s\rho}}{(\rho+\theta)(2+L(\rho+\theta)-\rho+\theta)}
\end{aligned}$$

$$\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,-,A_a,-) \quad = \quad & \frac{(1-f)\rho e^{-\tau_s L(\rho+\theta)}}{(1-\rho)(\rho+\theta)}(\\
& \log\left(\frac{1+(L-1)(\rho+\theta)}{1+\rho+\theta}\right)\left(\frac{1}{1+L(\rho+\theta)}-\frac{1}{2+L(\rho+\theta)-\rho}\right)\\
& -\log\left(\frac{1+L(\rho+\theta)-\rho}{1+\rho+2\theta}\right)\left(\frac{e^{-\tau_s\theta}}{1+L(\rho+\theta)+\theta}-\frac{e^{-\tau_s\theta}}{2+L(\rho+\theta)+\theta-\rho}\right)\\
& +\log\left(\frac{1+(L-1)(\rho+\theta)-\rho}{1+\theta}\right)\left(\frac{1}{1+L(\rho+\theta)}-\frac{e^{-\tau_s\theta}}{1+L(\rho+\theta)+\theta}\right.\\
& \left.-\frac{1}{2+L(\rho+\theta)-\rho}+\frac{e^{-\tau_s\theta}}{2+L(\rho+\theta)+\theta-\rho}\right)\Bigg)
\end{aligned}$$

The last terms $H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,-)$ and $H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,+)$, as mentioned before, involve exponential integrals that cannot be reduced to elementary functions. We will approximate them using a standard tight bracketing of the exponential integral. Since

$$\frac{1}{2}e^{-x}\log\left(1+\frac{2}{x}\right)<\int_x^\infty\frac{e^{-\tau}}{\tau}d\tau<e^{-x}\log\left(1+\frac{1}{x}\right),$$

(see Abramovitz & Stegun (1964)), we will let

$$\in \tau_a^b\frac{e^{-\tau}}{\tau}d\tau\approx e^{-a}\log\left(1+\frac{1}{a}\right)-\frac{1}{2}e^{-b}\log\left(1+\frac{2}{b}\right)$$

and use this to derive the approximation

$$\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,+) \quad = \quad & \frac{f(1-f)(1-e^{-\tau_a\rho})}{\rho+\theta}\left(\frac{1}{2+L(\rho+\theta)-\rho}\right(\\
& e^{\tau_s-\tau_a(1+L(\rho+\theta)-\rho)}\int_{u=(\tau_s-\tau_a)(1+\rho+\theta)}^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta))}\frac{e^{-u}}{u}du\\
& \left.+\,e^{\tau_a-\tau_s(1+L(\rho+\theta)-\rho)}\int_{u=(\tau_s-\tau_a)(1+\theta)}^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\frac{e^{u}}{u}du\right)\\
& -\frac{1}{2+L(\rho+\theta)+\theta-\rho}\Bigg(e^{\tau_s-\tau_a(1+L(\rho+\theta)+\theta-\rho)}\int_{u=(\tau_s-\tau_a)(1+2\theta+\rho)}^{(\tau_s-\tau_a)(1+L(\rho+\theta)-\rho)}\frac{e^{-u}}{u}du
\end{aligned}$$

$$\begin{aligned}
&+e^{\tau_a-\tau_s(1+L(\rho+\theta)+\theta-\rho)}\int_{u=(\tau_s-\tau_a)(1+\theta)}^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\frac{e^u}{u}du\Bigg)\Bigg)\Bigg)\\
\approx\quad&\frac{f(1-f)(1-e^{-\tau_a\rho})}{\rho+\theta}\Bigg(\frac{1}{2+L(\rho+\theta)-\rho}(\\
&e^{\tau_s-\tau_a(1+L(\rho+\theta)-\rho)}\Bigg(e^{-(\tau_s-\tau_a)(1+\rho+\theta)}\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+\rho+\theta)}\Bigg)\\
&-\frac{1}{2}e^{-(\tau_s-\tau_a)(1+(L-1)(\rho+\theta))}\log\Bigg(1+\frac{2}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta))}\Bigg)\Bigg)\\
&+e^{\tau_a-\tau_s(1+L(\rho+\theta)-\rho)}(\\
&e^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho}\Bigg)\\
&-\frac{1}{2}e^{(\tau_s-\tau_a)(1+\theta)}\log\Bigg(1+\frac{2}{(\tau_s-\tau_a)(1+\theta)}\Bigg)\Bigg)\\
&-\frac{1}{2+L(\rho+\theta)+\theta-\rho}\Bigg(e^{\tau_s-\tau_a(1+L(\rho+\theta)+\theta-\rho)}\\
&\Bigg(e^{-(\tau_s-\tau_a)(1+2\theta+\rho)}\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+2\theta+\rho)}\Bigg)\\
&-\frac{1}{2}e^{-(\tau_s-\tau_a)(1+L(\rho+\theta)-\rho)}\log\Bigg(1+\frac{2}{1+L(\rho+\theta)-\rho)}\Bigg)\Bigg)\\
&+e^{\tau_a-\tau_s(1+L(\rho+\theta)-\rho)}\Bigg(e^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\\
&\cdot\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\Bigg)\\
&-\frac{1}{2}e^{(\tau_s-\tau_a)(1+\theta)}\log\Bigg(1+\frac{2}{1+\theta}\Bigg)\Bigg)\Bigg)\Bigg)\Bigg),
\end{aligned}$$

which simplifies to

$$\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,-,R,+)\quad=\quad&\frac{f(1-f)(1-e^{-\tau_a\rho})}{\rho+\theta}\Bigg(\frac{e^{-\tau_a(L(\rho+\theta)-\rho)}}{2+L(\rho+\theta)-\rho}(\\
&e^{-(\tau_s-\tau_a)(\rho+\theta)}\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+\rho+\theta)}\Bigg)\\
&-\frac{1}{2}e^{-(\tau_s-\tau_a)(L-1)(\rho+\theta)}\log\Bigg(1+\frac{2}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta))}\Bigg)\\
&+e^{-(\tau_s-\tau_a)(\rho+\theta)}\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\Bigg)\\
&-\frac{1}{2}e^{-(\tau_s-\tau_a)(L-1)(\rho+\theta)}\log\Bigg(1+\frac{2}{(\tau_s-\tau_a)(1+\theta)}\Bigg)\Bigg)\\
&-\frac{e^{-\tau_a(L(\rho+\theta)+\theta-\rho)}}{2+L(\rho+\theta)+\theta-\rho}\Bigg(e^{-(\tau_s-\tau_a)(2\theta+\rho)}\log\Bigg(1+\frac{1}{(\tau_s-\tau_a)(1+2\theta+\rho)}\Bigg)\\
&-\frac{1}{2}e^{-(\tau_s-\tau_a)(L(\rho+\theta)-\rho)}\log\Bigg(1+\frac{2}{(\tau_s-\tau_a)(1+L(\rho+\theta)-\rho)}\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&+e^{-(\tau_s-\tau_a)(2\theta+\rho)}\log\left(1+\frac{1}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\right)\\
&-\frac{1}{2}e^{-(\tau_s-\tau_a)(L(\rho+\theta)-\rho)}\log\left(1+\frac{2}{(\tau_s-\tau_a)(1+\theta)}\right)\Bigg)\Bigg)\Bigg).
\end{aligned}
$$

We can approximate $H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,-)$ in the exact same way:

$$
\begin{aligned}
H^{(1)}_{\tau_a,\tau_s,f}(L,+,R,-) \;=\;& \frac{f(1-f)(1-e^{-\tau_a\rho})}{\rho+\theta}\left(\frac{1}{2+L(\rho+\theta)-\rho}\right(\\
& e^{\tau_s-\tau_a(1+L(\rho+\theta)-\rho)}\int_{u=(\tau_s-\tau_a)(1+\theta)}^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\frac{e^{-u}}{u}du\\
& + e^{\tau_a-\tau_s(1+L(\rho+\theta)-\rho)}\int_{u=(\tau_s-\tau_a)(1+\rho+\theta)}^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\frac{e^{u}}{u}du\Bigg)\\
& -\frac{1}{2+L(\rho+\theta)+\theta-\rho}\Bigg(\\
& e^{\tau_s-\tau_a(1+L(\rho+\theta)+\theta-\rho)}\int_{u=(\tau_s-\tau_a)(1+\theta)}^{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\frac{e^{-u}}{u}du\\
& +e^{\tau_a-\tau_s(1+L(\rho+\theta)+\theta-\rho)}\int_{u=(\tau_s-\tau_a)(1+2\theta+\rho)}^{(\tau_s-\tau_a)(1+L(\rho+\theta)-\rho)}\frac{e^{u}}{u}du\Bigg)\Bigg)\Bigg)\\
\approx\;& \frac{f(1-f)(1-e^{-\tau_a\rho})}{\rho+\theta}\left(\frac{e^{-\tau_a(L(\rho+\theta)-\rho)}}{2+L(\rho+\theta)-\rho}\right(\\
& e^{-(\tau_s-\tau_a)\theta}\log\left(1+\frac{1}{(\tau_s-\tau_a)(1+\theta)}\right)\\
& -\frac{1}{2}e^{-(\tau_s-\tau_a)((L-1)(\rho+\theta)-\rho)}\log\left(1+\frac{2}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\right)\\
& +e^{-(\tau_s-\tau_a)\theta}\log\left(1+\frac{1}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta))}\right)\\
& -\frac{1}{2}e^{-(\tau_s-\tau_a)((L-1)(\rho+\theta)-\rho)}\log\left(1+\frac{2}{(\tau_s-\tau_a)(1+\theta+\rho)}\right)\Bigg)\\
& -\frac{e^{-\tau_a(L(\rho+\theta)+\theta-\rho)}}{2+L(\rho+\theta)+\theta-\rho}\left(e^{-(\tau_s-\tau_a)\theta}\log\left(1+\frac{1}{(\tau_s-\tau_a)(1+\theta)}\right)\right.\\
& -\frac{1}{2}e^{-(\tau_s-\tau_a)((L-1)(\rho+\theta)-\rho)}\log\left(1+\frac{2}{(\tau_s-\tau_a)(1+(L-1)(\rho+\theta)-\rho)}\right)\\
& +e^{-(\tau_s-\tau_a)\theta}\log\left(1+\frac{1}{(\tau_s-\tau_a)(1+L(\rho+\theta)-\rho)}\right)\\
& -\frac{1}{2}e^{-(\tau_s-\tau_a)((L-1)(\rho+\theta)-\rho)}\log\left(1+\frac{2}{(\tau_s-\tau_a)(1+2\theta+\rho)}\right)\Bigg)\Bigg).
\end{aligned}
$$

A.2 Simulated admixture histories

To generate each colored curve in Figure 2.2 of the main text, we used MS to simulate 4.8×10^{10} bases of pairwise sequence alignment assuming a mutation rate of 2.5×10^{-8} per site per generation and a recombination rate of 1.0×10^{-8} per site per generation. Letting t_a denote the admixture time in units of $2N$ generations, the data were generated with the following command line:

```
./ms 2 4800 -t 10000 -r 4000 100000 -I 2 1 1 -es ta/2 1 0.95
-ej ta/2+0.000001 3 2 -ej 0.05 2 1
```

For each of $t_a = 200$ generations and $t_a = 400$ generations, we simulated 100 replicate datasets and inferred the set of demographic parameters t_a, t_s, f, and N. The mean and variance of the estimates for each parameter are reported in Table 2.1 of the main text. Figures A.2 and A.3 plot the full histogram of values estimated for each of the four parameters, suggesting that the times are estimated consistently, the admixture fraction is slightly biased downward, and the effective population size is biased downward more significantly.

A.3 Analysis of human data

A.3.1 Generating empirical tract spectra

We generated empirical spectra of IBS tract lengths from the 1000 Genomes pilot sequences reported in 1000 Genomes Project (2010) and available at:

```
ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2010_07/
```

We used VCF-tools to extract the haplotype sequences from the VCF files encoding the low-coverage haplotypes and trio haplotypes Danacek et al. (2011). We then used the CEU, CHBJPT, and YRI mask files available at

```
 ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_03/pilot1/
supporting/README_callability_masks
```

to excise all haplotype regions that were at least 10,000 bases long and annotated as inaccessible for SNP calling in any of the three low coverage data sets. In addition, we excised all regions annotated as gaps on the UCSC genome browser, a list available at:

```
http://cistrome.dfci.harvard.edu/browser/cgi-bin/hgTables
```

After these annotated gaps were removed, the remaining genome contained some conspicuously long regions with few or no SNP calls in the low coverage data, meaning a large fraction of the 358 total haplotypes were IBS. By visual inspection on the 1000 genomes browser, many of these SNP deserts had disappeared with the addition of new individuals sequenced after completion of the pilot phase, indicating that their sparsity of SNP calls was probably

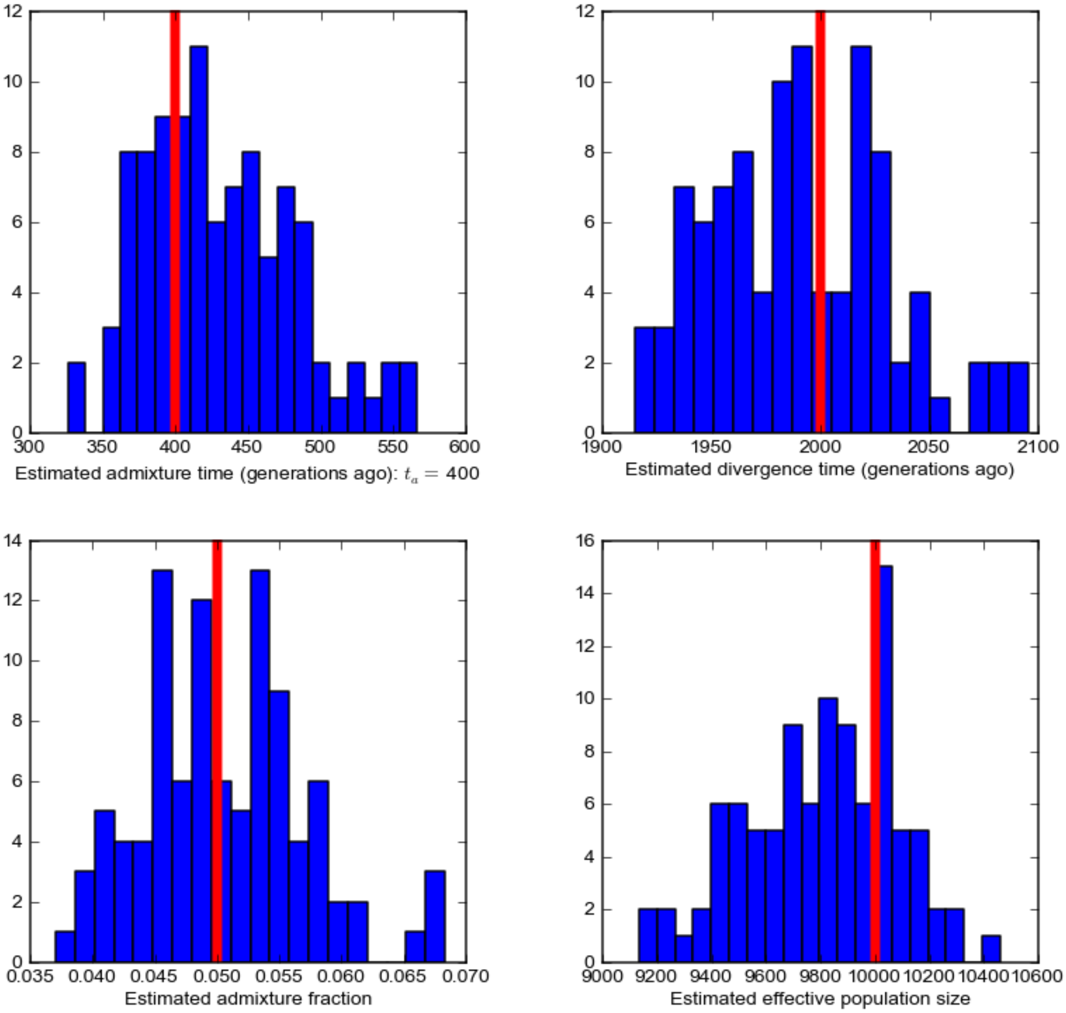


Figure A.2: Each of these histograms was generated from 100 simple admixture history datasets that were simulated with gene flow time $t_a = 400$. The true parameter value is shown in red. All parameter estimates have low variance, and all appear consistent with the exception of effective population size.

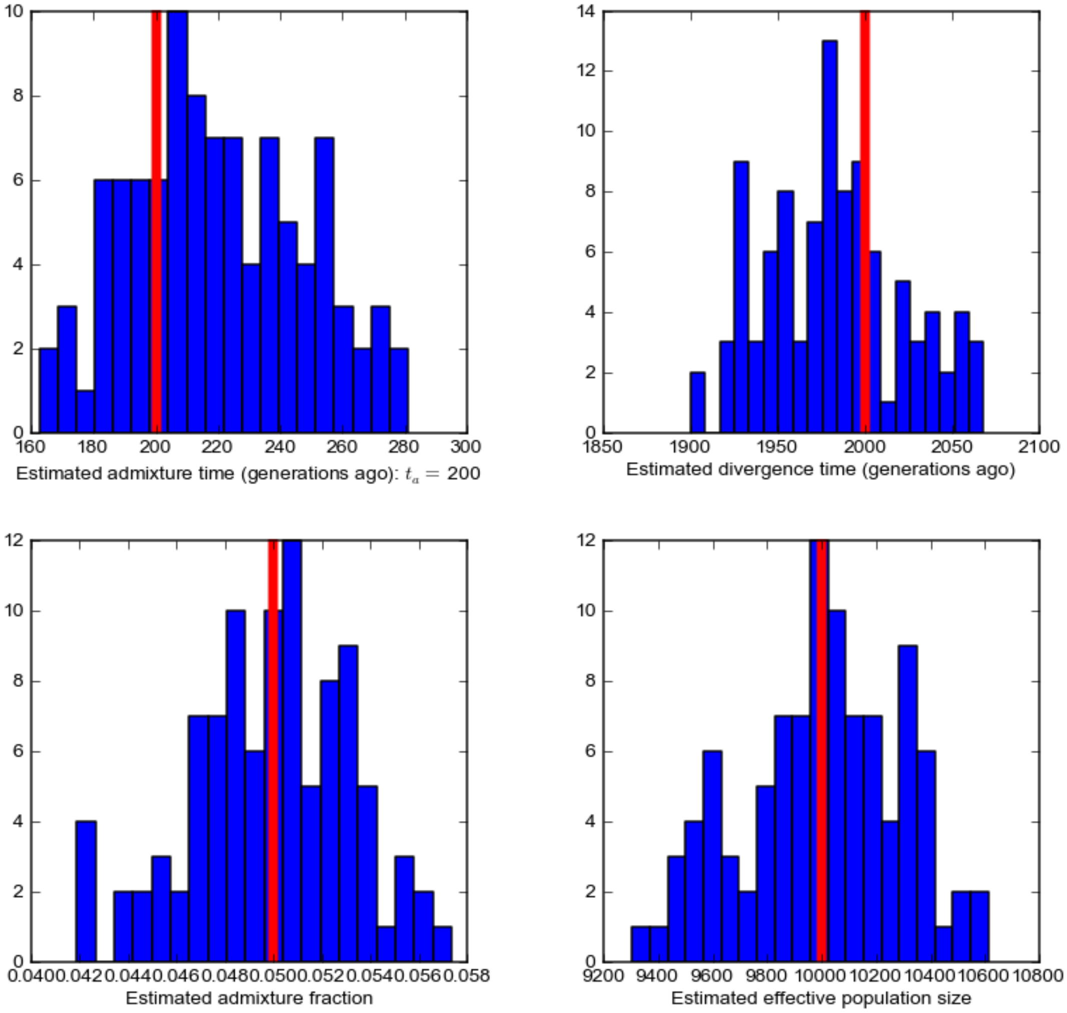


Figure A.3: These histograms record the distribution of parameter estimates for 100 simple admixture histories with gene flow time $t_a = 200$. The effective population size is estimated here with a greater downward bias than for the older admixture time $t_a = 400$. Smaller effective population size inflates the abundance of long tracts, an effect that is counteracted by a downward bias in the admixture fraction estimate.

a sequencing artifact. We therefore excised each 10^6-base region of the genome that did not have at least 66 SNP calls in each of the CEU, YRI, and JPTCHB low coverage data sets.

For the remaining portion of the genome, we generated within-population IBS length spectra as follows: For each low coverage population, we numbered the 120 haplotypes with consecutive integers and aligned haplotype n with haplotype $n+1$. This generated a total of 119 whole-genome alignments, totaling 3.05×10^{11} base pairs, which was cut up into IBS fragments at each of the sites where the two haplotypes differed. For each pair of populations, we also generated a between-population IBS length spectrum by aligning haplotype n from population A with haplotype n from population B, yielding 120 whole-genome alignments that totaled 3.08×10^{11} base pairs. Each alignment was parsed into an IBS length spectrum by cutting it up at the sites where the two haplotypes differed and sorting the resulting fragments by length.

The four parental haplotypes from each trio were numbered 1,2,3,4, and all six pairwise alignments (1 paired with 2, 1 paired with 3, etc.) were used to create the within-population tract spectra. All twelve possible pairwise alignments were used to create the spectrum of CEU-YRI trio sharing.

A.3.2 IBS tracts in low coverage data

For the CEU and YRI populations, we looked at IBS tract sharing within two subsets of the 1000 Genomes pilot data: four high quality whole genome haplotypes from the trio parents and 120 whole-genome haplotypes sequenced at low coverage. We were able to account for the excess of long IBS tracts in the high coverage trios by modeling the distribution of excess errors in the low coverage data (see Figure 2.5 of the main text).

One difference between the trios and the low coverage sequences was that the low coverage alignments had higher mean heterozygosity. We found that the YRI low coverage data had a mean heterozygosity of 8.47×10^{-4}, while the YRI trio parents had a mean heterozygosity of 6.98×10^{-4}. To determine whether the difference was significant, we bootstrapped the low coverage data as follows: for $0 \leq n \leq 30$, we subsampled haplotypes $4n, 4n+1, 4n+2$, and $4n+3$ from the low coverage YRI data and determined their shared IBS tract spectrum in the same way that was done with the four trio haplotypes. These bootstrapped low coverage data sets had mean heterozygosity 8.23×10^{-4} with standard deviation 1.81×10^{-5}, making the trio heterozygosity significantly lower. When we bootstrapped the CEU sequences in the same way, the subsample heterozygosities had mean 6.84×10^{-4} and standard deviation 2.3×10^{-5}. In contract, the CEU trio parents had mean heterozygosity 5.50×10^{-4}. In both populations, the low coverage data had an excess heterozygosity between 1.25×10^{-4} and 1.35×10^{-4}, probably due to sequencing errors. The mean heterozygosity between CEU and YRI was also higher in the low coverage data, at 9.33×10^{-4} compared to 8.05×10^{-4} in the trio data.

An error rate of 10^{-4} per base pair would destroy most 10^5- and 10^6-base IBS tracts if the errors were evenly Poisson-distributed throughout the low coverage sequences. However, the situation is somewhat better because of the imputation that was used to generate the low coverage 1000 Genomes sequences, preferentially calling haplotypes that are IBS with one another in regions where both appear IBD with one of the HapMap references. For this reason, as well as the empirical abundance of long IBS tracts in the low coverage data, we expect true IBS tracts in the low coverage alignments to be broken up by errors at some rate $\epsilon_{\rm IBS} \ll 10^{-4}$.

Let $f^{\rm YRI}_{\rm trio}(L)$ (resp. $f^{\rm YRI}_{\rm lc}$) be the frequency of differences between two high coverage YRI samples (resp. two low coverage YRI samples) that are followed by exactly L bases of IBS. If IBS tracts are accurately observed in the trio data but broken up by errors at rate $\epsilon^{\rm YRI}_{\rm IBS}$ in the low coverage data, then we should expect that $f^{\rm YRI}_{\rm lc}(L) \approx f^{\rm YRI}_{\rm trio}(L)e^{-L\epsilon^{\rm YRI}_{\rm IBS}}$. In this way, the data are consistent with an error rate of $\epsilon^{\rm YRI}_{\rm IBS} = 5 \times 10^{-6}$, with the function $f^{\rm YRI}_{\rm trio}(L)e^{-L\epsilon^{\rm YRI}_{\rm IBS}}$ falling within the realm of variation of $f^{\rm YRI}_{\rm lc}(L)$ in the bootstrapped low coverage datasets. The frequencies of long IBS tracts in the trio data are consistent with an identical error rate of $\epsilon^{\rm CEU}_{\rm IBS} = 5 \times 10^{-6}$ (see Figure A.4).

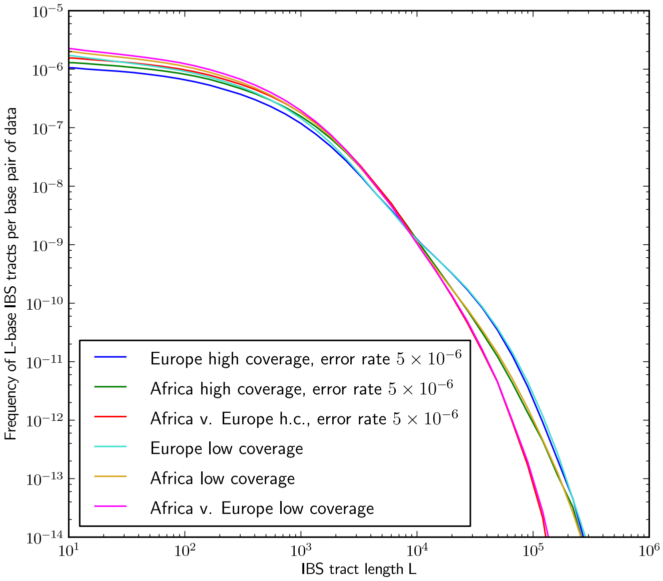


Figure A.4: This figure plots the low coverage IBS tract frequencies $f_{\rm lc}(L)$ along with the error-degraded high coverage trio frequencies $f_{\rm hc}(L)e^{-L\epsilon_{\rm IBS}}$. For $L > 1000$, we can see that $f^{\rm CEU}_{\rm lc}(L) \approx f^{\rm CEU}_{\rm hc}(L)e^{-L\epsilon^{\rm CEU}_{\rm IBS}}$, $f^{\rm YRI}_{\rm lc}(L) \approx f^{\rm YRI}_{\rm hc}(L)e^{-L\epsilon^{\rm YRI}_{\rm IBS}}$, and $f^{\rm CEU\text{-}YRI}_{\rm lc}(L) \approx f^{\rm CEU\text{-}YRI}_{\rm hc}(L)e^{-L\epsilon^{\rm CEU}_{\rm IBS}}$ when we let $\epsilon^{\rm CEU}_{\rm IBS} = \epsilon^{\rm YRI}_{\rm IBS} = 5 \times 10^{-6}$.

A.3.3 Human evolutionary model

The simulated human data plotted in Figure 2.8 of the main text were generated using the following MS command line:

```
./ms 2 1800 -t 10000 -r 4000 100000 -I 2 2 2 -en 0 2 N0_YRI -en t0_YRI 2 N2
-eN t_ancient N3 -en 0 1 N0_CEU -en t0_CEU 1 N1_CEU -en t_s 1 N2
-es t_m 1 (1-f_{Eu-As}/4) -es t_m 2 (1-f_{Eu-As}/4) -ej t_m*1.00001 3 2
-ej t_m*1.00001 4 1
-es t_m 1 1-f_{Eu-As}/4 -es t_m+(t_s-t_m)/4 2 1-f_{Eu-As}/4
-ej (t_m+(t_s-t_m)/4)*1.00001 5 2 -ej (t_m+(t_s-t_m)/4)*1.00001 6 1
-es (t_m+(t_s-t_m)/2) 1 (1-f_{Eu-As}/4)
-es (1-f_{Eu-As}/4) 2 (1-f_{Eu-As}/4)
-ej (t_m+(t_s-t_m)/2)*1.00001 7 2 -ej (t_m+(t_s-t_m)/2)*1.00001 8 1
-es (t_m+(t_s-t_m)*3/4) 1 (1-f_{Eu-As}/4)
-es (t_m+(t_s-t_m)*3/4) 2 (1-f_{Eu-As}/4)
-ej (t_m+(t_s-t_m)*3/4)*1.00001 9 2 -ej (t_m+(t_s-t_m)*3/4)*1.00001 10 1
-es t_s 1 1-f_{ghost} -ej t_s*1.00001 11 2 -ej t_ghost 1 2
```

The migration rate $m_{\text{Eu-As}}$ in Table 2.2 of the main text is calculated such that $f_{\text{Eu-As}} = m_{\text{Eu-As}}(t_s - t_m)$, letting continuous migration be approximated by four evenly spaced discrete pulses.

To estimate parameters efficiently, we used a two-step procedure. First, we estimate N_2, N_3, and t_{ancient} by fitting a two-epoch history to the IBS sharing with the YRI. Next, we fix N_2, N_3, and t_{ancient} and estimate the rest of the parameters by maximizing the likelihood of all three informative spectra: within the YRI, within the CEU and between YRI and CEU.

A.3.4 Assessing uncertainty via simulation

We simulated 27 replicate datasets under the human evolutionary model described in section A.3.3 with the maximum likelihood parameters inferred from the trio data. We then estimated parameters from each replicate dataset to gauge our accuracy at inferring a complex history. Figure A.5 illustrates the differences between the parameters inferred from the trio data and the mean estimates obtained from replicate simulations. Figures A.6, A.7 and A.8 record the full distribution of parameter estimates obtained from simulated data.

A.4 Comparison to ∂a∂i (Diffusion approximations for demographic inference)

A.4.1 Simulated data

Like our method, the program ∂a∂i can compute a composite likelihood of genomic data given a wide range of parametric histories Gutenkunst et al. (2009). We therefore evaluated ∂a∂i's ability to infer the parameters of the simple admixture history considered in the main text.

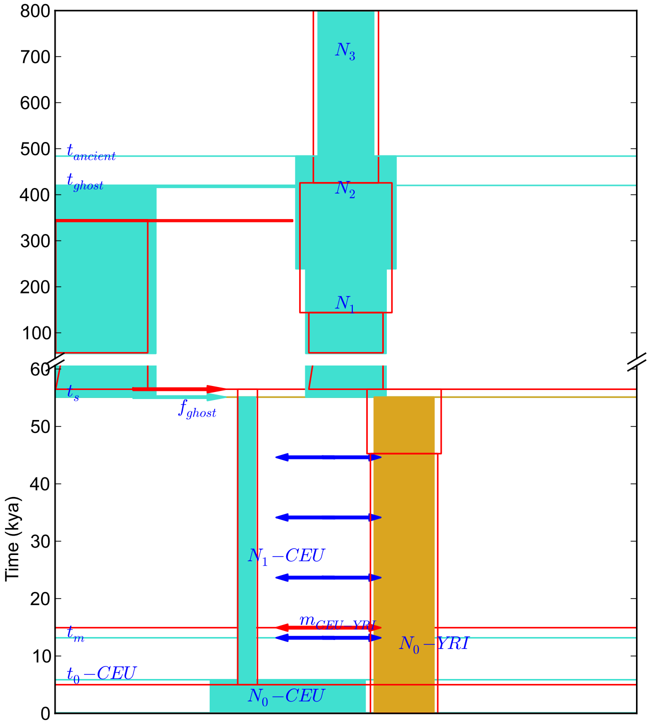


Figure A.5: The solid blue and yellow blocks in this figure represent the demographic history that was inferred from the 1000 genomes trio data. The overlaid red lines depict the mean history inferred from replicate MS simulations.

In Figure A.9, we compare composite likelihood surfaces generated by ∂a∂i from the joint allele frequency spectrum and by our method from the joint IBS tract spectrum. Each point in the ∂a∂i likelihood surface is generated by fixing τ_a and τ_s and then optimizing f (∂a∂i deterministically optimizes the size N). Similarly, each point in the IBS tract likelihood surface is generated by fixing τ_a and τ_s and jointly optimizing f and N. These likelihood surfaces show that both methods allow for accurate grid search estimation of demographic parameters. However, the ∂a∂i numerical optimization routines usually fail to arrive at a good estimate starting from a random point in demographic parameter space. For the $\tau_a = 0.01$ history, the optimal parameters located by grid search have a Poisson log likelihood greater than $-5,000$, the best parameters obtained from 20 random Nelder-Mead optimizations have Poisson log likelihood $-8,329$. Nelder-Mead optimization was chosen for this comparison because it is the routine recommended in the ∂a∂i manual for optimization starting far from the true optimum. In contrast, if we sample Θ uniformly at random

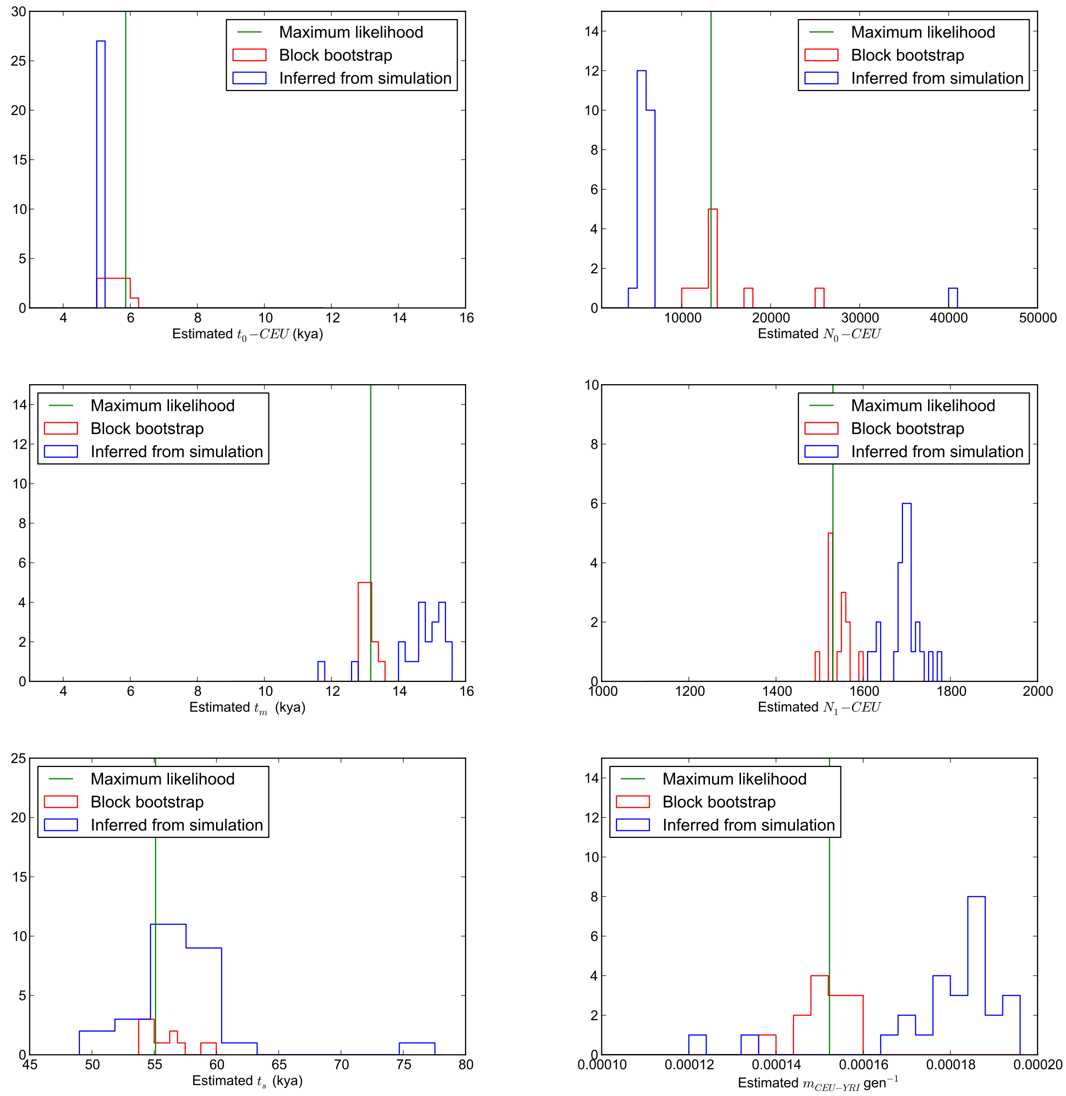


Figure A.6: **Results of block bootstrapping and parametric bootstrapping: Part I of III.** In each of these figures, the green line marks a parameter estimate obtained from the 1000 genomes trio data. Each data point contributing to the red "block bootstrap" histogram was estimated from a dataset that was created by sampling 100 bootstrap blocks from the trio data with replacement. The blue histogram records the results of inference from simulated data: each dataset was generated using the MS command line in section A.3.3 and the maximum likelihood parameter values shown in green.

from a bounded range and maximize the likelihood of observed IBS tracts, the optimization consistently terminates very close to the global maximum (see Tables A.1 and A.2. Both the SFS likelihood surface and the IBS tract likelihood surface allow for parameter estimation by

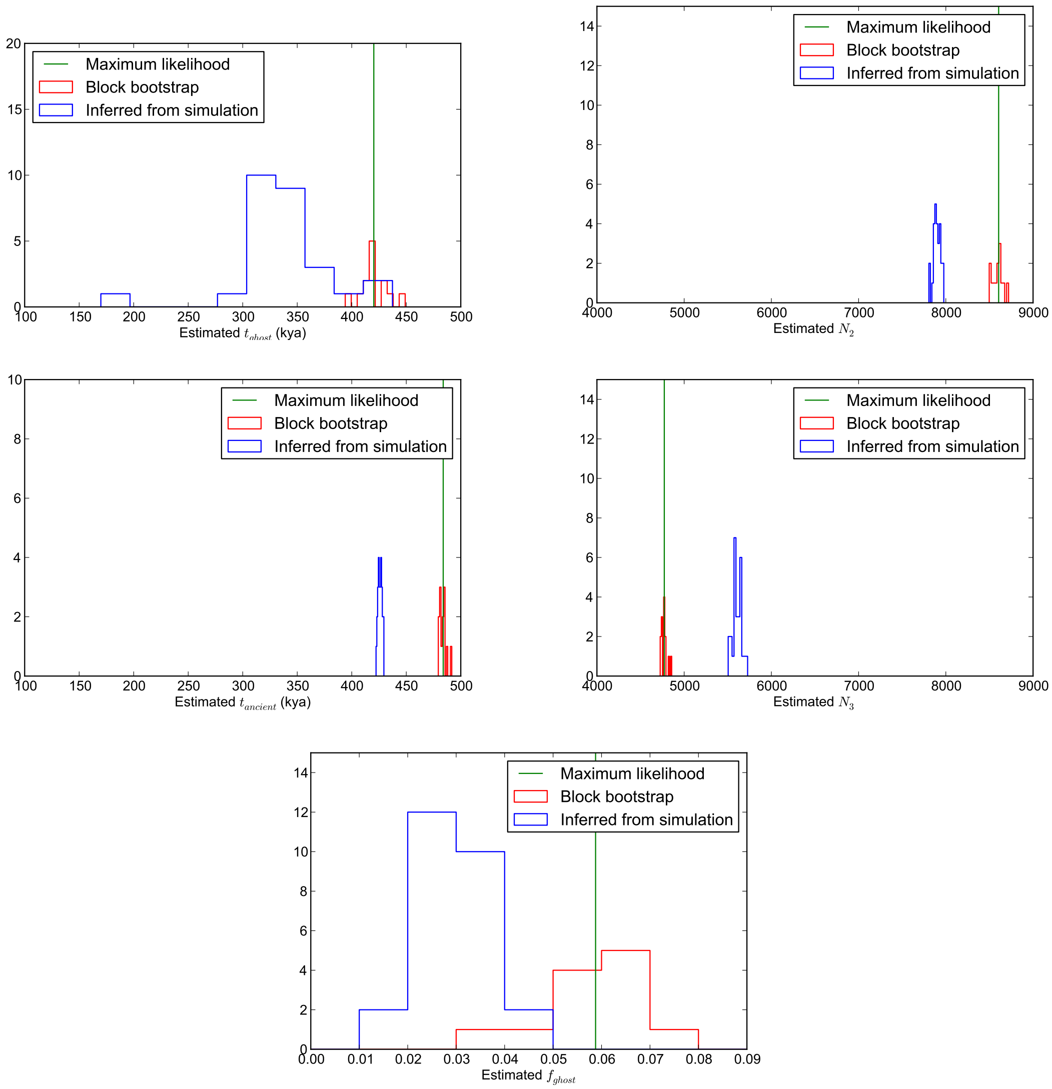


Figure A.7: **Results of block bootstrapping and parametric bootstrapping: Part II of III.**

grid search, but the two likelihood surfaces have different shapes that suggest complementary demographic sensitivities. The SFS likelihood is more sensitive to variation in divergence time than to changes in admixture time, while the IBS tract likelihood is more sensitive to variation in the time of last gene flow.

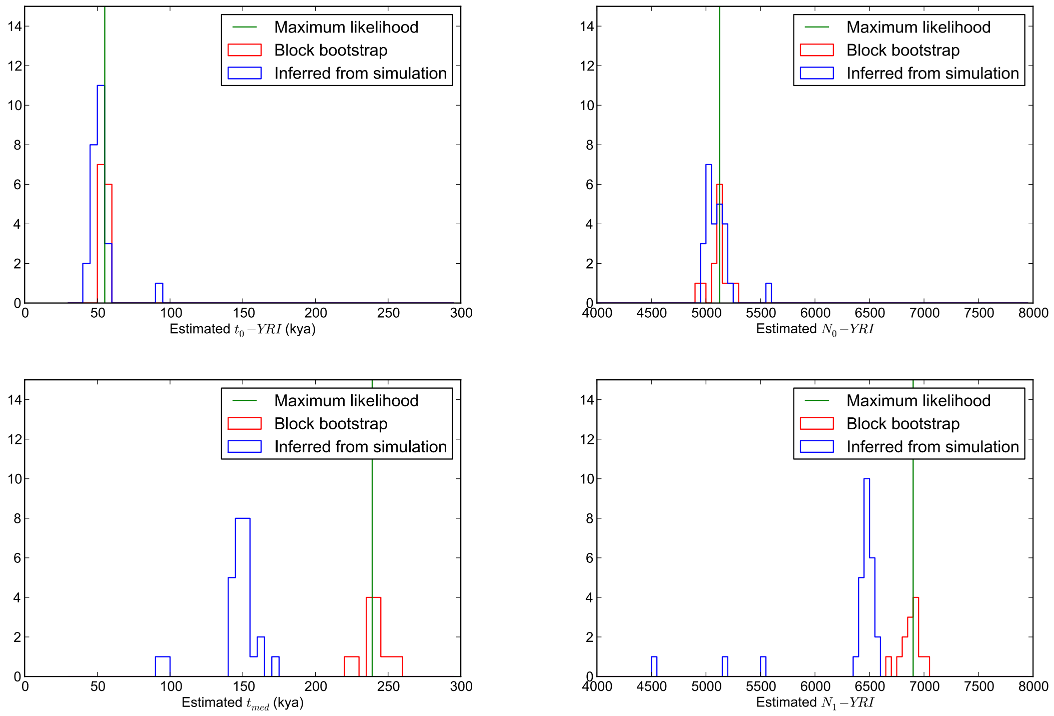


Figure A.8: **Results of block bootstrapping and parametric bootstrapping: Part III of III.**

A.4.2 The 1000 Genomes trios

Before inferring a new demographic history from the trio data, we examined IBS tracts simulated under histories published by Gutenkunst, *et al*, Gravel, *et al.* and Li and Durbin Gutenkunst et al. (2009); Gravel et al. (2011); Li & Durbin (2011). As seen in Figure A.10, the Gutenkunst history predicts too few shared IBS tracts in the tens of kilobases range shared within the CEU and between the CEU and YRI. The results are similar if we simulated under the Gravel, *et al.* history that was inferred from site frequencies in the 1000 Genomes low coverage pilot Gravel et al. (2011), and are commensurate with the fact that these models predict a weaker out of Africa bottleneck and shorter period of CEU-YRI gene flow than we do.

A.4.3 The NIEHS site frequency spectrum

Although the Gutenkunst, *et al.* and Gravel, *et al.* histories fit human site frequency spectra well, Figure A.10 illustrates that they do not predict the right spectrum of shared IBS tracts. Similarly, there is no guarantee that our inferred demographic history should predict the right CEU-YRI site frequency spectrum. To test this, we used MS to simulate

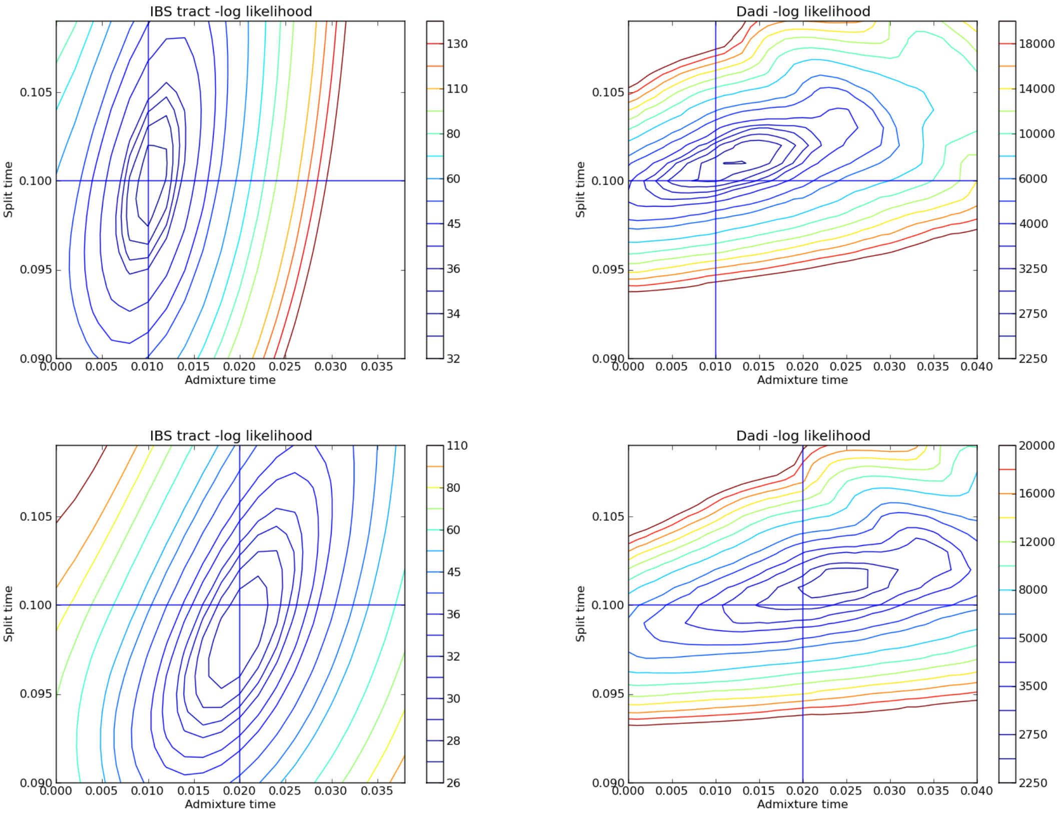


Figure A.9: The IBS tract likelihood surfaces were generated from two of the 200 simulated data sets that were analyzed to produce Table 2.1 in the main text, while the ∂a∂i log likelihood surfaces were generated from an equivalent amount of simulated allele frequency data. In each case, a grid search will accurately estimate the parameters of the simple admixture history in Figure 2.2 of the main text.

a 20-by-20 site frequency spectrum under our demographic model and compared it to the National Institute of Environmental Health Science (NIEHS) frequency spectrum that was analyzed in Gutenkunst et al. (2009) and generously made available by Ryan Gutenkunst. The simulated SFS had an FST of 0.205, significantly larger than the FST value of 0.158 that was computed from the NIEHS frequency spectrum data (see Figure A.11). We discuss possible reasons for the discrepancy in the main text.

-log likelihood	τ_a	τ_s	f
Simul. params:	0.01	0.1	0.05
8329.9620198	0.03659	0.1055	0.104
26328.1005948	0.06069	0.1120	0.232
26333.3741567	0.06071	0.1120	0.232
27515.9085976	0.05863	0.1106	0.208
40172.3596033	0.08467	0.1106	0.542
45041.8469532	0.08896	0.1291	0.856
45041.8469846	0.08896	0.1291	0.856
66521.0187249	0.09032	0.09040	0.333
66521.0187249	0.09032	0.09040	0.333
66521.0187253	0.09032	0.09040	0.333
66521.18999	0.09035	0.09043	0.668
66521.1899907	0.09035	0.09043	0.668
66521.1899912	0.09035	0.09043	0.668
66521.2117573	0.09032	0.09040	0.327
66841.6311539	0.09037	0.09037	0.849
66841.6916678	0.09037	0.09037	0.849
66847.2382421	0.09037	0.09037	0.896
66849.1579895	0.09037	0.09037	0.911
66851.6329195	0.09037	0.09037	0.0709
66855.8711617	0.09037	0.09037	0.959

Table A.1: This table contains the results of 20 ∂a∂i Nelder-Mead optimizations attempting to guess demographic parameters from an allele frequency spectrum. There is no population size estimate because ∂a∂i estimates it analytically before the optimization begins. See Table A.2 for a comparison to parameters estimated using IBS tracts.

-log likelihood	τ_a	$\tau_s - \tau_a$	f	$N \times 10^{-4}$
Simul. params:	0.01	0.09	0.05	1
44.69326588422074	0.01083835	0.08870295	0.05276519	1.00355111
44.693265893441264	0.01083864	0.08870302	0.05276541	1.0035453
44.693265917129885	0.01083854	0.08870251	0.05276585	1.00355611
44.69326593481351	0.01083841	0.08870352	0.05276549	1.00354219
44.69326597041676	0.01083884	0.08870345	0.05276642	1.00353741
44.69326598971337	0.01083823	0.08870245	0.05276505	1.00356078
44.693266033799766	0.01083897	0.0887035	0.05276659	1.00353431
44.693266061466886	0.0108382	0.08870225	0.0527649	1.00356397
44.6932661730406	0.0108392	0.08870364	0.05276679	1.00352958
44.69326620042027	0.01083882	0.08870404	0.05276646	1.00352885
44.69326649533553	0.01083919	0.08870424	0.05276697	1.00352098
44.6932671113402	0.01083957	0.0887046	0.05276771	1.00350983
44.693268280577676	0.01083968	0.08870552	0.05276809	1.00349454
44.693268832064774	0.01083732	0.08870015	0.05276338	1.00360853
44.69326911467046	0.01083995	0.08870584	0.05276843	1.00348585
44.693269115375614	0.01083886	0.08870406	0.05276158	1.00352361
44.69327005738606	0.01083705	0.08869962	0.05276269	1.00361987
44.69327180464335	0.01083675	0.08869895	0.05276187	1.00363333
44.69327369365423	0.01084035	0.08870774	0.05276928	1.00345149
44.693274370560026	0.01083657	0.0886981	0.05276198	1.00364998

Table A.2: This table contains the result of 20 optimizations that use an equivalent amount of IBS tract data to the ∂a∂i optimizations in Table A.2 (one of the 100 replicates used to generate Table 2.1 of the main text). All optimizations start from random parameter guesses–initial τ_a $\tau_s - \tau_a$ values are chosen uniformly between 0 to 20,000 generations; f is chosen uniformly on $(0, 1)$; N is chosen uniformly between 100 and 100,000. Our numerical routine for finding the optimum of the IBS tract likelihood surface is generally more successful at finding the optimum than the analogous routines that are part of the ∂a∂i package.

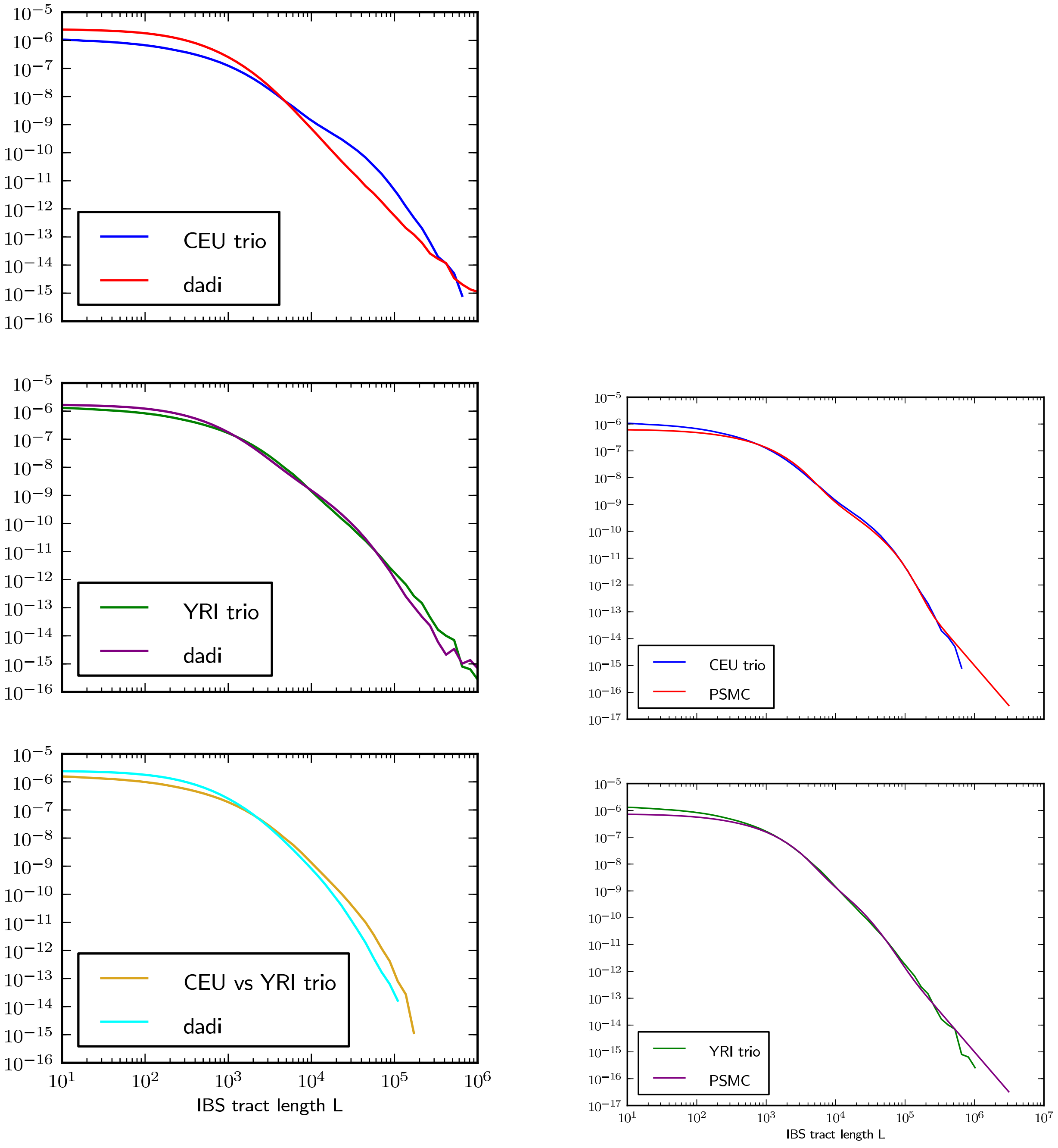


Figure A.10: IBS tract sharing in the 1000 Genomes trio parents vs. data simulated under the Gutenkunst, 2009 demographic model Gutenkunst et al. (2009).

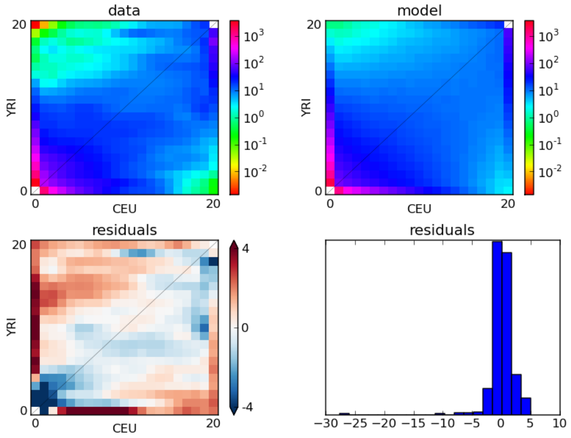


Figure A.11: **A SFS simulated under our inferred demographic history.** This model has an excess of high frequency derived alleles compared to the NIEHS data. The excess produces red off-diagonal regions in this Anscombe residual plot produced by ∂a∂i.

Appendix B

Supporting Information for Chapter 3: Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach

B.1 Formula derivations

B.1.1 HMM formulas:

The expression $R(i,t;j,t')$ in (3.8) is defined as

$$R(i,t;j,t') = \begin{cases} \left(R^{(i)}(t) + \sum_{k=0}^{i-1} R^{(k)}\right), & \text{if } i < j, \\ \left(R^{(j)}(t') + \sum_{k=0}^{j-1} R^{(k)}\right), & \text{if } i > j, \\ \left(R^{(i)}(t \wedge t') + \sum_{k=0}^{i-1} R^{(k)}\right), & \text{if } i = j, \end{cases}$$

where $\wedge$ denotes the min operator and, for $u \in [t_{k-1}, t_k)$,

$$\begin{aligned} R^{(k)}(u) &:= \frac{\rho_b \lambda_k}{n - \rho_b \lambda_k} \left(e^{-\rho_b u + n(u - t_{k-1})/\lambda_k} - 1\right) \prod_{m=1}^{k-1} e^{n(t_m - t_{m-1})/\lambda_m}, \\ R^{(k)} &:= \frac{\rho_b \lambda_k}{n - \rho_b \lambda_k} \left(e^{-\rho_b t_k + n(t_k - t_{k-1})/\lambda_k} - 1\right) \prod_{m=1}^{k-1} e^{n(t_m - t_{m-1})/\lambda_m}. \end{aligned}$$

After the state space has been discretized, we compute the transition probabilities using $y^{(i)}$ (the probability no recombination occurs), and $z^{(i,j)}$ (the probability recombination does occur):

$$\begin{aligned} y^{(i)} &= \frac{1}{\hat{\zeta}^{(\lambda)}(D_i,h)}\int_{t_{i-1}}^{t_i}\zeta^{(\lambda)}(t,h)e^{-\rho_b t}dt \\ &= \frac{1}{w^{(i)}}\frac{n}{n+\rho_b\lambda_i}\prod_{k=1}^{i-1}e^{-n(t_k-t_{k-1})/\lambda_k}\left(e^{-\rho_b t_{i-1}}-e^{-\rho_b t_i - n(t_i-t_{i-1})/\lambda_i}\right) \text{ and} \end{aligned}$$

$$\begin{aligned} z^{(i,j)} &= \frac{n}{w^{(i)}n_{h_{\ell-1}}}\int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i}\int_0^{t_{\ell-1}\wedge t_\ell}\rho_b e^{-\rho_b t_r}\frac{\zeta^{(\lambda)}(t_\ell,h_\ell)}{\int_{t_r}^\infty \zeta^{(\lambda)}(\tau)d\tau}\zeta^{(\lambda)}(t_{\ell-1},h_{\ell-1})dt_r dt_{\ell-1}dt_\ell \\ &:= Z^{(i,j)}+w^{(j)}\sum_{k=1}^{i\wedge j-1}R^{(k)}, \end{aligned}$$

where $Z^{(i,j)}$ corresponds to the case when the recombination event occurs during the time interval $D_{i\wedge j}$ (i.e. the latest it could) and $R^{(k)}$ corresponds to a recombination event in the time interval D_k. $R^{(k)}$ is defined as before, and $Z^{(i,j)}$ is

$$Z^{(i,j)} = \frac{n}{w^{(i)}n_{h_{\ell-1}}}\int_{t_{j-1}}^{t_j}\int_{t_{i-1}}^{t_i}\int_{t_{(i\wedge j)-1}}^{t_{\ell-1}\wedge t_\ell}\rho_b e^{-\rho_b t_r}\frac{\zeta^{(\lambda)}(t_\ell,h_\ell)}{\int_{t_r}^\infty \zeta^{(\lambda)}(\tau)d\tau}\zeta^{(\lambda)}(t_{\ell-1},h_{\ell-1})dt_r dt_{\ell-1}dt_\ell.$$

To evaluate $Z^{(i,j)}$, we must separate the computation into the cases $i<j$, $i>j$, and $i=j$,

$$Z^{(i,j)} = \begin{cases} \dfrac{w^{(j)}}{w^{(i)}}f^{(i)}, & \text{if } i<j \\ f^{(j)}, & \text{if } i>j \\ \frac{1}{w^{(i)}}\left(\frac{\rho_b\lambda_i}{n+\rho_b\lambda_i}e^{-\rho_b t_{i-1}} - 2e^{-n(t_i-t_{i-1})/\lambda_i-\rho_b t_{i-1}} - \frac{\rho_b\lambda_i}{n-\lambda_i\rho}e^{-\rho_b t_{i-1}-2n(t_i-t_{i-1})/\lambda_i} \right. \\ \left. +\frac{2n^2}{(n-\lambda_i\rho)(n+\lambda_i\rho)}e^{-\rho_b t_i-n(t_i-t_{i-1})/\lambda_i)}\right)\prod_{m=1}^{i-1}e^{-n(t_m-t_{m-1})/\lambda_m}, & \text{if } i=j, \end{cases}$$

where we define

$$f^{(i)} := e^{-\rho_b t_{i-1}} + \frac{\lambda_i\rho_b}{n-\lambda_i\rho_b}e^{-n(t_i-t_{i-1})/\lambda_i-\rho_b t_{i-1}} - \frac{n}{n-\lambda_i\rho_b}e^{-\rho_b t_i}.$$

To compute the emission probabilities we define $v^{(i)}(k)$ below:

$$v^{(i)}(k) := \frac{n(\theta_\ell)^k}{\lambda_i w^{(i)}k!}e^{nt_{i-1}/\lambda_i}\prod_{j=1}^{i-1}e^{-n(t_j-t_{j-1})/\lambda_j}\sum_{j=0}^k c_i^{-(j+1)}\frac{k!}{(k-j)!}\left[e^{-c_i t_{i-1}}t_{i-1}^{k-j}-e^{-c_i t_i}t_i^{k-j}\right],$$

where

$$c_i := \theta_\ell + \frac{n}{\lambda_i}.$$

B.1.2 Computation of the expected transition counts during the E-step:

Naively, if we compute the expected number of transitions from state $s_{\ell-1} = (D_i, h_{\ell-1})$ to state $s_\ell = (D_j, h_\ell)$, then marginalize over the haplotypes, we obtain an $O(n^2)$ algorithm. To improve the runtime, we can decompose the probability a transition is used between locus $\ell - 1$ and ℓ into a part that depends on the absorption haplotype and a part that depends on the absorption time interval, and thus we can reduce the run time to $O(n)$. First we compute the posterior probability $A^{(\ell)}(s_{\ell-1}, s_\ell)$ that a particular transition is used between locus $\ell - 1$ to ℓ, in terms of the discretized forward and backward probabilities $F_\ell(\cdot)$ and $B_\ell(\cdot)$. Let the newly sampled haplotype have allele a at locus ℓ, so $\alpha[\ell] = a$. Then

$$A^{(\ell)}(s_{\ell-1}, s_\ell) = \frac{1}{\hat{\pi}(\alpha)} \cdot F_\ell(s_{\ell-1}) \cdot \hat{\phi}^{(\lambda)}(s_\ell|s_{\ell-1}) \cdot \hat{\xi}^{(\lambda)}(a|s_\ell) \cdot B_\ell(s_\ell).$$

Now we marginalize over the haplotypes, plugging in the transition density formula

$$\begin{aligned}
\sum_{h_{\ell-1}} \sum_{h_\ell} A^{(\ell)}(s_{\ell-1}, s_\ell) &= \frac{1}{\hat{\pi}(\alpha)} \sum_{h_{\ell-1}} \sum_{h_\ell} F_\ell(s_{\ell-1}) \cdot \hat{\phi}^{(\lambda)}(s_\ell|s_{\ell-1}) \cdot \hat{\xi}^{(\lambda)}(a|s_\ell) \cdot B_\ell(s_\ell) \\
A^{(\ell)}(D_i, D_j) &= \frac{1}{\hat{\pi}(\alpha)} \sum_{h_{\ell-1}} \sum_{h_\ell} F_\ell(s_{\ell-1}) \cdot \hat{\xi}^{(\lambda)}(a|s_\ell) \cdot B_\ell(s_\ell) \left(y^{(i)} \delta_{s_{\ell-1}, s_\ell} + z^{(i,j)} \frac{n_{h_\ell}}{n} \right) \\
&= \frac{1}{\hat{\pi}(\alpha)} \left[\delta_{i,j} y^{(i)} \left(\sum_h F_\ell(D_i, h) \hat{\xi}^{(\lambda)}(a|D_i, h) B_\ell(D_i, h) \right) \right. \\
&\quad \left. + z^{(i,j)} \left(\sum_{h_{\ell-1}} F_\ell(s_{\ell-1}) \right) \left(\sum_{h_\ell} \frac{n_{h_\ell}}{n} \hat{\xi}^{(\lambda)}(a|s_\ell) B_\ell(s_\ell) \right) \right]
\end{aligned}$$

which is linear in n since we are only ever summing over one haplotype. To get the expected transition counts, we then sum over all the loci, so $A_{ij} = \sum_\ell A^{(\ell)}(D_i, D_j)$.

B.2 Simulation details

The following `ms` commands were used to simulate data under three population size change histories.

S1: `ms 10 1 -T -r 10000 1000000 -eN 0.05 0.1 -eN 0.2 0.5 -eN 0.5 1.25`
S2: `ms 10 1 -T -r 10000 1000000 -eN 0 10 -eN 0.05 0.1 -eN 0.2 0.5 -eN 0.5 1.25`
S3: `ms 10 1 -T -r 10000 1000000 -eN 0 0.75`

Note that `ms` times are in units of $4N_0$ generations, so we multiplied the raw times above by 2 to compare to PSMC and our SMCSD. Mutation rates were not specified above, since the only ms output used was tree at each base (-T flag). Mutations were then added to the trees using a finite sites model, the mutation matrix in Table B.1, and a mutation rate $\theta = 0.01 \times 1.443$. The factor of 1.443 accounts for the fact that this mutation matrix allows mutations that do not actually change the base (i.e., an A → A transition); see Chan et al. (2012) for further explanation. This mutation matrix was also used for the real data analysis.

The following style of command was used to run PSMC. We used 20 iterations as described in the PSMC paper, and the same pattern of parameters we used for our SMCSD.

```
psmc -p 3+2+2+2+2+2+3 -t 7 -N 20 -r 1 -o output.psmc input.psmcfa
```

To run our method, the following style of command was used.

```
java -Xmx25G -d64 SMCSD_EM -i input.fasta
-p params.txt -n 9 -t 5 -a "3 2 2 2 2 2 3"
```

The parameter file includes the number of loci in each sequence, the number of alleles (4 in our case), an estimate of the mutation rate, mutation matrix, and recombination rate, and the discretization. The -n flag specifies the number of haplotypes to use in the trunk, so there are $n + 1$ total. The -t flag specifies the number of threads to use; memory requirements scale linearly with this parameter. If -t 1 was specified in the case, then -Xmx5G could be used for the memory requirement. The -a flag specifies the pattern of parameters, in an analogous fashion to PSMC.

Table B.1: Mutation matrix for realistic human data. The rows represent the original base, and the columns represent the mutated base.

base	A	C	G	T
A	0.503	0.082	0.315	0.100
C	0.186	0.002	0.158	0.655
G	0.654	0.158	0	0.189
T	0.097	0.303	0.085	0.515

Appendix C

Supporting Information for Chapter 4: Decoding coalescent hidden Markov models in linear time

C.1 Explicit Computation of Transition Probabilities

In Equations 2 and 12 from the main text, we decompose the transition probabilities $\phi(s \mid s')$ and the stationary probability $\zeta(s)$ into the component terms $\mathbb{P}(C_i \mid C_{>i-1})$, $\mathbb{P}(C_{>i} \mid C_{>i-1})$, $\mathbb{P}(R_i, C_i \mid T_\ell = i)$, $\mathbb{P}(R_i, C_i \mid T_\ell > i)$, $\mathbb{P}(R_i, C_{>i} \mid T_\ell = i)$, $\mathbb{P}(R_i, C_{>i} \mid T_\ell > i)$, and $\mathbb{P}(\overline{R} \mid T_\ell = i)$. Here we give explicit formulae for these transition probabilities in terms of scaling factors $\lambda_1, \ldots, \lambda_d$ that specify the relative effective population sizes within each time interval. These formulae are specific to the method diCal with variable population size but no population structure. Very similar computations could be used for diCal with population structure, as well as for PSMC, CoalHMM, and related methods.

In addition to $\lambda_1, \ldots, \lambda_d$, these formulae will include the recombination rate ρ, scaled with respect to an implicit population size N_0 such that $\lambda_i \cdot N_0$ is the effective population size in interval i and $\rho = 4N_0 r$, where r is the recombination rate per site per generation. Time intervals are defined with respect to a fixed sequence of time points $t_0 = 0 < t_1 < \cdots < t_d = \infty$, where the ith time state is the interval between t_{i-1} and t_i. In addition, $\overline{n}_i$ denotes the average number of lineages that are present at time t_{i-1} in an n-leaf coalescent tree, and is computed in Sheehan et al. (2013).

We compute the components of the stationary and transition probabilities as follows:

$$\begin{aligned}
\mathbb{P}(C_{>i} \mid C_{>i-1}) &= e^{-(t_i - t_{i-1})\overline{n}_i/\lambda_i} \\
\mathbb{P}(C_i \mid C_{>i-1}) &= 1 - e^{-(t_{i-1} - t_i)\overline{n}_i/\lambda_i} \\
\mathbb{P}(\overline{R} \mid T_\ell = i) &= \frac{1}{1 - e^{-\overline{n}_i(t_i - t_{i-1})/\lambda_i}} \int_{t=t_{i-1}}^{t_i} \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_i(t - t_{i-1})/\lambda_i - t\rho} dt \\
&= \frac{\overline{n}_i}{\overline{n}_i + \lambda_i \rho} \frac{e^{-t_{i-1}\rho} - e^{-t_i\rho - \overline{n}_i(t_i - t_{i-1})/\lambda_i}}{1 - e^{-\overline{n}_i(t_i - t_{i-1})/\lambda_i}}
\end{aligned}$$

$$
\begin{aligned}
\mathbb{P}(R_i, C_i \mid T_\ell = i) &= \frac{1}{\int_{t_{i-1}}^{t_i} \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_i (t - t_{i-1})/\lambda_i} dt} \int_{t_0=t_{i-1}}^{t_i} \int_{t=t_{i-1}}^{t_i} \int_{t_r=t_{i-1}}^{t_0 \wedge t} \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_\ell (t_0 - t_{i-1})/\lambda_i} \\
&\quad \cdot \rho e^{-t_r \rho} \cdot \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_i (t - t_r)/\lambda_i} dt_r dt dt_0 \\
&= \frac{1}{1 - e^{-\overline{n}_i (t_i - t_{i-1})/\lambda_i}} \left(\frac{\rho \lambda_i}{\overline{n}_i + \rho \lambda_i} e^{-\rho t_{i-1}} - 2 e^{-\overline{n}_i (t_i - t_{i-1})/\lambda_i - \rho t_{i-1}} \right. \\
&\quad \left. - \frac{\rho \lambda_i}{\overline{n}_i - \lambda_i \rho} e^{-\rho t_{i-1} - 2\overline{n}_i (t_i - t_{i-1})/\lambda_i} + \frac{2\overline{n}_i^2}{(\overline{n} - \lambda_i \rho)(\overline{n}_i + \lambda_i \rho)} e^{-\rho t_i - \overline{n}_i (t_i - t_{i-1})/\lambda_i} \right) \\
\mathbb{P}(R_i, C_i \mid T_\ell > i) &= \int_{t=t_{i-1}}^{t_i} \int_{t_r=t_{i-1}}^{t} \rho e^{-\rho t_r} \cdot \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_i (t - t_r)/\lambda_i} dt_r dt \\
&= e^{-t_{i-1}\rho} + \frac{\lambda_i \rho}{\overline{n}_i - \lambda_i \rho} e^{-\overline{n}_i (t_i - t_{i-1})/\lambda_i - t_{i-1}\rho} - \frac{\overline{n}_i}{\overline{n}_i - \lambda_i \rho} e^{-t_i \rho} \\
\mathbb{P}(R_i, C_{>i} \mid T_\ell = i) &= \frac{1}{\int_{t_{i-1}}^{t_i} \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_i (t - t_{i-1})/\lambda_i} dt} \int_{t_r=t_{i-1}}^{t_i} \int_{t_0=t_r}^{t_i} \frac{\overline{n}_i}{\lambda_i} e^{-\overline{n}_i (t_0 - t_{i-1})/\lambda_i} \\
&\quad \cdot \rho e^{-t_r \rho} \cdot e^{-\overline{n}_i (t_i - t_r)/\lambda_i} dt_0 dt_r \\
&= \frac{1}{1 - e^{-\overline{n}_i (t_i - t_{i-1})/\lambda_i}} \cdot e^{-\overline{n}_i (t_i - t_{i-1})/\lambda_i} \left(e^{-t_{i-1}\rho} - \frac{\overline{n}_i}{\overline{n}_i - \lambda_i \rho} e^{-t_i \rho} \right. \\
&\quad \left. + \frac{\rho \lambda_i}{\overline{n}_i - \rho \lambda_i} e^{-\rho t_{i-1} - \overline{n}_i (t_i - t_{i-1})/\lambda_i} \right)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}(R_i, C_{>i} \mid T_\ell > i) &= \mathbb{P}(R_i \mid T_\ell > i) - \mathbb{P}(R_i, C_i \mid T_\ell > i) \\
&= (1 - e^{-(t_i - t_{i-1})\rho}) - (e^{-t_{i-1}\rho} + \frac{\lambda_i \rho}{\overline{n}_i - \lambda_i \rho} e^{-\overline{n}_i (t_i - t_{i-1})/\lambda_i - t_{i-1}\rho} - \frac{\overline{n}_i}{\overline{n}_i - \lambda_i \rho} e^{-t_i \rho}).
\end{aligned}
$$

C.2 Posterior Expectations for the Augmented HMM

In this section of the appendix, we discuss how to compute the posterior expectations in (4.15). We express these posterior expectations in terms of usual forward and backward probabilities $f(x_{1:\ell}, S_\ell)$ and $b(x_{\ell+1:L} \mid S_\ell)$, and also the combined forward probabilities $f\,(x_{1:\ell}, T_\ell = i)$, $f\,(x_{1:\ell}, T_\ell > i)$, and $f\,(x_{1:\ell}, R_{\leq i}, C_{>i})$ introduced in Section 2.1 of the main text. In addition, we need to define the combined backward probabilities,

$$
\begin{aligned}
b(x_{\ell+1:L} \mid T_{\ell+1} = i) = \frac{\mathbb{P}(x_{\ell+1:L}, T_{\ell+1} = i)}{\mathbb{P}(T_{\ell+1} = i)} \\
= \frac{f(x^{(r)}_{1:L-\ell}, T^{(r)}_{L-\ell} = i)}{\mathbb{P}(T = i)},
\end{aligned}
$$

$$\begin{aligned} b(x_{\ell+1:L} \mid T_{\ell+1} > i) &= \frac{\mathbb{P}(x_{\ell+1:L}, T_{\ell+1} > i)}{\mathbb{P}(T_{\ell+1} > i)} \\ &= \frac{f(x^{(r)}_{1:L-\ell}, T^{(r)}_{L-\ell} > i)}{\mathbb{P}(T > i)} \\ &= \frac{f(x^{(r)}_{1:L-\ell}, T^{(r)}_{L-\ell} > i)}{\mathbb{P}(C_{>1})\mathbb{P}(C_{>2} \mid C_{>1}) \cdots \mathbb{P}(C_{>i} \mid C_{>i-1})}. \end{aligned}$$

We start by showing how to express $\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > i, T_{\ell+1} > i\} \mid x_{1:L}\Big]$ in terms of these forward and backward probabilities:

$$\begin{aligned} &\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > i, T_{\ell+1} > i\} \mid x_{1:L}\Big] \\ &= \sum_{\ell=1}^{L-1} \mathbb{P}(\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > i, T_{\ell+1} > i \mid x_{1:L}) \\ &= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) \\ &\qquad \times \sum_{j>i} \left(\prod_{k=i+1}^{j-1} \mathbb{P}(C_{>k} \mid C_{>k-1}) \right) \mathbb{P}(C_j \mid C_{>j-1}) \mathbb{P}(x_{\ell+1:L} \mid T_{\ell+1} = j) \\ &= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) \sum_{j>i} \frac{\mathbb{P}(x_{\ell+1:L}, T_{\ell+1} = j)}{\mathbb{P}(C_{>1})\mathbb{P}(C_{>2} \mid C_{>1}) \cdots \mathbb{P}(C_{>i} \mid C_{>i-1})} \\ &= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f\left(x_{1:\ell}, T_\ell > i\right) \mathbb{P}(R_i, C_{>i} \mid T_\ell > i) b(x_{\ell+1:L} \mid T_{\ell+1} > i), \end{aligned}$$

where $\pi(x) := \mathbb{P}(x) = \sum_s f(x_{1:L}, s)$.

Computing the other posterior expectations is similarly straightforward. We list the derived expressions for them here:

$$\begin{aligned} &\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = \overline{R}, T_\ell = i\} \mid x_{1:L}\Big] \\ &\qquad = \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} \sum_{h \in \mathcal{H}} f(x_{1:\ell}, (h,i)) \mathbb{P}(\overline{R} \mid T_\ell = i) \xi(x_{\ell+1} \mid h_{\ell+1}, i) b(x_{\ell+2:L} \mid (h,i)), \\ &\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell = T_{\ell+1} = i\} \mid x_{1:L}\Big] \\ &\qquad = \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f(x_{1:\ell}, T_\ell = i) \mathbb{P}(R_i, C_i \mid T_\ell = i) b(x_{\ell+1:L} \mid T_{\ell+1} = i), \\ &\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > T_{\ell+1} = i\} \mid x_{1:L}\Big] \end{aligned}$$

$$= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f(x_{1:\ell}, T_\ell > i)\mathbb{P}(R_i, C_i \mid T_\ell > i) b(x_{\ell+1:L} \mid T_{\ell+1} = i),$$

$$\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_{\ell+1} > T_\ell = i\} \mid x_{1:L}\Big]$$

$$= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f(x_{1:\ell}, T_\ell = i)\mathbb{P}(R_i, C_{>i} \mid T_\ell = i) b(x_{\ell+1:L} \mid T_{\ell+1} > i),$$

$$\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} = R_i, T_\ell > i, T_{\ell+1} > i\} \mid x_{1:L}\Big]$$

$$= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f(x_{1:\ell}, T_\ell > i)\mathbb{P}(R_i, C_{>i} \mid T_\ell > i) b(x_{\ell+1:L} \mid T_{\ell+1} > i),$$

$$\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} \in R_{<i}, T_{\ell+1} > i\} \mid x_{1:L}\Big]$$

$$= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f\left(x_{1:\ell}, R_{\leq i-1}, C_{>i-1}\right) \mathbb{P}(C_{>i} \mid C_{>i-1}) b(x_{\ell+1:L} \mid T_{\ell+1} > i),$$

$$\mathbb{E}\Big[\#\ell : \{\mathcal{R}_{\ell,\ell+1} \in R_{<i}, T_{\ell+1} = i\} \mid x_{1:L}\Big]$$

$$= \frac{1}{\pi(x)} \sum_{\ell=1}^{L-1} f\left(x_{1:\ell}, R_{\leq i-1}, C_{>i-1}\right) \mathbb{P}(C_i \mid C_{>i-1}) b(x_{\ell+1:L} \mid T_{\ell+1} = i),$$

$$\mathbb{P}(T_1 = i \mid x_{1:L})$$

$$= \frac{1}{\pi(x)} \sum_{h \in \mathcal{H}} f(x_1, (h, i)) b(x_{2:L} \mid (h, i)),$$

$$\mathbb{P}(T_1 > i \mid x_{1:L})$$

$$= \frac{1}{\pi(x)} \sum_{h \in \mathcal{H}} \sum_{j=i+1}^{d} f(x_1, (h, j)) b(x_{2:L} \mid (h, j))$$

$$= \mathbb{P}(T_1 > i + 1 \mid x_{1:L}) + \mathbb{P}(T_1 = i + 1 \mid x_{1:L}).$$

C.3 Running diCal on Simulated and Real Data

To run diCal on the simulated data, we used the commandlines below ($d = 9$ and $d = 21$):

```
java -jar diCal.jar -F data.fasta -I params.txt
-n 10 -p "3+3+3" -t 1.0 -u 0
java -jar diCal.jar -F data.fasta -I params.txt
-n 10 -p "8+2+2+2+2+2+3" -t 2.0 -u 1
```

where n is the total number of haplotypes, p is the grouping of discretization intervals into parameters, t is the start point of the last discretization interval (in coalescent units) and u

is a flag for using the linear vs. quadratic method.

To run diCal on the 1000 Genomes data, we used the commandlines:

```
java -jar diCal.jar -F data.fasta -I params.txt -n 10 -p "3+3+3" -t 1.0 -u 0
java -jar diCal.jar -F data.fasta -I params.txt -n 10 -p "5+3+2+2+2+3+4" -t 2.0 -u 1
```

For both the simulated and real data, the parameter groupings were chosen such that the number of parameters inferred would be $d/3$, with minimal runaway behavior.

Appendix D

Supplemental Information for Chapter 5: Error-prone polymerase activity causes multinucleotide mutations in humans

D.1 Exponential decay of LD over short genomic distances

In data simulated under the standard coalescent with recombination using `ms` (Hudson 2002), we saw that the count $N_{\mathrm{LD}}(L)$ of SNPs in perfect LD L bp apart decays approximately exponentially for L between 1 and 100 bp. Here, we give a heuristic argument why this should be true in the asymptotic limit $L \ll 1/\rho$, where ρ is the population-scaled recombination rate.

Let $T_1, \ldots, T_L$ be the sequence of n-leaf coalescence trees that occur at the sites of a sequence of length L that has been evolving with mutation and recombination parameters θ and ρ. For simplicity, we assume a constant effective population size N. The rates θ and ρ are population-scaled such that $\mu = \theta/(4N)$ is the mutation rate per site per generation and $r = \rho/(4N)$ is the recombination rate per site per generation. Given any of these trees T_i, let $P(T_i)$ be the set of points $(x, y) \in T_i$ with the property that x and y lie on the same branch of T_i. The sequential coalescent yields a natural map from points on T_i to points on T_{i+1}, though not every point on T_i necessarily maps to a point on T_{i+1} if a recombination has occurred between the sites. Let $\epsilon_{x,y}(T_i)$ be defined such that $1 - \epsilon_{x,y}(T_i)$ is the probability that x and y both map to T_{i+1}, $(x, y) \in P(T_{i+1})$, and the branch containing (x, y) subtends the same set of lineages in both T_i and T_{i+1}.

A pair of points $(x, y) \in P(T_i)$ can give rise to a pair of SNPs in perfect LD at sites i and j if the following events occur: E1) a mutation occurs at position x on tree T_i, E2) x and y map to a single branch of each tree between T_i and T_j that subtends the same set of lineages, and E3) A mutation occurs at position y on tree T_j. Not every pair of SNPs in perfect LD must correspond to a pair of points x, y satisfying E1–E3; for example, the integrity of the clade by the branch containing x and y could be broken up and re-formed by two separate recombinations occurring between sites i and j. If the sample size n is relatively large, however, it will be combinatorially unlikely for any clade to re-form after it has been broken up by recombination, particularly within a very short genomic window. Motivated by this, we will estimate $N_{\mathrm{LD}}(L)$ assuming that all linked SNP pairs arise at pairs of points (x, y) that satisfy E1–E3 for some T_i and T_j.

Integrating over x, y and $T_i, \ldots, T_{i+L}$, we compute that the probability of observing a pair of SNPs in perfect LD at sites i and $i + L$ is the following:

$$\begin{aligned} N_{\mathrm{LD}}(i, i+L) &= \theta^2 \int_{T_i,\ldots,T_{i+L}} \int_{(x,y)\in P(T_i)} (1 - \epsilon_{x,y}(T_i)) \cdots (1 - \epsilon_{x,y}(T_{i+L})) d_{(x,y)} d_{(T_i,\ldots,T_{i+L})} \\ &= \theta^2 + \theta^2 \sum_{k=1}^{L} (-1)^k \int_{T_i,\ldots,T_{i+L}} \int_{(x,y)\in P(T_i)} \sum_{i \le j_1 < \cdots < j_k \le i+L} \epsilon_{x,y}(T_{j_1}) \cdots \epsilon_{x,y}(T_{j_k}) d_{(x,y)} d_{(T_i,\ldots,T_{i+L})}. \end{aligned}$$

Let $\ell(T)$ denote the total branch length of tree T. Since any alteration of tree structure requires a recombination event, $\epsilon_{x,y}(T) \le \rho \cdot \ell(T)$. This implies that

$$\sum_{i \le j_1 < \cdots < j_k \le i+L} \epsilon_{x,y}(T_{j_1}) \cdots \epsilon_{x,y}(T_{j_k}) \le (\epsilon_{x,y}(T_i) + \cdots + \epsilon_{x,y}(T_{i+L}))^k \le (L\rho(\ell(T_i) + \cdots + \ell(T_{i+L})))^k$$

for every k. Letting $T^{(2)}$ denote the sum of squares of the branch lengths of a coalescent tree T, this implies that

$$\begin{aligned} N_{\mathrm{LD}}(i, i+L) &= \theta^2 - \theta^2 \int_{T_i,\ldots,T_{i+L}} \int_{(x,y)\in P(T_i)} \rho(\epsilon_{x,y}(T_i) + \cdots + \epsilon_{x,y}(T_{i+L})) d_{(x,y)} d_{T_i,\ldots,T_{i+L}} + O((\rho L)^2) \\ &= \theta^2 \mathbb{E}(T^{(2)})(1 - \rho L \cdot \mathbb{E}(\epsilon_{x,y}(T_i))) + O((\rho L)^2). \end{aligned}$$

In human-like data where $N = 10,000$ and $\rho = 0.0004$, we can see that $\rho L \le 0.04 \ll 1$ when $L < 100$. Therefore, the first-order linear decay rate of $N_{\mathrm{LD}}(i, i+L)$ is small compared to L. In addition, we can see from equation (D.2) that the $O((\rho L)^2)$ term of the Taylor expansion will be positive, meaning that $N_{\mathrm{LD}}(i, i+L)$ has concave upward shape. This makes it reasonable, for our purposes, to approximate $N_{\mathrm{LD}}(i, i+L)$ by an exponential function.

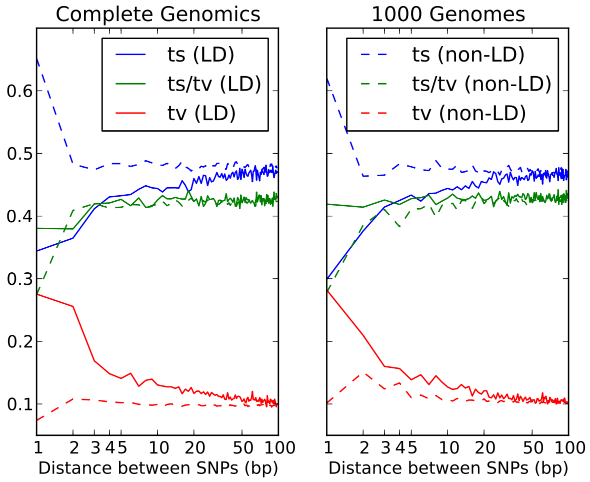


Figure D.1: **Consistency of the transition: transversion ratio across linked SNPs from different sequencing platforms.** This figure shows that excess transversions in perfect LD are not an artifact of Illumina sequencing or the 1000 Genomes pipeline, but are also present in a set of 54 human genomes sequenced by Complete Genomics (CG). To make this comparison, we subsampled 54 genomes from the 1000 Genomes Phase I dataset that had approximately the same population breakdown as the 54 CG individuals. Because the CG data are unphased, we ignore all 1000 Genomes phasing information, classifying each SNP pair as being in perfect LD if it is in perfect LD with respect to at least one possible haplotype phasing. We ignore all CG SNPs at which more than 10% of the samples have a missing genotype. Note that most pairs of nearby SNPs in the CG data are not annotated as "SNPs" in the MasterVarBeta files that are publicly availably, but as "complex" substitutions where a string of two or more bases is regarded as one polymorphic unit. We ignored all complex substitutions that included indels, but extracted SNPs from each substitution multi nucleotide substitution where all variant alleles had the same length and a one-to-one mapping between sites was possible.

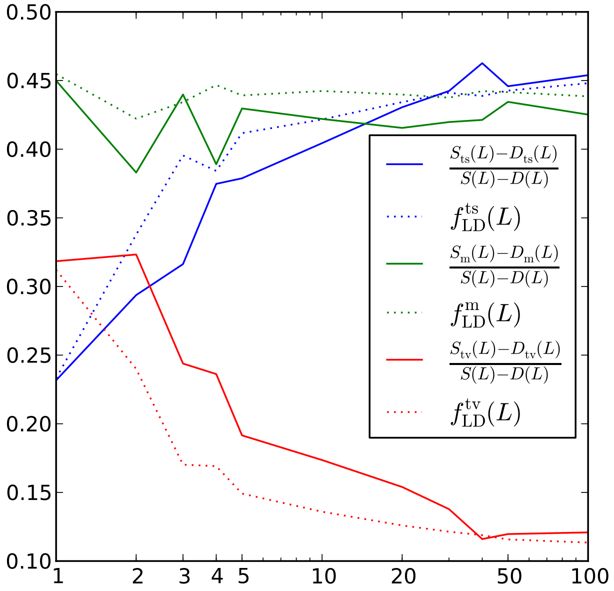


Figure D.2: **Quantifying simultaneous transitions vs. transversions.** This figure plots the relative abundances of transitions, transversions, and mixed pairs as fractions of the quantity $S(L) - D(L)$. The transversion fraction $(S_{\mathrm{tv}}(L) - D_{\mathrm{tv}}(L))/(S(L) - D(L))$ is slightly higher than $f_{\mathrm{LD}}^{\mathrm{tv}}(L)$, especially for small L where MNMs are the most apparent.

Table D.1: **Quantifying MNMs spanning $>$ 100 bp**

L	$S_{\text{ts}}^{(\text{LD})}(L)$	$m_{\text{ts}}(L)$	$S_{\text{m}}^{(\text{LD})}(L)$	$m_{\text{m}}(L)$	$S_{\text{tv}}^{(\text{LD})}(L)$	$m_{\text{tv}}(L)$
100	4.507487e+10	0.3704	4.414249e+10	0.3542	1.142233e+10	0.3891
200	4.217310e+10	0.3798	4.100119e+10	0.3723	1.054684e+10	0.3770
300	3.961781e+10	0.3954	3.880264e+10	0.3776	9.884130e+09	0.3716
400	3.807976e+10	0.3992	3.726721e+10	0.3950	9.351356e+09	0.3898
1000	2.094000e+10	0.4262	2.096024e+10	0.4237	5.275246e+09	0.3977
3000	1.287797e+10	0.4435	1.270760e+10	0.4183	3.234083e+09	0.3820
5000	9.147849e+09	0.4033	9.088510e+09	0.3921	2.377678e+09	0.3860
10000	5.147443e+09	0.2906	5.139715e+09	0.2801	1.330389e+09	0.2672

For each distance L listed above, we counted all SNPs in perfect LD between L and $L + 100$ bp apart. For each pair type t, we subsampled haplotype pairs in order to calculate $S_t(L), D_t(L), S_t^{(\text{LD})}(L)$, and $m_t(L)$ aggregated over this 100 bp window. The results suggest that the ratio of MNMs to perfect LD SNPs achieves its minimum value around $L = 100$ and then stops decreasing with the distance between SNPs.

D.2 Enrichment of MNMs in large datasets

As a consequence of the argument in Section D.1, we saw that the abundance of linked independent mutations in a sample of n lineages is proportional to the expected sum of squared branch lengths in an n-leaf coalescence tree. This is a simple consequence of the fact that two mutations must affect a single branch to create SNPs in perfect LD. In contrast, the abundance of MNMs should be proportional to the total tree length, just as the total number of segregating sites is proportional to the expected tree length.

It is a standard result in population genetics that the expected total tree length $\mathbb{E}(T_{\text{total}})$ equals the harmonic number $\sum_{i=1}^{n-1} 1/i$ (Watterson 1975). To show this, let T_i be the length of time that the a random genealogy has exactly i lineages, which has distribution function $f_i(t) = \binom{i}{2} \exp(-t\binom{i}{2})$. It follows that

$$\mathbb{E}(T_{\text{total}}) = \mathbb{E}\left(\sum_{i=2}^{n} iT_i\right) = \sum_{i=2}^{n} i\mathbb{E}(T_i) = \sum_{i=2}^{n} i \cdot \frac{2}{i(i-1)} = \sum_{i=1}^{n-1} \frac{1}{i} \approx \log(n-1)$$

Therefore, if μ_{MNM} is the rate of MNMs per coalescent time unit, the expected number of MNMs approaches infinity with increasing n at the asymptotic rate $\mu_{\text{MNM}} \log(n)$.

In contrast, if μ is the rate of ordinary point mutations, linked independent mutations appear at the rate $\mu^2\mathbb{E}(T_{\text{total}}^{(2)})$, where $T_{\text{total}}^{(2)}$ is the sum of squares of the coalescent tree branch lengths. We can show that $\mathbb{E}(T_{\text{total}}^{(2)})$ approaches a constant as $n \to \infty$. To proceed, we let $\ell_1, \dots, \ell_n$ denote the lengths of the n leaves of the tree and $b_{n-1}, \dots, b_2$ denote the lengths

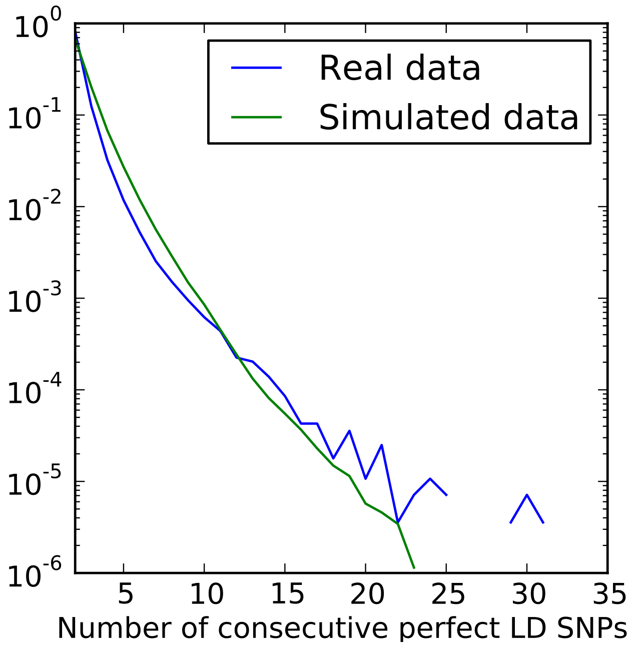


Figure D.3: **Clusters of 2 or more perfect LD SNPs.** In the 1000 Genomes data, we found all clusters of 2 or more perfect LD SNPs with fewer than 1 kb between adjacent pairs. We plot the resulting distribution of cluster sizes and compare it to the distribution of cluster sizes in data simulated under the Harris & Nielsen (2013) model using ms. The cluster sizes from real data are slightly more dispersed toward very small and very large clusters. It is possible that the longest clusters formed by error-prone replication of single-stranded DNA following double-strand breakage as proposed by Roberts et al. (2012).

of the $n-2$ internal branches, indexed such that the more recent endpoint of branch i is the first time when the tree has i lineages:

$$\mathbb{E}(T^{(2)}_{\text{total}}) = n\mathbb{E}(\ell_n^2) + \sum_{i=2}^{n-1} \mathbb{E}(b_i^2).$$

Given that a branch is present when the tree has i lineages, the probability that the branch is ended by the next coalescence event is $(i-1)/\binom{i}{2} = 2/i$. Therefore, given $j < i$, the probability that $b_i = T_i + \cdots + T_j$ is

$$\mathbb{P}(b_i = T_i + \cdots + T_j) = \left(1 - \frac{2}{i}\right) \cdots \left(1 - \frac{2}{j+1}\right) \cdot \frac{2}{j} = \frac{(i-2)\cdots(j-1)\cdot 2}{i \cdots (j+1)\cdot j} = \frac{2(j-1)}{i(i-1)}.$$

It follows that

$$\begin{aligned}
\mathbb{E}(b_i^2) &= \sum_{j=2}^{i} \mathbb{P}(b_i = T_i + \cdots + T_j) \cdot \mathbb{E}((T_i + \cdots + T_j)^2) \\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\sum_{k=j}^{i} \mathbb{E}(T_k^2) + 2 \sum_{j \le k < \ell \le i} \mathbb{E}(T_k)\mathbb{E}(T_\ell) \right) \\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\sum_{k=j}^{i} \frac{8}{k^2(k-1)^2} + \sum_{j \le k < \ell \le i} \frac{8}{k(k-1)\ell(\ell-1)} \right) \\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\sum_{k=j}^{i} \frac{4}{k^2(k-1)^2} + \left(\sum_{k=j}^{i} \frac{2}{k(k-1)} \right)^2 \right). \\
&= \sum_{j=2}^{i} \frac{2(j-1)}{i(i-1)} \left(\frac{4}{3} \left(\frac{1}{j^3} - \frac{1}{i^3} \right) + \left(\frac{2}{j-1} - \frac{2}{i} \right)^2 + O\left(\frac{1}{j^4} + \frac{1}{i^4} \right) \right). \\
&= \frac{2}{i(i-1)} (4 \log(i-1) - 3/2) + O(i^{-3}).
\end{aligned}$$

This implies that

$$\begin{aligned}
\mathbb{E}(T_{\text{total}}^{(2)}) &= n\mathbb{E}(b_n^2) + \sum_{i=2}^{n-1} \mathbb{E}(b_i^2) \\
&= \frac{8\log(n-1)}{n-1} + \sum_{i=2}^{n-1} \frac{8\log(i-1)}{i(i-1)} + O(1/n) \\
&= \frac{1}{2}(\log(2)+1) + \frac{7\log(n-1)}{n-1} + O(1/n),
\end{aligned}$$

which decreases asymptotically to the limit $(\log(2)+1)/2$ as n approaches infinity.

It may seem counterintuitive that $\mathbb{E}(T_{\text{total}}^{(2)})$ decreases as more lineages are sampled and $\mathbb{E}(T_{\text{total}})$ increases unboundedly, but in both simulated and real data we observe fewer SNPs in perfect LD in a sample of 2,184 haplotypes than in a subset of e.g. 1,000 haplotypes. To explain why, we note that the total tree length grows at rate $\log(n)$ as more lineages are sampled, but the tree length is subdivided among distinct branches at the faster rate $O(1/n)$. Because branch subdivision occurs faster than the growth rate of the total tree length, the sum of squared branch lengths decreases with increasing sample size, reducing the prevalence of independent linked SNPs and enhancing the signature of MNMs.

D.3 Simulating data with a realistic MNM distribution

We argue that MNM affects many features of genetic data including SNP density, the local transition/transverion ratio, and linkage disequilibrium. To capture these effects, it may be useful for readers to incorporate MNM into simulations of human-like SNP data. Here, we outline a strategy for doing this. Using the equations in D.2, one can first obtain empirical estimates $P_{\text{MNM}}(L, t)$ of the frequency of MNMs among SNP pairs of type t and spacing L. Then, to simulate a dataset with θ total SNPs, one should first use a program such as `ms` to generate a dataset with $\theta \times (1 - 0.009)$ total SNPs. After this, one should select a fraction $P(L, t)$ of SNPs uniformly at random to be MNMs of type t and spacing L. For each selected SNP, a new SNP should be introduced in perfect LD exactly L bp to the left.

Appendix E

Supporting Information for Chapter 6: Evidence for recent, population-specific evolution of the human mutation rate

E.1 A branch length ratio test for mutation rate change

Consider two haplotypes h_1 and h_2 sampled from populations P_1 and P_2, respectively. At each locus, h_1 and h_2 are related by a very simple coalescent tree consisting of two branches b_1 and b_2 (Figure E.1). If no mutation rate changes have occurred since the divergence of P_1 and P_2, the number of mutations falling on branches b_1 and b_2 are expected to be equal regardless of population demographic history. This reasoning motivates a simple branch length ratio test statistic (**BLR**).

Given a panel of sequences from P_1 and a panel of sequences from P_2, $\mathbf{BLR}(P_1, P_2)$ is computed by subsampling one haplotype from each panel, counting the derived alleles that appear on only the P_1 haplotype, and likewise counting the derived alleles that appear on only the P_2 haplotype. These counts can be summed across all haplotype pairs sampled from the two panels to yield derived allele counts D_1 and D_2. $\mathbf{BLR}(P_1, P_2)$ is then defined to be the ratio $\frac{D_1}{D_2}$, which is expected to equal 1 in the case of no mutation rate change.

Given a particular mutation type m (i.e. C→T transitions or TCC→T mutations), we can similarly define a branch length ratio $\mathbf{BLR}_m(P_1, P_2)$ that only counts derived alleles of mutation type m. It should hold that $\mathbf{BLR}_m(P_1, P_2) = 1$ for every mutation type m that has the same rate in P_1 and P_2.

Branch length ratios comparing Europeans (EUR) to Asians (ASN) were computed by subsampling pairs of haplotypes from the 1000 Genomes data and using the chimp PanTro4 reference to identify ancestral alleles. When $\mathbf{BLR}_m(\text{EUR, ASN})$ was computed separately for each transition and transversion type m, each derived allele count was greater in European sequences than Asian sequences, indicating that diverse mutation types appear to have higher rates in Europe than Asia. C→T mutations exhibited the largest apparent rate difference (Figure E.2).

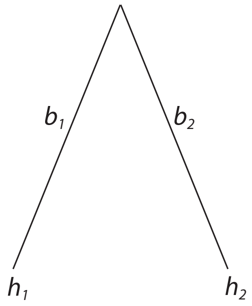


Figure E.1: This simple two-lineage coalescent tree shows the branches b_1 and b_2 that lead backward in time from haplotypes h_1 and h_2 to their most recent common ancestor at a given locus. Branches b_1 and b_2 have equal length; in the absence of any mutation rate differences, h_1 and h_2 should contain equal numbers of derived alleles.

Nonparametric bootstrapping was used to estimate branch length ratio variance for each mutation type. The genome was divided into 100 bins with approximately equal SNP counts, and 100 replicates were generated by resampling 100 bins with replacement. A branch length ratio of 1 lies within the 95% confidence interval for only two mutation types, A→T and C→A. Each other mutation type appears to have a significantly higher rate in Europe than Asia.

Branch length ratios for context-dependent mutations yield a more complex picture, with numerous mutation types appearing to have equal rates in Asia and Europe and a few types (notably GAA→ GTA and CCG→CAG) appearing to have higher rates in Asia (Figure E.3). As expected, TCC→T has the highest branch length ratio ($\mathbf{BLR}_{\text{TCC}\rightarrow\text{TTC}}(\text{EUR, ASN}) > 1.04$). Unexpectedly, the transversion type GAC→GCC has nearly as high a branch length ratio, indicating that this mutation is also a prime candidate for recent acceleration in Europe (or recent rate reduction in Asia). This signature merits further investigation, but since each transversion type is four times less common than a given transition type, GAC→GCC rate change makes less of an impact on total diversity than TCC→TTC rate change does.

E.2 Gradient of f(TCC) within Europe

One pattern that is visible in Figure 6.2.3 is a north-to-south gradient of f(TCC) within Europe. The southern Spanish and Italian populations have the highest mean f(TCC) values (0.0335 and 0.0337, respectively), while the central British and CEU values (0.0325 and 0.0326) are intermediate and the northern Finnish value (0.0313) is lowest.

One demographic event that might have contributed to this f(TCC) gradient is gene flow from Asia into northeast Europe. Lazaridis et al. (2014) and Hellenthal et al. (2014) each inferred that the Finns have an Asian ancestry component, probably the result of gene flow from Siberia. In support of this, f(TCC) appears to be inversely correlated with

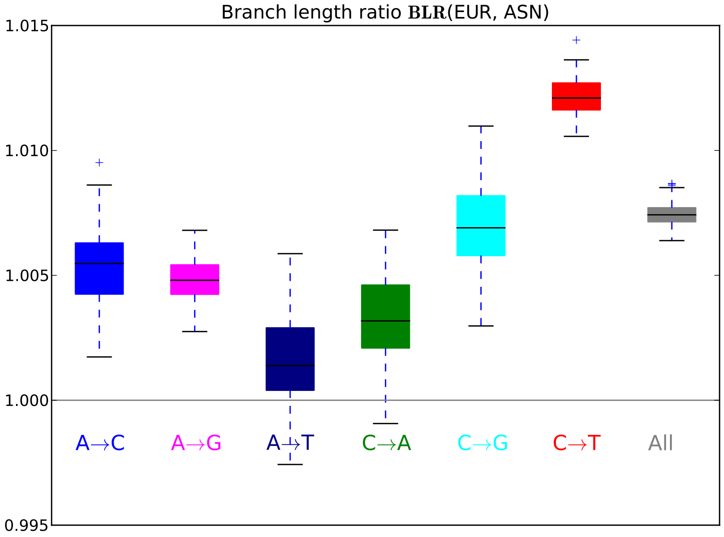


Figure E.2: This box plot shows branch length ratio distributions for each type of transition and transversion. For simplicity, each pair of complementary mutations (e.g. C→T and G→A) has been merged into a single category.

European/Asian allele sharing. Specifically, I looked at the variant set PAsE of 913,662 SNPs that are fixed in Africa but variable in both Asia and Europe. For each European haplotype h, I counted the number dPAsE(h) of derived alleles from the set PAsE that occur on haplotype h. As expected given gene flow from Siberia into Finland, dPAsE(h) is highest in the Finns, lowest in the Spanish and Italians, and inversely correlated with f_h(TCC) across all haplotypes h sampled in Europe (regression $p < 2.17 \times 10^{-4}$, Figure 6.2.4).

A second possibility, not exclusive of the first, is that the f(TCC) gradient was created by ancient admixture among early European founder populations. Lazaridis, *et al.* have argued that Europeans are the admixed descendants of three genetically distinct groups: early European farmers (EEF), west European hunter-gatherers (WHG), and Ancient north Eurasians (ANE). The Italians, Spanish, and other southern European populations are inferred to have relatively high EEF ancestry fractions, compared to intermediate EEF ancestry levels in the English and low EEF ancestry in the Finns. This is consistent with a scenario where the TCC→TTC mutation rate change first occurred within the EEF population.

The light skin pigmentation alleles that are widespread in modern Europe are similarly believed to have originated in the EEF population, perhaps to compensate for low Vitamin D levels in a diet of cultivated grains. The hunter-gatherer populations that admixed with these early farmers show genetic indications of dark skin and hair (Lazaridis et al. 2014; Olalde et al. 2014). This is interesting in light of the association mentioned in the main text discussion between TCC→T mutations and melanoma, a skin cancer whose incidence is strongly predicted by light skin and European ancestry. There might also be some causal

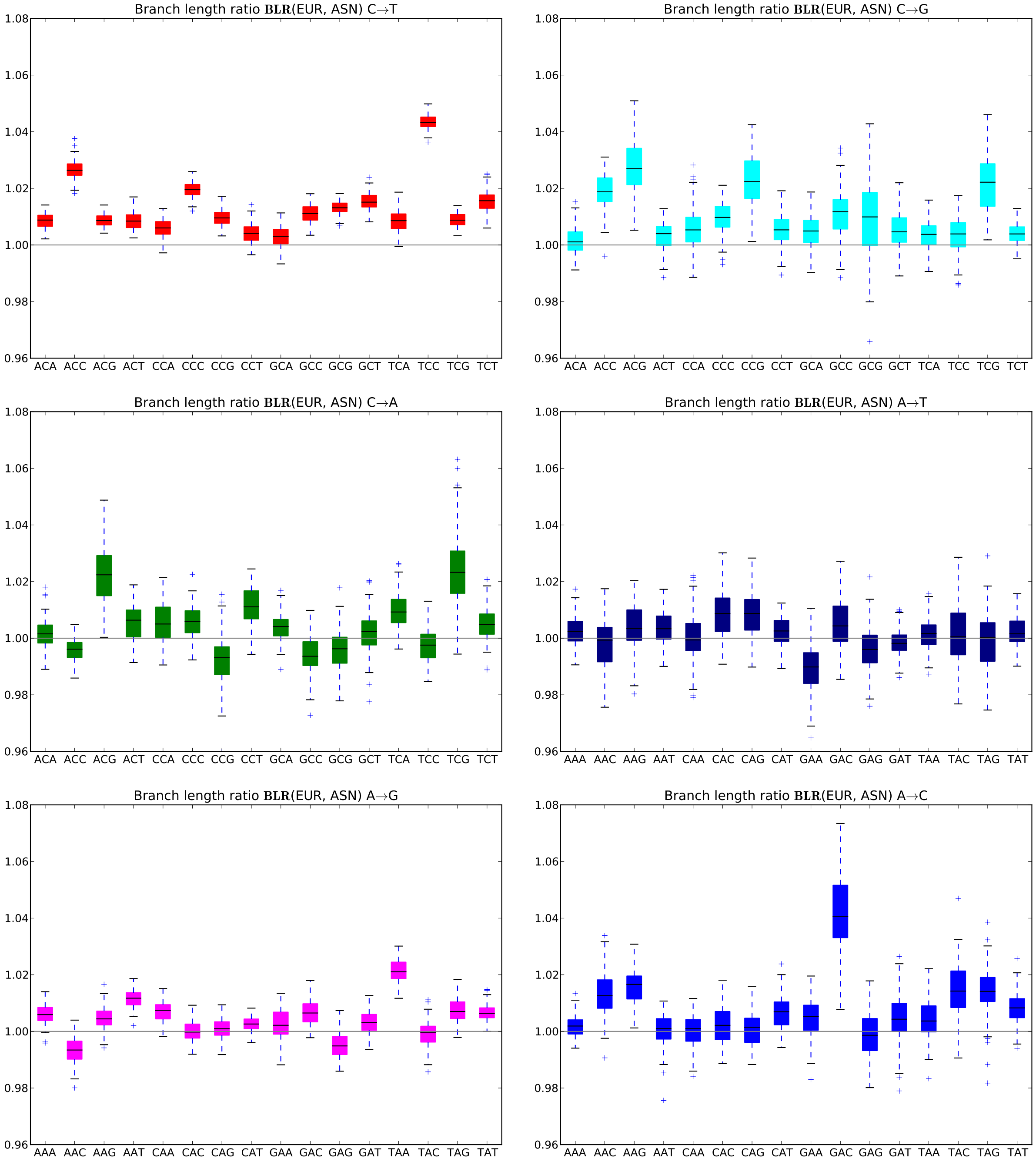


Figure E.3: These box plots shows branch length ratio distributions for context-dependent transition and transversion types. Each set of axes is restricted to a single mutation type (e.g. C→T transitions, C→G transversions, etc), and each individual bar plot is labeled with the ancestral nucleotide flaked by its 3' and 5' neighbors. Each complementary mutation pair (e.g. CCC→T and GGG→A) has been merged into a single mutation category.

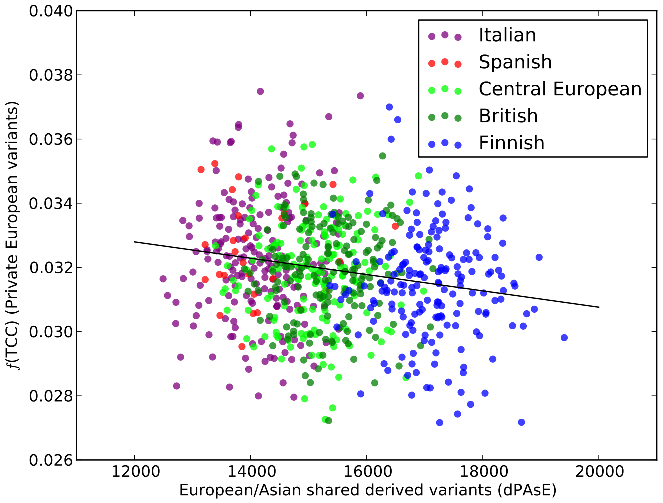


Figure E.4: Each point in this scatterplot shows $f_h(\text{TCC})$ and $\text{dPAsE}(h)$ for a particular European haplotype h, revealing a negative correlation between the frequency of private European m_{TCC} variants and the number of SNPs shared with Asia but not with Africa (solid line, regression slope -2.54×10^{-7}, $p < 2.17 \times 10^{-4}$).

link between higher TCC→T frequencies and higher UV exposure at southern latitudes.

E.3 TCC→T mutations in Complete Genomics data

To ensure that the results presented in this paper are not specific to a single sequencing platform or consortium, 62 human genomes sequenced by by Complete Genomics (CG) were downloaded from `www.completegenomics.com /public-data/69-Genomes/` and analyzed (Drmanac et al. 2010). These 62 unrelated individuals include the 54-member CG diversity panel, the parents of the Yoruban trio and Puerto Rican trio, and the four grandparents of the 17-member CEPH pedigree. This dataset contains representatives of 11 populations: two European (CEU, TSI), three Asian (CHB, JPT, and GIH (Gujarati Indians from Houston)), three African (YRI, LWK, and MKK (Maasai from Kenya)), and three admixed (ASW, MXL, PUR). There are 13 Europeans, 12 Asians, 17 Africans, and 12 admixed individuals from the Americas.

Population-private SNP sets within the CG panel were defined independently of the 1000 Genomes data. Looking only at variation within the CG panel, the private European set PE(CG) contains SNPs that are variable in CEU or TSI and not variable in CHB, JPT, GIH, YRI, LWK, or MKK. Similarly, the private Asian set PAs(CG) contains SNPs that are variable in CHB, JPT, or GIH and not variable in the CEU, TSI, YRI, LWK, or MKK. The private African set PAf(CG) contains SNPs that are variable in YRI, LWK, or MKK and not variable in CEU, TSI, CHB, JPT, or GIH. Singletons were excluded to minimize the impact of sequencing error. Using the private SNP sets PE(CG), PAs(CG), and PAf(CG), frequency differences $(f_{\mathrm{PE}}(m) - f_{\mathrm{PAf}}(m))/f_{\mathrm{PAf}}(m)$ and $(f_{\mathrm{PAs}}(m) - f_{\mathrm{PAf}}(m))/f_{\mathrm{PAf}}(m)$ were computed along with χ^2-based p values from the contingency tables in Figure6.2.1C,D of the main text.

The results are depicted in Figure E.5A,B, a volcano plot analogous to Figure 6.2.1A,B from the main text. It can be seen that TCC→T is again the major outlier in the comparison of Europe to Africa, with a significance that dwarfs that of all outliers in the Asia-to-Africa comparison. The minor outliers TCT→ TTT, AGA→ AAA, GGT→GAT, and ACC→ ATC are also significantly more abundant in Europeans than Asians or Africans in both CG and 1000 Genomes.

All outliers in Figure E.5A have lower p-value significance than the corresponding outliers in Figure 6.2.1 of the main text. There are two reasons why the difference can be attributed to sample size. Intuitively, the CG data contains fewer individuals than the 1000 Genomes data and contains proportionately fewer SNPs. In addition, population-private SNPs in the CG data are ascertained with less certainty than population-private SNPs in 1000 Genomes. For example, if a particular SNP originated in Africa and is segregating today in both Africa and Europe, there is a chance that no Africans carrying the derived allele will be sampled, leading the SNP to be classified as private European variation. This misclassification should happen more often in a 62-genome panel than in a 1,092-genome panel.

To demonstrate that sample size can account for the difference between Figure 6.2.1 and Figure E.5A,B, I subsampled 13 Europeans, 12 Asians, and 17 Africans from the 1000 Genomes data and ascertained the sets of mutations that appear to be continent-private with respect to this dataset. The population composition of the CG panel was mirrored except for substitution of 4 CHS individuals for GIH and 4 LWK individuals for MKK. Since the Complete Genomics data are not imputed and thus cannot be filtered for imputation quality, no imputation quality filtering was performed on the 1000 Genomes data for the purpose of this analysis. As shown in Figure E.5C,D, all p values calculated from the subsampled 1000 Genomes panel are very close to the p values calculated from the CG panel.

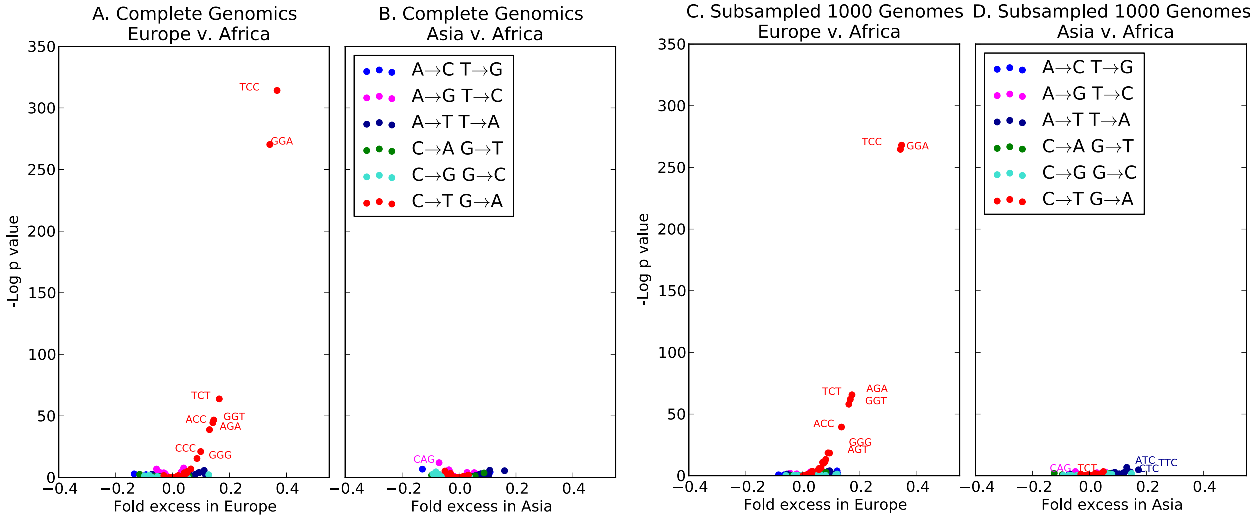


Figure E.5: Panels A and B, analogous to Figure 6.2.1A,B of the main text, show differences in context-dependent mutation frequencies between the Complete Genomics populations. Panels C and D represent the same frequency differences in a panel subsampled from the 1000 Genomes data to mirror the sample size and population makeup of the CG panel.

E.4 Estimating the time of TCC→TTC acceleration in Europe

The program ARGweaver recently developed by Rasmussen et al. (2014) uses patterns of haplotype diversity to infer the approximate genealogical history of a collection of haplotypes. This makes it possible to estimate the age of a particular derived allele based on the extent of divergence among the sequences that carry it. Rasmussen, et al. applied ARGweaver to the Complete Genomics data utilized in Section E.3 and published the resulting estimates of genealogy and allele age on the following server:

`http://compgen.bscb.cornell.edu/ARGweaver/CG_results/`

Rather than estimating a single genealogy for the Complete Genomics data, ARGweaver generates a distribution of probable genealogies that reflect the uncertainty of estimating history from present-day diversity. This makes it possible to extract a mean derived allele age estimate for each SNP. ARGweaver estimates a posterior distribution of allele age by computing a posterior distribution of genealogies, identifying the branch of each genealogy on which the mutation must have occurred, and placing the mutation uniformly at random on this branch.

I downloaded the mean ARGweaver allele age estimates for each TCC→T variant that appears private to Europeans or Asians in both the Complete Genomics data and the 1000 Genomes data. As a control, I also downloaded mean allele age estimates for private CCT→CTT variants, which are similarly abundant to TCC→TTC variants but do not show much evidence of mutation rate divergence between Asia and Europe. These allele age estimates were grouped into bins such that each age bin spans at least 100 generations and contains at least 500 private European CCT→CTT variants. Figure E.6A shows the number of private European and private Asian TCC→TTC and CCT→CTT variants that fall into each bin. ARGweaver's age estimates (given in generations) were converted to years before the present by assuming a human generation time of 29 years.

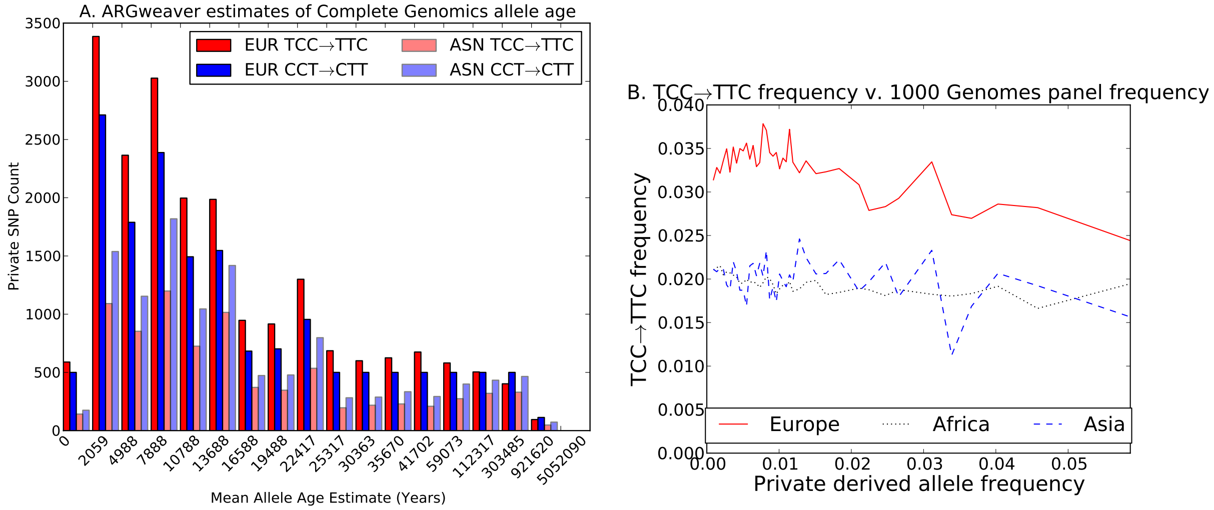


Figure E.6: A. This bar plot shows the distribution of private European (EUR) and Asian (ASN) allele ages computed by Rasmussen et al. (2014) from Complete Genomics data using ARGweaver. For all but the most ancient alleles, TCC→TCC is more frequent in Europe in than the control mutation type CCT→CTT, indicating that TCC→TTC acceleration in Europe probably happened relatively soon after Europeans and Asians diverged. B. For this figure, private European, Asian, and African alleles from the 1000 Genomes data were partitioned into bins based on derived allele frequency (DAF). Within each DAF bin, the frequency of TCC→TTC (plus GGA→GAA) was calculated and plotted as a function of DAF. For private Asian and African alleles, TCC→TTC frequency appears to be independent of DAF (hovering around 2%). In contrast, private European TCC→TTC frequency is a decreasing function of DAF, indicating that higher DAF categories contain a lower percentage variants that arose later than the time of TCC→TTC acceleration in Europe.

As shown in Figure E.6A, each time bin contains slightly more private Asian CCT→CTT variants compared to private Asian TCC→TTC variants. However, the only bins containing more private European CCT→CTT variants than private European TCC→TTC variants are the two most ancient time bins ($>$ 303 thousand years ago (kya)). The next-most-ancient bin (112–303 kya) contains 1% more private European TCC→TTC variants compared to CCT→CTT, and all other bins contain 16–37% more private European TCC→TTC variants compared to CCT→CTT, with the TCC→TTC-to-CCT→CTT ratio peaking around 25 kya. These ARGweaver estimates suggest that the European mutation rate changed very soon after Europeans diverged from Asians between 40,000 and 80,000 years ago (Scally & Durbin 2012), a result that is consistent with the relatively uniform distribution of excess TCC→TTC variants among diverse European populations.

Due to sheer sample size, the 1000 Genomes dataset contains more information about allele age than the Complete Genomics dataset does. Figure E.6B plots TCC→TTC frequency as a function of minor allele frequency in the 1000 Genomes data, showing that TCC→TTC SNPS comprise close to to 3.5% of private European variants that have less than 1% frequency in the entire 1000 Genomes panel. In contrast, TCC→TTC SNPs comprise fewer than 2.5% of older private European variants that occur in 5–6% of the 1000 Genomes haplotypes. Unfortunately, methods like ARGweaver are not yet scalable to datasets containing hundreds of haplotypes, making it hard to rigorously incorporate this information into a better estimate of the TCC→TTC acceleration time.

E.5 Stratifying the genome by GC content

To assess the effect of GC content on mutation spectrum differences between the 1000 Genomes populations, the hg19 reference genome was partitioned into 100 kb bins. Within each bin, the ratio of G/C base pairs to total base pairs was computed, excluding sites annotated as N's. Bins containing more than 50% N's were excluded from the analysis. This yielded a distribution of GC content percentages ranging from 0% GC to 63.6% GC, which was partitioned into 10 quantiles containing the same number (2,862) of hg19 bins. Table E.5 lists the upper and lower GC content percentages for each of the 10 quantiles.

Qtl	%GC	Qtl	%GC	Qtl	%GC	Qtl	%GC	Qtl	%GC
1	0–35.5	2	35.5–36.6	3	36.6–37.7	4	37.7–38.7	5	38.7–39.8
6	39.8–41.1	7	41.1–42.6	8	42.6–44.7	9	44.7–48.1	10	48.1–63.6

Lists of private European, Asian, and African SNPs were compiled within each GC quantile, and the χ^2 metric from Figure 6.2.1 of the main text was used to assess mutation spectrum differences. Tables E.1–E.10, one for each quantile, list the top-ranked 10 SNPs that show frequency differences between Europe and Africa. Similarly, Tables E.11–E.20 contain lists of the SNPs that show the most significant frequency differences between Asia and Africa.

The Europe v. Africa results show much more consistency across GC-content bins than do the Asia v. Africa results. Almost all Europe v. Africa outliers are C→T/G→A transitions, whereas the Asia v. Africa outliers contain a variety of transitions and transversions. In addition, 6 of the top 10 outliers for Europe v. Africa Quantile 1, including TCC→TTC/GGA→GAA, appear in the top 10 outliers for at least 9 out of 10 Europe v. Africa GC quantiles. In contrast, GAT→GTT is the only mutation that appears in the top

10 outliers for every Asia vs. Africa quantile. In each table, the column “T10-EUR" records the number of distinct GC quantiles for which a given mutation appears in the Europe v. Africa top 10; similarly, “T10-ASN" records the number of GC quantiles for which the mutation appears in the Europe v. Asia top 10. The Europe v. Africa outliers are mostly disjoint from the Asia v. Africa outliers.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.464e-02	>	7.986e-03	1.833e+00	6.33e-39	10	0
GGA	A	1.336e-02	>	8.131e-03	1.644e+00	5.00e-24	10	0
AGA	A	1.279e-02	>	1.052e-02	1.216e+00	7.49e-04	9	0
ACC	T	1.066e-02	>	8.683e-03	1.227e+00	0.001	9	2
TCT	T	1.256e-02	>	1.068e-02	1.175e+00	0.010	10	0
ATT	C	2.050e-02		2.282e-02	8.983e-01	0.037	2	1
TCA	G	2.866e-03		3.779e-03	7.584e-01	0.048	1	0
GAA	G	4.950e-03		6.061e-03	8.167e-01	0.064	1	0
CAA	G	6.200e-03		7.400e-03	8.379e-01	0.074	1	0
GTT	C	7.347e-03		8.632e-03	8.511e-01	0.078	1	0

Table E.1: Quantile 1: 0–35.5% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.418e-02	>	8.501e-03	1.668e+00	3.62e-29	10	0
GGA	A	1.444e-02	>	8.740e-03	1.653e+00	1.21e-28	10	0
GGT	A	1.207e-02	>	8.844e-03	1.365e+00	2.12e-09	9	0
CTG	C	6.096e-03		8.035e-03	7.586e-01	3.83e-04	2	1
TCT	T	1.283e-02	>	1.069e-02	1.200e+00	0.001	10	0
AAG	G	6.143e-03		7.531e-03	8.157e-01	0.019	1	2
TCT	G	8.562e-03	>	7.193e-03	1.190e+00	0.021	1	1
AAT	G	1.924e-02		2.148e-02	8.957e-01	0.028	1	2
CCT	G	4.506e-03	>	3.582e-03	1.258e+00	0.032	1	1
ACT	A	2.467e-03		3.294e-03	7.489e-01	0.044	1	0

Table E.2: Quantile 2: 35.5–36.6% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.484e-02	>	8.671e-03	1.711e+00	9.55e-36	10	0
GGA	A	1.418e-02	>	8.717e-03	1.626e+00	4.24e-28	10	0
TCT	T	1.296e-02	>	1.045e-02	1.241e+00	2.41e-05	10	0
GGT	A	1.105e-02	>	9.036e-03	1.222e+00	4.87e-04	9	0
ACC	T	1.129e-02	>	9.412e-03	1.199e+00	0.002	9	2
CTT	C	5.997e-03		7.544e-03	7.948e-01	0.004	1	0
CCA	A	2.183e-03		3.013e-03	7.244e-01	0.023	1	0
TGG	A	8.752e-03		1.023e-02	8.560e-01	0.032	1	1
AGA	A	1.208e-02	>	1.064e-02	1.135e+00	0.049	9	0
GCT	T	1.082e-02	>	9.525e-03	1.136e+00	0.068	1	0

Table E.3: Quantile 3: 36.6–37.7% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.552e-02	>	9.034e-03	1.718e+00	5.00e-40	10	0
GGA	A	1.457e-02	>	8.800e-03	1.655e+00	1.11e-32	10	0
ACC	T	1.208e-02	>	9.602e-03	1.258e+00	6.45e-06	9	2
GGT	A	1.175e-02	>	9.519e-03	1.234e+00	6.72e-05	9	0
AGA	A	1.241e-02	>	1.016e-02	1.221e+00	1.09e-04	9	0
TCT	T	1.245e-02	>	1.066e-02	1.168e+00	0.005	10	0
CCG	A	3.108e-04		7.300e-04	4.257e-01	0.012	1	0
ATT	G	2.528e-03		3.360e-03	7.524e-01	0.028	1	1
TAC	G	6.278e-03		7.383e-03	8.503e-01	0.064	1	1
AAC	G	6.692e-03		7.804e-03	8.576e-01	0.074	1	2

Table E.4: Quantile 4: 37.7–38.7% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
GGA	A	1.648e-02	>	9.475e-03	1.739e+00	2.08e-45	10	0
TCC	T	1.569e-02	>	9.329e-03	1.682e+00	2.88e-38	10	0
TCT	T	1.302e-02	>	1.005e-02	1.296e+00	3.49e-08	10	0
AGA	A	1.343e-02	>	1.042e-02	1.288e+00	4.33e-08	9	0
GGT	A	1.179e-02	>	1.005e-02	1.173e+00	0.005	9	0
TGA	A	1.179e-02	>	1.017e-02	1.160e+00	0.010	2	0
ACC	T	1.151e-02	>	9.961e-03	1.155e+00	0.015	9	2
TGC	C	1.555e-03		2.267e-03	6.858e-01	0.017	1	0
CGG	A	1.212e-02		1.377e-02	8.797e-01	0.030	1	2
CCC	T	9.732e-03	>	8.526e-03	1.141e+00	0.060	3	1

Table E.5: Quantile 5: 38.7–39.8% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.606e-02	>	9.553e-03	1.681e+00	4.03e-40	10	0
GGA	A	1.598e-02	>	9.509e-03	1.680e+00	6.99e-40	10	0
GGT	A	1.274e-02	>	9.954e-03	1.279e+00	1.82e-07	9	0
AGA	A	1.275e-02	>	1.026e-02	1.243e+00	6.73e-06	9	0
ACC	T	1.152e-02	>	9.839e-03	1.170e+00	0.005	9	2
TCT	T	1.217e-02	>	1.045e-02	1.164e+00	0.006	10	0
CAC	G	5.405e-03		6.648e-03	8.130e-01	0.013	2	2
TTT	C	8.352e-03		9.805e-03	8.518e-01	0.019	1	0
CCA	G	3.538e-03	>	2.757e-03	1.283e+00	0.022	1	1
GGG	A	1.028e-02	>	8.917e-03	1.153e+00	0.026	3	1

Table E.6: Quantile 6: 39.8–41.1% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
GGA	A	1.740e-02	>	9.840e-03	1.769e+00	8.11e-53	10	0
TCC	T	1.709e-02	>	9.983e-03	1.712e+00	2.97e-46	10	0
GGT	A	1.341e-02	>	1.037e-02	1.293e+00	1.32e-08	9	0
TCT	T	1.296e-02	>	1.022e-02	1.269e+00	3.36e-07	10	0
AGA	A	1.283e-02	>	1.015e-02	1.264e+00	6.72e-07	9	0
ACC	T	1.234e-02	>	1.034e-02	1.193e+00	5.93e-04	9	2
GTG	C	5.353e-03		6.830e-03	7.838e-01	0.002	2	1
CCC	T	1.139e-02	>	9.599e-03	1.186e+00	0.002	3	1
TAA	T	1.577e-03		2.330e-03	6.769e-01	0.009	1	0
AGC	A	1.119e-02	>	9.781e-03	1.144e+00	0.027	1	0

Table E.7: Quantile 7: 41.1–42.6% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) − r(PAf)	T10-EUR	T10-ASN
TCC	T	1.696e-02	>	9.915e-03	1.711e+00	1.47e-47	10	0
GGA	A	1.691e-02	>	9.985e-03	1.693e+00	1.25e-45	10	0
AGA	A	1.295e-02	>	9.472e-03	1.367e+00	1.27e-12	9	0
ACC	T	1.278e-02	>	1.023e-02	1.250e+00	1.70e-06	9	2
TCT	T	1.237e-02	>	9.970e-03	1.241e+00	6.26e-06	10	0
GGT	A	1.306e-02	>	1.071e-02	1.220e+00	2.41e-05	9	0
TAC	T	5.039e-04		9.842e-04	5.120e-01	0.008	1	0
ATT	C	1.433e-02		1.631e-02	8.790e-01	0.009	2	1
GCG	G	3.546e-04		7.752e-04	4.574e-01	0.009	1	0
GAC	G	2.743e-03		3.645e-03	7.527e-01	0.012	1	0

Table E.8: Quantile 8: 42.6–44.7% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) $-$ r(PAf)	T10-EUR	T10-ASN
GGA	A	1.813e-02	>	1.061e-02	1.708e+00	1.61e-50	10	0
TCC	T	1.741e-02	>	1.049e-02	1.660e+00	1.45e-43	10	0
GGT	A	1.456e-02	>	1.085e-02	1.343e+00	1.24e-12	9	0
AGA	A	1.271e-02	>	9.617e-03	1.321e+00	6.91e-10	9	0
ACC	T	1.386e-02	>	1.097e-02	1.264e+00	1.06e-07	9	2
GGG	A	1.377e-02	>	1.175e-02	1.171e+00	0.001	3	1
GTG	C	6.186e-03		7.713e-03	8.020e-01	0.002	2	1
CCC	T	1.377e-02	>	1.187e-02	1.159e+00	0.003	3	1
TGA	A	1.105e-02	>	9.396e-03	1.176e+00	0.003	2	0
TCT	T	1.116e-02	>	9.592e-03	1.164e+00	0.007	10	0

Table E.9: Quantile 9: 44.7–48.1% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PE)		r(PAf)	r(PE)/r(PAf)	p value r(PE) $-$ r(PAf)	T10-EUR	T10-ASN
TCC	T	2.019e-02	>	1.141e-02	1.770e+00	9.36e-62	10	0
GGA	A	1.913e-02	>	1.123e-02	1.704e+00	4.00e-51	10	0
ACC	T	1.449e-02	>	1.093e-02	1.325e+00	3.83e-11	9	2
GGT	A	1.483e-02	>	1.129e-02	1.314e+00	9.66e-11	9	0
TCT	T	1.098e-02	>	8.455e-03	1.298e+00	2.56e-07	10	0
GGG	A	1.783e-02	>	1.462e-02	1.220e+00	5.49e-07	3	1
AGA	A	1.090e-02	>	8.722e-03	1.250e+00	2.11e-05	9	0
CAG	G	9.767e-03		1.223e-02	7.986e-01	3.09e-05	1	2
CAC	G	6.198e-03		8.007e-03	7.741e-01	2.28e-04	2	2
CTG	C	1.017e-02		1.229e-02	8.277e-01	6.29e-04	2	1

Table E.10: Quantile 10: 48.1–100% GC. Top-ranked SNPs showing frequency differences between Europe and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
GAT	C	7.286e-04		1.559e-03	4.674e-01	0.052	0	1
CCT	T	1.384e-02	>	1.157e-02	1.196e+00	0.055	0	1
GAA	T	2.914e-03	>	2.019e-03	1.443e+00	0.084	0	1
TGG	C	3.195e-03	>	2.255e-03	1.417e+00	0.087	0	1
CAC	G	4.259e-03		5.699e-03	7.474e-01	0.098	2	2
GGA	T	3.923e-03	>	2.908e-03	1.349e+00	0.114	0	1
CAT	G	2.303e-02	>	2.052e-02	1.123e+00	0.149	0	2
AAT	G	1.962e-02		2.210e-02	8.875e-01	0.179	1	2
CTT	G	3.867e-03	>	2.949e-03	1.311e+00	0.186	0	1
AAG	G	8.911e-03	>	7.499e-03	1.188e+00	0.210	1	2

Table E.11: Quantile 1: 0–35.5% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
GTC	A	2.490e-03	>	1.486e-03	1.676e+00	0.005	0	3
GTA	A	2.149e-03	>	1.325e-03	1.622e+00	0.022	0	1
ATT	C	1.875e-02		2.193e-02	8.549e-01	0.027	2	1
CGT	A	1.626e-02		1.904e-02	8.542e-01	0.045	0	1
AGG	C	4.834e-03	>	3.617e-03	1.337e+00	0.050	0	2
GCG	T	6.788e-03		8.507e-03	7.979e-01	0.076	0	1
GGT	C	3.223e-03	>	2.331e-03	1.383e+00	0.090	0	1
TAA	G	1.109e-02	>	9.382e-03	1.182e+00	0.112	0	1
CGT	T	5.860e-04		1.127e-03	5.199e-01	0.159	0	1
CAT	G	2.261e-02	>	2.040e-02	1.109e+00	0.193	0	2

Table E.12: Quantile 2: 35.5–36.6% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
TGC	T	2.762e-03		4.107e-03	6.724e-01	0.017	0	1
AAT	G	1.778e-02		2.080e-02	8.552e-01	0.018	1	2
TCC	G	4.352e-03	>	3.199e-03	1.360e+00	0.027	0	1
AGT	C	6.277e-03	>	4.921e-03	1.276e+00	0.040	0	2
TCT	G	9.373e-03	>	7.788e-03	1.204e+00	0.065	1	1
TTG	A	2.594e-03	>	1.821e-03	1.424e+00	0.066	0	1
CAT	T	5.565e-03	>	4.391e-03	1.267e+00	0.073	0	2
GGG	C	3.431e-03	>	2.559e-03	1.341e+00	0.087	0	3
CCT	A	1.674e-03		2.478e-03	6.755e-01	0.112	0	2
GTT	G	1.297e-03		2.019e-03	6.426e-01	0.115	0	2

Table E.13: Quantile 3: 36.6–37.7% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
TAC	G	5.646e-03		7.383e-03	7.647e-01	0.022	1	1
AGC	C	3.723e-03	>	2.715e-03	1.371e+00	0.038	0	1
ACT	G	5.973e-03	>	4.750e-03	1.257e+00	0.067	0	1
ACC	T	8.019e-03		9.602e-03	8.351e-01	0.106	9	2
TGG	A	8.673e-03		1.029e-02	8.429e-01	0.114	1	1
GAC	T	2.046e-03	>	1.452e-03	1.409e+00	0.142	0	3
GGG	A	6.873e-03		8.214e-03	8.368e-01	0.164	3	1
GTT	A	2.496e-03	>	1.849e-03	1.350e+00	0.165	0	1
TGT	A	1.591e-02		1.787e-02	8.904e-01	0.168	0	1
TAG	C	2.086e-03	>	1.511e-03	1.381e+00	0.177	0	1

Table E.14: Quantile 4: 37.7–38.7% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
AGT	C	6.272e-03	>	4.757e-03	1.318e+00	0.008	0	2
GAC	T	2.464e-03	>	1.587e-03	1.553e+00	0.008	0	3
GTC	C	2.539e-03		3.678e-03	6.902e-01	0.029	0	1
AGG	C	5.414e-03	>	4.189e-03	1.292e+00	0.032	0	2
TCC	A	3.771e-03	>	2.854e-03	1.321e+00	0.065	0	1
ATT	G	2.091e-03		3.011e-03	6.943e-01	0.066	1	1
CAC	G	5.190e-03		6.538e-03	7.938e-01	0.069	2	2
CGG	T	3.360e-04		7.468e-04	4.499e-01	0.118	0	1
CCT	A	1.643e-03		2.373e-03	6.922e-01	0.124	0	2
CTT	A	2.240e-03	>	1.648e-03	1.359e+00	0.157	0	1

Table E.15: Quantile 5: 38.7–39.8% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
CCA	T	8.341e-03		1.059e-02	7.876e-01	0.006	0	1
GGC	A	6.155e-03		8.119e-03	7.582e-01	0.006	0	1
GTC	G	1.311e-03	>	8.498e-04	1.543e+00	0.102	0	1
CTA	A	7.649e-04		1.284e-03	5.958e-01	0.137	0	1
CAG	G	8.377e-03		9.756e-03	8.586e-01	0.165	1	2
GCA	A	4.480e-03	>	3.641e-03	1.230e+00	0.180	0	1
AAC	G	8.341e-03	>	7.196e-03	1.159e+00	0.199	1	2
GAA	C	2.914e-03	>	2.271e-03	1.283e+00	0.205	0	1
GTC	A	1.967e-03	>	1.463e-03	1.344e+00	0.226	0	3
CGG	A	1.362e-02		1.519e-02	8.967e-01	0.236	1	2

Table E.16: Quantile 6: 39.8–41.1% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
ATG	A	5.017e-03	>	3.976e-03	1.262e+00	0.068	0	1
GCA	T	1.003e-02	>	8.525e-03	1.177e+00	0.070	0	2
CCT	G	5.578e-03	>	4.516e-03	1.235e+00	0.087	1	1
GTT	G	1.052e-03		1.677e-03	6.274e-01	0.095	0	2
GTG	C	5.578e-03		6.830e-03	8.167e-01	0.102	2	1
ACC	T	8.806e-03		1.034e-02	8.515e-01	0.103	9	2
GTA	G	1.438e-03	>	9.615e-04	1.496e+00	0.108	0	1
GAC	T	1.965e-03	>	1.419e-03	1.384e+00	0.143	0	3
GAG	G	3.578e-03		4.513e-03	7.929e-01	0.155	0	2
AAG	G	6.104e-03		7.288e-03	8.376e-01	0.158	1	2

Table E.17: Quantile 7: 41.1–42.6% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
GGG	C	4.816e-03	>	3.625e-03	1.329e+00	0.013	0	3
CTC	C	3.365e-03		4.599e-03	7.317e-01	0.023	0	1
ACA	A	4.486e-03	>	3.546e-03	1.265e+00	0.075	0	1
CCC	T	8.940e-03		1.039e-02	8.604e-01	0.121	3	1
TTT	A	1.748e-03		2.421e-03	7.221e-01	0.146	0	1
ATC	A	3.794e-03	>	3.038e-03	1.249e+00	0.157	0	1
GGT	T	5.872e-03	>	4.931e-03	1.191e+00	0.171	0	1
ATA	G	1.188e-03		1.732e-03	6.858e-01	0.177	0	1
GTA	C	4.618e-03		5.595e-03	8.254e-01	0.180	0	1
CCG	T	1.768e-02		1.950e-02	9.069e-01	0.180	0	1

Table E.18: Quantile 8: 42.6–44.7% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
CCA	G	4.220e-03	>	3.141e-03	1.343e+00	0.016	1	1
CAT	T	4.350e-03	>	3.260e-03	1.334e+00	0.017	0	2
AAC	G	5.096e-03		6.396e-03	7.968e-01	0.053	1	2
GAT	T	3.506e-03	>	2.657e-03	1.319e+00	0.055	0	1
CGA	A	1.772e-02	>	1.569e-02	1.129e+00	0.057	0	1
GCA	T	1.045e-02	>	8.976e-03	1.164e+00	0.075	0	2
GAG	G	3.928e-03		4.979e-03	7.889e-01	0.091	0	2
TCG	T	1.733e-02	>	1.551e-02	1.117e+00	0.105	0	1
TGA	T	1.883e-03		2.611e-03	7.212e-01	0.115	0	1
CGC	C	4.869e-04		9.045e-04	5.383e-01	0.127	0	1

Table E.19: Quantile 9: 44.7–48.1% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

$B_{5'}B_AB_{3'}$	B_D	r(PAs)		r(PAf)	r(PAs)/r(PAf)	p value r(PAs) − r(PAf)	T10-EUR	T10-ASN
CTG	C	1.006e-02		1.229e-02	8.184e-01	0.006	2	1
ACG	T	3.382e-02	>	3.046e-02	1.110e+00	0.010	0	1
GTG	G	1.288e-03		2.140e-03	6.017e-01	0.014	0	1
CAG	G	1.024e-02		1.223e-02	8.372e-01	0.019	1	2
GTC	A	2.177e-03	>	1.516e-03	1.436e+00	0.039	0	3
CGG	A	3.642e-02	>	3.352e-02	1.087e+00	0.051	1	2
CGT	C	9.198e-04		1.537e-03	5.985e-01	0.053	0	1
CCC	A	4.967e-03	>	3.960e-03	1.254e+00	0.056	0	1
GTG	A	2.299e-03	>	1.654e-03	1.390e+00	0.062	0	1
GGG	C	6.714e-03	>	5.550e-03	1.210e+00	0.064	0	3

Table E.20: Quantile 10: 48.1–100% GC. Top-ranked SNPs showing frequency differences between Asia and Africa.

E.6 Singleton variants in 1000 Genomes and Complete Genomics

Singleton variants that occur within only a single genome were excluded from the analyses in this paper because of concerns about their quality. Such concerns appear significant given the alterations to Figure 6.2.1A,B and Figure E.5A,B that are produced by including singletons in the analysis. The volcano plots in Figure E.7A,B, which compare the 1000 Genomes populations with singletons included, show many apparent frequency differences between populations that do not show up when singletons are excluded, particularly in the Asia versus Africa comparison. The plots in Figure E.7C,D show that the same is true for the Complete Genomics data, despite its higher coverage. Figure E.7A,B is qualitatively very different from Figure E.7C,D, suggesting that the Illumina/SOLID 1000 Genomes pipeline and the Complete Genomics pipeline have very different error patterns with regard to calling singletons.

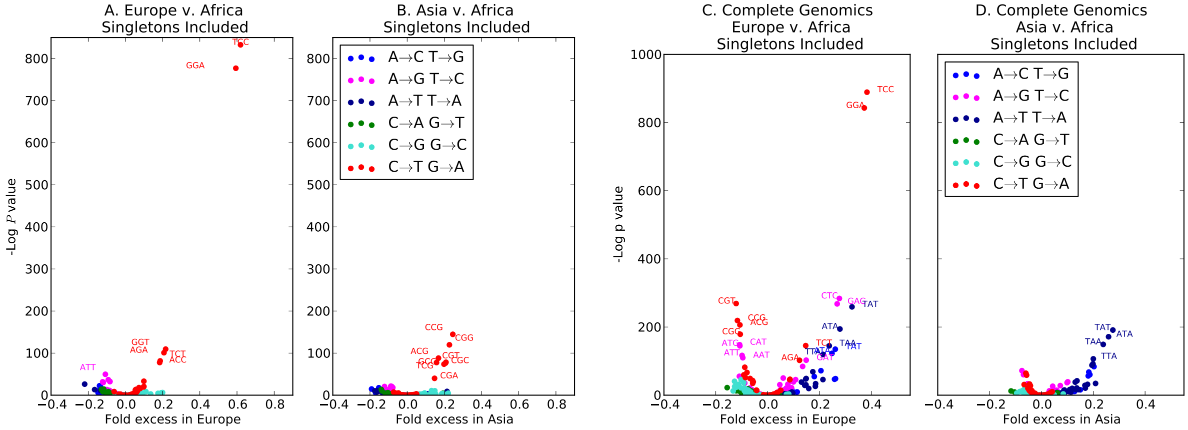


Figure E.7: Panels A and B display frequency differences between continental groups in the 1000 Genomes Phase I data. These figures were produced in the same way as Figure 6.2.1A,B of the main text except that singletons (variants of minor allele count 1) were included. Similarly, Panels C and D show differences between the same groups in the Complete Genomics data. These panels were produced in the same way as Supplementary Figure E.5A,B except that singletons were included. In both cases, singletons show extensive frequency differentiation that is not reproducible in non-singletons.

E.7 Imputation accuracy of TCC→TTC mutations

Each genotype call in the 1000 Genomes Phase I data is associated with an RSQ quality score, the estimated correlation coefficient between true and imputed genotypes at a given locus (Li et al. 2010b). To assess the effect of imputation error on mutation frequency differences between 1000 Genomes populations, I repeated the volcano plot analysis from Figure 6.2.1 on medium imputation quality SNPs (RSQ between 0.8 and 0.95) as well as low imputation quality SNPs (RSQ less than 0.8). As shown in Figure E.8, excess TCC→TTC mutations in Europe are evident across all imputation quality ranges. However, the medium-quality and low-quality volcano plots show many more minor outliers, mutation type frequency differences between Europe and Africa or Asia and Africa that are not reproducible in the high-quality 1000 Genomes SNPs or Complete Genomics data. These minor outliers might be indicative of real mutation rate change that occurred very recently and affects only alleles of very low frequency, too low-frequency to appear in the Complete Genomics dataset or to have high average imputation quality. However, these outliers might also be bioinformatic artifacts, particularly the CpG outliers that appear susceptible to ancestral misidentification errors.

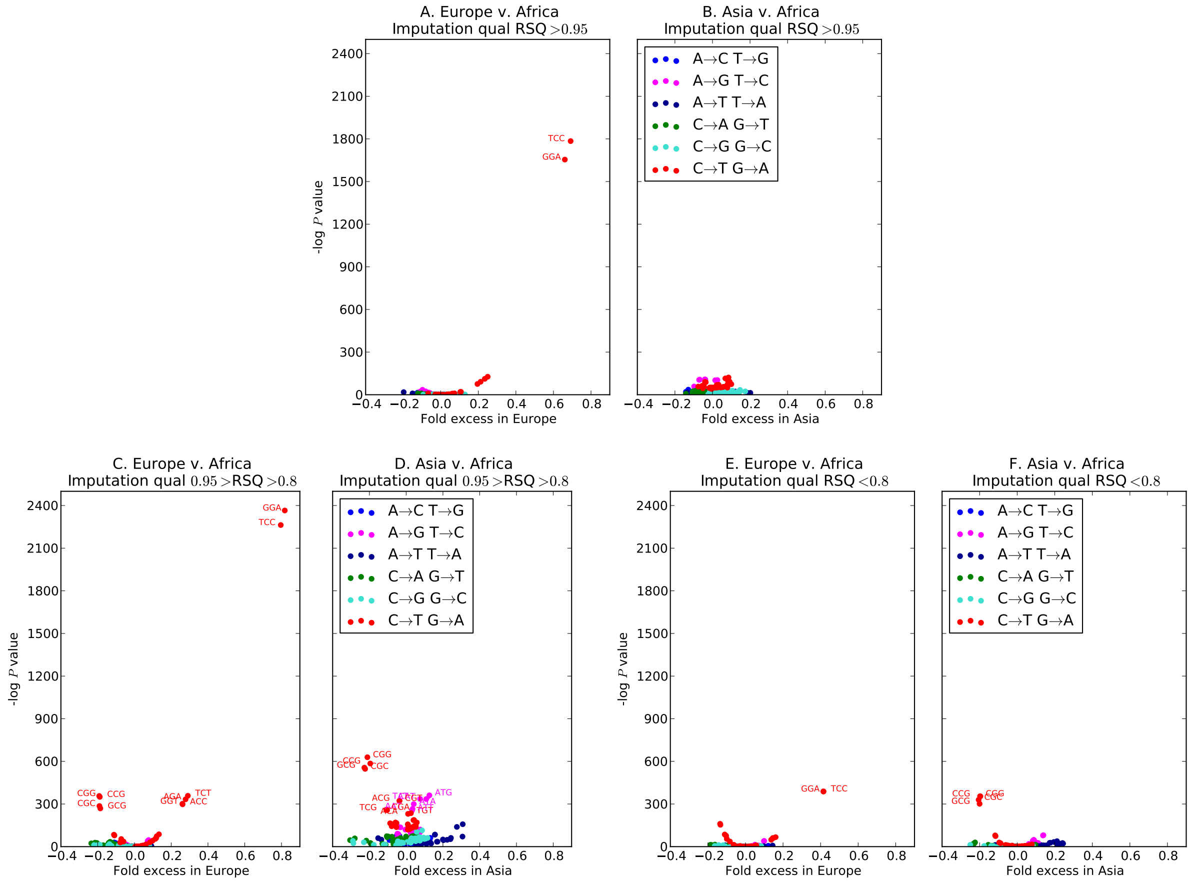


Figure E.8: Panels A and B show differently scaled versions of Figure 6.2.1A and 6.2.1B from the main text, illustrating that high imputation quality SNPs (RSQ > 0.95) support little frequency differentiation other than a C→T transition excess in Europe compared to Africa. In contrast, Panels C and D show volcano plots generated in the same way except that medium-imputation-quality SNPs (RSQ between 0.8 and 0.95) were used instead of high-quality SNPs. These medium-quality SNPs reveal an even clearer excess of various C→T transitions in Europe, but they also suggest evidence of other frequency differences between Europe and Africa and between Asia and Africa that are not reproducible in higher-quality data. Finally, Panels E and F reproduce the same analysis using only low-imputation-quality SNPs (RSQ < 0.8). Again, the low-quality SNPs show evidence of differences that are not reproducible in the highest-quality data. Further work will be required to assess whether these differences are real or artifactual.

E.8 Comparison to somatic mutations in melanoma

In 2013, Alexandrov et al. (2013) introduced the concept of cancer mutational signatures: collections of mutation types that are each characteristic of one or more cancer types and sometimes associated with exposure to a known carcinogen. They discovered that melanoma has a unique mutational signature composed almost entirely of C→T transitions, almost 30% of which are TCC→TTC mutations. I downloaded the mutational spectrum of a melanoma skin cancer described by Pleasance et al. (2010) to make a more detailed comparison between its mutational signature and the differences between PE and PAf. As shown in the scatterplot below, CCC→CTC mutations and TCT→TTT mutations are candidates for mutation rate acceleration in Europe that also contribute substantially to both the spectrum of melanoma (Figure E.9). To a lesser extent, the same is true of ACC→ATC and TCA→TTA. However, the correlation between melanoma and European mutation rate change is far from perfect overall.

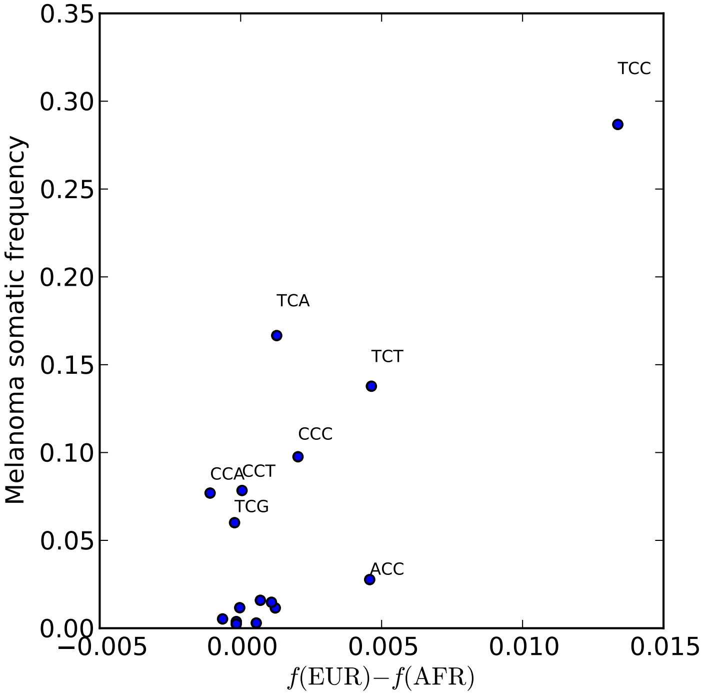


Figure E.9: Each dot in this scatterplot represents a different type of C→T transition, merged with its G→A strand complement for simplicity. The y coordinate of the dot representing mutation type m is the frequency of m in melanoma, while the x coordinate is the difference between the frequency of m in PE and the frequency of m in PAf.

Bibliography

1000 Genomes Project. 2010, Nature, 467, 1061

—. 2012, Nature, 491, 56

Abramovitz, M., & Stegun, I. 1964, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (New York: Dover)

Alexandrov, L., Nik-Zainal, S., Wedge, D., et al. 2013, Nature, 500, 415

Allen, A., Gillooly, J., Savage, V., & Brown, J. 2006, Proc Natl Acad Sci USA, 103, 9130

Awadalla, P., Gauthier, J., Myers, R., et al. 2010, American Journal of Human Genetics, 87, 316

Bakos, L., Masiero, N., Bakos, R., et al. 2009, JEADV, 23, 304

Barton, N. June 28, 2012, Personal communication

Beaumont, M., Zhang, W., & Balding, D. 2002, Genetics, 192, 2025

Beerli, P., & Felsenstein, J. 2001, Proc Natl Acad Sci USA, 98, 4563

Beleza, S., Santos, A., McEvoy, B., et al. 2013, Mol Biol Evol, 30, 24

Bemark, M., Khamlichi, A., Davies, S., & Neuberger, M. 2000, Curr Biol, 10, 1213

Blount, B., Mack, M., Wehr, C., et al. 1997, Proc Natl Acad Sci USA, 94, 3290

Branda, R., & Eaton, J. 1978, Science, 201, 625

Brash, D., Seetharam, S., Kraemer, K., Seidman, M., & Bredberg, A. 1987, Proc Natl Acad Sci USA, 84, 3782

Browning, B., & Browning, S. 2011, Am J Hum Gen, 88, 173

Browning, S., & Browning, B. 2009, Am J Hum Gen, 84, 210

Cahill, J. A., Green, R. E., Fulton, T. L., et al. 2013, PLoS Genetics, 9, e1003345

Chan, A. H., Jenkins, P. A., & Song, Y. S. 2012, PLoS Genet., 8, e1003090

Charlesworth, D. 2006, PLoS Genetics, 2, e64

Charlesworth, D., Charlesworth, B., & Morgan, M. 1995, Genetics, 141, 1619

Chen, J., Ferec, C., & Cooper, D. 2009, Hum Mutat, 30, 1435

Coventry, A., Bull-Otterson, L. M., Liu, X., et al. 2010, Nat Commun, 1, 131

Cox, M., Woerner, A., Wall, J., & Hammer, M. 2008, BMC Genetics, 9, 1471

Crawford, D. C., Bhangale, T., Li, N., et al. 2004, Nature Genetics, 36, 700

Cress, R., & Holly, E. 1997, Cancer Causes and Control, 8, 246

Crombie, I. 1979, Br J Cancer, 40, 185

Daly, J., Bebenek, K., Watt, D., et al. 2012, The Journal of Immunology, 188, 5528

Danacek, P., Auton, A., Abecasis, G., Albers, C., & Banks, E. 2011, Bioinformatics, 27, 2156

De Iorio, M., & Griffiths, R. C. 2004a, Adv. in Appl. Probab., 36, 417

—. 2004b, Adv. in Appl. Probab., 36, 434

Denver, D., Morris, K., Lynch, M., & Thomas, W. 2004, Nature, 430, 679

Denver, D., Dolan, P., Wilhelm, L., et al. 2009, Proc Natl Acad Sci USA, 106, 16310

Drake, J., Bebenek, A., Kissling, G., & Peddada, S. 2005, Proc Natl Acad Sci USA, 102, 12849

Drmanac, R., Sparks, A., Callow, M., et al. 2010, Science, 327, 78

Drobetsky, E., Grosovsky, A., & Glickman, B. 1987, Proc Natl Acad Sci USA, 84, 9103

Drobetsky, E., & Sage, E. 1993, Mut Res, 289, 131

Dutheil, J. Y., Ganapathy, G., Hobolth, A., et al. 2009, Genetics, 183, 259

Esposito, G., Godin, I., Klein, U., et al. 2000, Curr Biol, 10, 1221

Fearnhead, P., & Smith, N. G. C. 2005, 77, 781

Flanagan, M. T. 2010, www.ee.ucl.ac.uk/mflanaga

Fraser, H. 2013, Genome Res, 23, 1089

Gan, G., Wittschieben, J., Wittschieben, B., & Wood, R. 2008, Cell Res, 18, 174

Gillooly, J., Allen, A., West, G., & Brown, J. 2005, Proc Natl Acad Sci USA, 102, 140

Goodman, M. 1961, Human Biol, 33, 131

—. 2002, Annu Rev Biochem, 71, 17

Gravel, S. 2012, Genetics, 191, 607

Gravel, S., Henn, B., Gutenkunst, R., et al. 2011, Proc Natl Acad Sci USA, 108, 11983

Green, P., Ewing, B., Miller, W., et al. 2003, Nature Genetics, 514

Green, R., Krause, J., Briggs, A., et al. 2010, Science, 328, 710

Griffiths, R., & Tavaré, S. 1994, Phil. Trans. R. Soc. Lond. B, 344, 403

Griffiths, R. C., Jenkins, P. A., & Song, Y. S. 2008, Adv. in Appl. Probab., 40, 473

Griffiths, R. C., & Tavaré, S. 1994a, Philos. Trans. R. Soc. Lond. B Biol. Sci., 344, 403

—. 1994b, Theor. Popul. Biol., 46, 131

Gronau, I., Hubisz, M., Gulko, B., Danko, C., & Siepel, A. 2011, Nature Genetics, 43, 1031

Gusev, A., Lowe, J., Stoffel, M., et al. 2009, Genome Res, 19, 318

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. 2009, PLoS Genet, 5, e1000695

Haddrill, P. R., Thornton, K. R., Charlesworth, B., & Andolfatto, P. 2005, Genome Res., 15, 790

Hailer, F., Kutschera, V. E., Hallstrom, B. M., et al. 2012, Science, 336, 344

Harris, K., & Nielsen, R. 2013, PLoS Genetics, 9, e1003521

Hayes, B., Visscher, P., McPartlan, H., & Goddard, M. 2003, Genome Res, 13, 635

Hellenthal, G., Busby, G., Band, G., et al. 2014, Science, 343, 747

Hill, W., & Robertson, A. 1968, Theoretical and Applied Genetics, 38, 226

Hobolth, A., Christensen, O., Mailund, T., & Schierup, M. 2007, PLoS Genetics, 3, e7

Hodgkinson, A., & Eyre-Walker, A. 2010, Genetics, 184, 233

Hodgkinson, A., Ladoukakis, E., & Eyre-Walker, A. 2009, PLoS Biology, 7, e1000027

Hu, D., Yu, G., McCormick, S., Schneider, S., & Finger, P. 2005, Am J Ophthalmology, 140,

612.e1
Hudson, R. 2002, Bioinformatics, 18, 337
Hwang, D., & Green, P. 2004, Proc Natl Acad Sci USA, 101, 13994
Jablonski, N., & Chaplin, G. 2000, J Hum Evol, 39, 57
—. 2010, Proc Natl Acad Sci USA, 107, 8962
Keightley, P., Trivedi, U., Thomson, M., et al. 2009, Gen Res, 19, 1195
Keinan, A., & Clark, A. G. 2012, Science, 336, 740
Kimura, M. 1980, J Mol Evol, 16, 111
Kingman, J. 1982, Stochastic processes and their applications, 13, 235
Kong, A., Gudbjartsson, D., Sainz, J., et al. 2002, Nature, 31, 241
Kong, A., Frigge, M., Masson, G., et al. 2012, Nature, 488, 471
Kuhner, M., Yamato, J., & Felsenstein, J. 1995, Genetics, 140, 1421
Lange, S., Takata, K., & Wood, R. 2011, Nature Rev Cancer, 11, 96
Lazaridis, I., Patterson, N., Mittnik, A., et al. 2014, Nature, 513, 409
Leffler, E., Gao, Z., Pfeifer, S., et al. 2013, Science, 29, 1578
Levy, S., Sutton, G., Ng, P., et al. 2007, PLoS Biology, 5, e254
Li, H., & Durbin, R. 2011, Nature, 475, 493
Li, N., & Stephens, M. 2003, Genetics, 165, 2213
Li, W., & Tanimura, M. 1987, Nature, 326, 93
Li, Y., Willer, C., Ding, J., Scheet, P., & Abecasis, G. 2010a, Gen Epidem, 34, 816
Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. 2010b, Genetic Epidemiology, 34, 816
Lohmueller, K., Albrechtsen, A., Li, Y., et al. 2011, PLoS Genetics, 7, e1002326
Lynch, M., Sung, W., Morris, K., et al. 2008, Proc Natl Acad Sci USA, 105, 9272
MacLeod, I., Meuwissen, T., Hayes, B., & Goddard, M. 2009, Genet Res, 91, 413
Mailund, T., Dutheil, J. Y., Hobolth, A., et al. 2011, PLoS Genetics, 7, e1001319
Marionnet, C., Benoit, A., Benhamou, S., Sarasin, A., & Stary, A. 1995, J Mol Biol, 252, 550
Marjoram, P., & Wall, J. 2006, BMC Genetics, 7, 16
McDonald, M., Wang, W., Huang, H., & Leu, J. 2011, PLoS Biology, 9, e1000622
McVean, G., & Cardin, N. 2005, Phil Trans Royal Soc B, 360, 1387
McVean, G. A. T., Myers, S. R., Hunt, S., et al. 2004, Science, 304, 581
McVicker, G., Gordon, D., Davis, C., & Green, P. 2009, PLoS Genetics, 5, e1000471
Meyer, M., et al. 2012, Science, 338, 222
Miller, W., Schuster, S., Welch, A., et al. 2012, Proc Natl Acad Sci USA, 109, E2382
Moltke, I., Albrechtsen, A., Hansen, T., Nielsen, F., & Nielsen, R. 2011, Genome Res, 21, 1168
Nachman, M., & Crowell, S. 2000, Genetics, 156, 297
Nelson, M. R., Wegmann, D., Ehm, M. G., et al. 2012, Science, 337, 100
Nielsen, R. 1997, Genetics, 146, 711
—. 1998, Theor Pop Biol, 53, 143
—. 2000, Genetics, 154, 931

Nielsen, R., & Wakeley, J. 2001, Genetics, 158, 885

Nielsen, R., & Wiuf, C. 5–12 April 2005, in ISI Conference Proceedings, Sydney, Australia

Noonan, J., Coop, G., Kudarvalli, S., et al. 2006, Science, 314, 1113

Northam, M., Robinson, H., Kochenova, O., & Scherbakova, P. 2010, Genetics, 184, 27

Off, M., Steindal, A., Porojnicu, A., et al. 2005, J Photochem Photobiol B, 82, 47

Ogawara, D., Muroya, T., Yamauchi, K., et al. 2010, DNA Repair, 9, 90

Olalde, I., Allentoft, M., Sánchez-Quinto, F., et al. 2014, Nature, 507, 225

Ossowski, S., Schneeberger, K., Locas-Liedo, J., et al. 2010, Science, 327, 92

Palamara, P., Lencz, T., Darvasi, A., & Pe'er, I. 2012, Am J Hum Gen, 91, 809

Paul, J. S., & Song, Y. S. 2010, Genetics, 186, 321

—. 2012, Bioinformatics, 28, 2008

Paul, J. S., Steinrücken, M., & Song, Y. S. 2011, Genetics, 187, 1115

Pickerell, J., Coop, G., Novembre, J., et al. 2009, Genome Res, 19, 826

Pleasance, E., Cheetham, R., Stephens, P., et al. 2010, Nature, 463, 191

Pool, J., & Nielsen, R. 2009, Genetics, 181, 711

Press, W., Teukolsky, S., Vetterling, W., & Flannery, B. 2007, Numerical Recipes: The Art of Scientific Computing, 3rd edn. (Cambridge University Press)

Pritchard, J. 2011, Nature Genetics, 43, 923

Pritchard, J., Seielstad, M., Perez-Lezun, A., & Feldman, M. 1999, Mol Biol Evol, 16, 1791

Prüfer, K., Racimo, F., Patterson, N., et al. 2014, Nature, 505, 43

Purcell, S., Neale, B., Todd-Brown, K., et al. 2007, Am J Hum Gen, 81, 559

Ralph, P., & Coop, G. 2013, PLoS Biology, 11, e1001555

Rasmussen, M., Hubisz, M., Gronau, I., & Siepel, A. 2014, PLoS Genetics, 10, e1004342

Rasmussen, M., Guo, X., Wang, Y., et al. 2011, Science, 334, 94

Roach, J., Glusman, G., Smit, A., et al. 2010, Science, 328, 636

Roberts, S., Sterling, J., Thompson, C., et al. 2012, Mol Cell, 46, 424

Sabeti, P., Reich, D., Higgins, J., et al. 2002, Nature, 419, 832

Sakamoto, A., Stone, J., Kissling, G., et al. 2007, DNA Repair, 6, 1829

Sankararaman, S., Patterson, N., Li, H., Pääblo, S., & Reich, D. 2012, PLoS Genetics, 8, e1002947

Saribasak, H., Maul, R., Cao, Z., et al. 2012, J Exp Med, 209, 1075

Scally, A., & Durbin, R. 2012, Nature Rev Genetics, 13, 745

Schaffner, S., Foo, C., Gabriel, S., et al. 2005, Genome Res, 15, 1576

Schierup, M., & Hein, J. 2000, Genetics, 156, 879

Schrider, D., Houle, D., Lynch, M., & Hahn, M. 2013, Genetics, 194, 937

Schrider, D., Hourmozdi, J., & Hahn, M. 2011, Curr Biol, 21, 1051

Ségurel, L., Wyman, M., & Przeworski, M. 2014, Annu Rev Genomics Hum Genet, 15, 19.1

Ségurel, L., Thompson, E., Flutre, T., et al. 2012, Proc Natl Acad Sci USA, 109, 18493

Seplyarskiy, V. B., Bazykin, G. A., & Soldatov, R. A. 2014, bioRxiv

Sheehan, S., Harris, K., & Song, Y. 2013, Genetics, 194, 647

Skandalis, A., Ford, B., & Glickman, B. 1994, Mutation Research/DNA Repair, 314, 21

Slatkin, M., & Hudson, R. 1991, Genetics, 129, 555

Slatkin, M., & Madison, W. 1989, Genetics, 123, 603

Steinrücken, M., Paul, J., & Song, Y. 2012, Theor Popul Biol, doi:10.1016/j.tpb.2012.08.004

Steinrücken, M., Paul, J. S., & Song, Y. S. 2013, Theor. Popul. Biol., 87, 51

Stone, J., Lujan, S., & Kunkel, T. 2012, Environmental and Molecular Mutagenesis, 53, 777

Stover, P. 2009, J Nutr, 139, 2402

Strasburg, J., & Rieseberg, L. 2010, Mol Biol Evol, 27, 297

Tajima, F. 1983, Genetics, 105, 437

Tavaré, S. 1984, Theoretical Population Biology, 26, 119

Tavaré, S., Balding, D., Griffiths, R., & Donnelly, P. 1997, Genetics, 505

Templeton, A. 2002, Nature, 416, 45

Terekhanova, N., Bazykin, G., Neverov, A., Kondrashov, A., & Seplyarsky, V. 2013, Mol Biol Evol, 30, 1315

Thornton, K., & Andolfatto, P. 2006, Genetics, 172, 1607

Venn, O., Turner, I., Mathieson, I., et al. 2014, Science, 13, 1272

Wakeley, J., & Hey, J. 1997, Genetics, 145, 847

Wallock, L., Tamura, T., Mayr, C., et al. 2001, Fertility and Sterility, 75, 252

Wang, Y., & Hey, J. 2010, Genetics, 184, 363

Waters, L., Minesinger, B., Wiltrout, M., et al. 2009, Microbiol and Mol Biol Rev, 73, 134

Waters, L., & Walker, G. 2006, Proc Natl Acad Sci USA, 103, 8971

Watterson, G. 1975, Theor Pop Biol, 7, 256

Williamson, S., Hernandez, R., Fledel-Alon, A., et al. 2005, Proc Natl Acad Sci USA, 102, 7882

Wittschieben, J., Shivji, M., Lalani, E., et al. 2000, Curr Biol, 10, 1217

Wiuf, C. 2006, Math Biol, 53, 821

Wiuf, C., & Hein, J. 1999, Theor Popul Biol, 55, 248

Wright, S., Keeling, J., & Gillman, L. 2006, Proc Natl Acad Sci USA, 103, 7718

Yang, Z., & Rannala, B. 1997, Mol Biol Evol, 14, 717