OPEN ACCESS Freely available online

PLOS | GENETICS

Inferring Demographic History from a Spectrum of Shared Haplotype Lengths

Kelley Harris[1]*, **Rasmus Nielsen**[2,3,4]

1 Department of Mathematics, University of California Berkeley, Berkeley, California, United States of America, **2** Department of Integrative Biology, University of California Berkeley, Berkeley, California, United States of America, **3** Department of Statistics, University of California Berkeley, Berkeley, California, United States of America, **4** Center for Bioinformatics, University of Copenhagen, Copenhagen, Denmark

Abstract

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. A central challenge is to estimate the timing of past admixture and divergence events, for example the time at which Neanderthals exchanged genetic material with humans and 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 coalescent approximation. Recent gene flow between populations leaves behind long tracts of identity by descent (IBD), and these tracts give our method 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 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 African-European 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 less population structure.

Citation: Harris K, Nielsen R (2013) Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. PLoS Genet 9(6): e1003521. doi:10.1371/journal.pgen.1003521

Editor: Jeffery D. Jensen, Ecole Polytechnique Federale de Lausanne, Switzerland

Received October 11, 2012; **Accepted** April 6, 2013; **Published** June 6, 2013

Funding: KH was supported by a U.C. Berkeley Regents' Fellowship and a NSF Graduate Research Fellowship. RN received support from NIH grant 2R14003229-07. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kharris@math.berkeley.edu

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 [1,2] 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 [3,4] and polymorphic sites [5] 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 [6–11]. With the availability of likelihood methods, complex models including both gene flow and population divergence [12], 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 [12–14], 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 [15–16].

A parallel advancement to likelihood methods has been the production 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 [17] and the antiquity of the Aboriginal Australian population [18].

Motivated by the limitations of full likelihood methods and the power of large datasets, there is great interest in developing

Author Summary

In this paper, we study the length distribution of tracts of identity by state (IBS), which are the gaps between pairwise differences in an alignment of two DNA sequences. These tract lengths contain information about the amount of genetic diversity that existed at various times in the history of a species and can therefore be used to estimate past population sizes. IBS tracts shared between DNA sequences from different populations also contain information about population divergence and past gene flow. By looking at IBS tracts shared within Africans and Europeans, as well as between the two groups, we infer that the two groups diverged in a complex way over more than 40,000 years, exchanging DNA as recently as 12,000 years ago. Besides having anthropological importance, the history we infer predicts the distribution of pairwise differences between humans extremely accurately at a fine-scale level, which may aid future scans for natural selection in the genome. Despite our current focus on human data, the method is general enough to use on other organisms and has the potential to shed light on the demography of many more species.

scalable approximate methods for population genetic inference across many recombining loci. One popular strategy is approximate Bayesian computation (ABC) [19–21], 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 [22] or by numerically solving the Wright-Fisher diffusion equation [23,24].

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 [25,26], 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) [27–30]. 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 [31,32]. 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 [28], 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 [33,34].

Linkage information can be especially revealing about recent demographic history and recent common ancestry. Many HMM-based methods have been devised to identify long haplotype tracts inherited *identical by descent* (IBD) from a single common ancestor without recombination [35–38], and downstream analyses can harness IBD blocks to infer recent demographic events [39–42]. Of particular interest are *migrant tracts* that were inherited IBD between individuals from different populations as a result of recent migration [42–44]; Gravel used migrant tracts to show that at least two migration "pulses" are needed to account for tracts admixed from Europe into African Americans [44]. In addition to migrant tracts, allele frequency correlations over long genetic distances (>0.1 cM) have been used to study recent gene flow between European populations [45].

It is a challenging problem to infer recent and ancient demography within a unified theoretical framework, bridging the time gap between IBD-based accounts of recent demography and the various methods that interpret older demographic signals. To this end, we present an analytic method that draws power from linked sites over a wide range of genomic length scales, not just short blocks where linkage is strong or long sequences inherited from recent common ancestors. Specifically, we study the set of distances between neighboring SNPs in a sample of two haplotypes. The distance between adjacent polymorphisms is inversely correlated with local TMRCA; an L-base-long 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 of each resulting tract of *identity by state* (IBS); for every L, we obtain the total abundance of L-base-long IBS tracts, where an L-base IBS tract is defined to be L contiguous identical base pairs bracketed by SNPs on the left and right (see Figure 1).

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 the expected distribution of IBS tract lengths, 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 [32,46], 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. Our source code is available for download at https://github.com/kelleyharris/Inferring-demography-from-IBS.

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 and 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 datasets by subsampling all haplotype pairs and regarding them as independent. Given

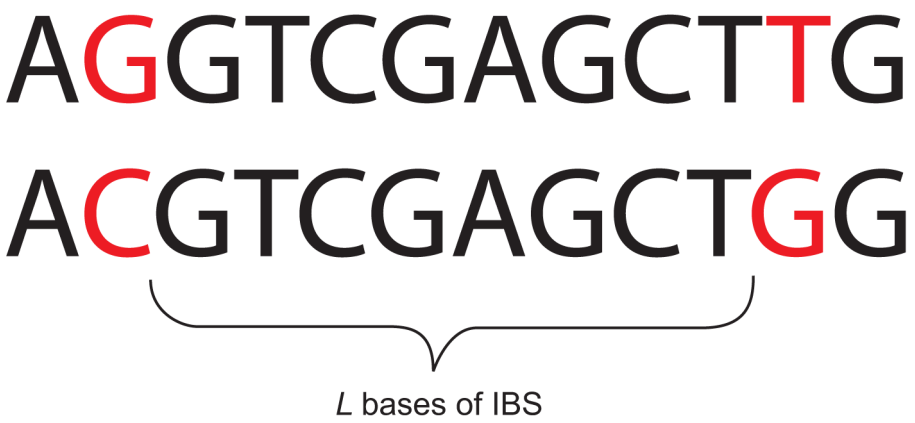

Figure 1. An eight base-pair tract of identity by state (IBS).
doi:10.1371/journal.pgen.1003521.g001

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 [47]. 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.* use a likelihood analysis of many short loci 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 [14]. The older Schaffner, *et al.* demographic model contains no recent European-African gene flow either [48], but Gutenkunst, *et al.* and Gravel, *et al.* use SFS data to infer divergence times and gene flow levels that are intermediate between these two extremes [22,49]. We aim to contribute to this discourse by using IBS tract lengths to study the same class of complex demographic models employed by Gutenkunst, *et al.* and Gronau, *et al.*, models that have only been previously used to study allele frequencies and short haplotypes that are assumed not to recombine. Our method is the first to use these models in conjunction with haplotype-sharing information similar to what is used by the PSMC and other coalescent HMMs, fitting complex, high-resolution demographic models to an equally high-resolution summary of genetic data.

Results

An accurate analytic IBS tract length distribution

In the methods section, we derive a formula for the expected length distribution of IBS tracts shared between two DNA sequences from the same population, as well as the length distribution of tracts shared between sequences from diverging populations. Our formula approximates the distribution expected under the SMC' model of Marjoram and Wall [46], which in turn approximates the coalescent with recombination. We evaluate the accuracy of the approximation by simulating data under the full coalescent with recombination and comparing the results to our analytical predictions. In general, we find that the approximations are very accurate as illustrated for two example histories in Figures 2 and 3. To create each plot in Figure 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 2 of Text S1). Figure 3 was generated by simulating population bottlenecks of varying duration and intensity. In both of these scenarios the analytical approximations closely follow the distributions obtained from full coalescent simulations.

If we wish to infer demography from IBS tract lengths, the following must be true: 1) IBS tract length distributions 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 and 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, two population bottlenecks with the same strength-to-duration ratio appear distinguishable if their population sizes differ by at least a factor of two during the bottleneck. As expected, longer IBS tracts are shared between populations that exchanged DNA more recently, suggesting that IBS tracts are highly informative about past admixture times and motivating the development of a statistical demographic inference method.

Estimates from simulated data

Inferring simulated population histories. Figures 2 and 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 [50].

To assess the power and accuracy of the method, we simulated 100 replicate datasets for each of two histories with different admixture times. From each dataset, we jointly inferred four parameters: admixture time, split time, admixture fraction, and effective population size. We obtained estimates that are extremely accurate and low-variance (see Table 1); supplementary Figures S1 and S2 show the full distributions of estimated parameter values.

Comparison to ∂a∂i. We compared the new method to the method implemented in ∂a∂i, which can evaluate demographic

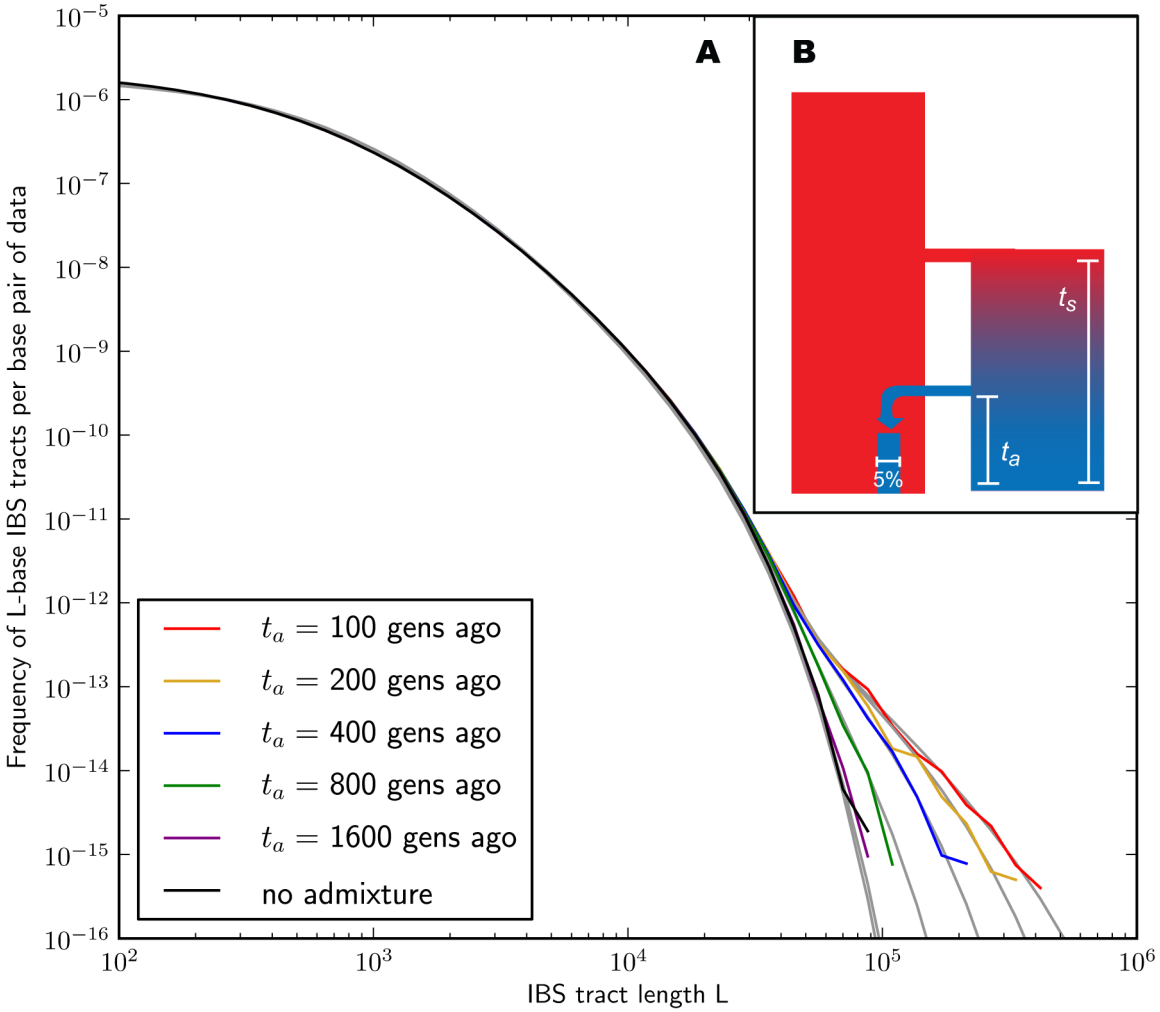

Figure 2. Spectra of IBS sharing between simulated populations that differ only in admixture time. Each of the colored tract spectra in Figure 2A was generated from 4.8×10^{10} base pairs of sequence alignment simulated with Hudson's MS [68]. The IBS tracts are shared between two populations of constant 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 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 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 $\{\lfloor 1.25^n \rfloor\}_{n \geq 1}$.
doi:10.1371/journal.pgen.1003521.g002

scenarios with the same parameterization as ours, focusing on the simple admixture history summarized in Table 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 of the IBS tract likelihood were almost always successful, converging to the global optimum, but optimizations performed using default ∂a∂i settings often terminated near random initial starting points (see Section 4.1 of Text S1 and Table S1). This suggests that the analytic 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 the numerical solution of partial differential equations.

For a simple four-parameter history, it is feasible to identify maximum-likelihood parameters through a grid search that is robust to minor numerical instabilities. Using this type of optimization strategy, both methods provide similar results (see Supplementary Figure S3). 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 changes more steeply along the divergence time axis.

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 who are Utah residents of central European descent (CEU) and another family recruited from the Yorubans of Ibadan, Nigeria (YRI).

We recorded the spectrum of IBS tracts shared between each pair sampled from the eight parental haplotypes, which were

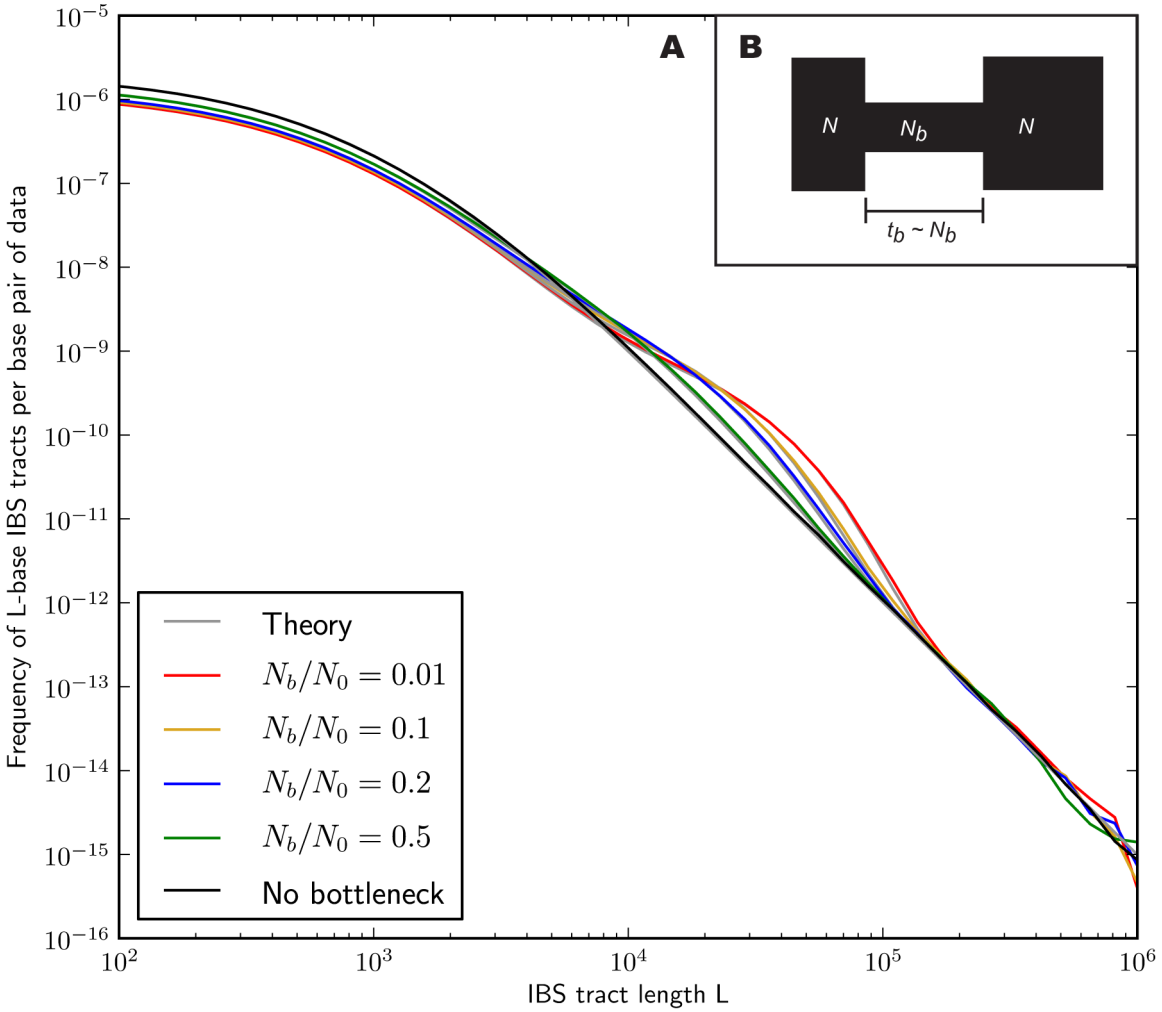

Figure 3. Shared IBS tracts within bottlenecked populations. As in Figure 2, each colored spectrum in Figure 3A was generated by using MS to simulate 4.8×10^{10} base pairs of pairwise alignment. Both sequences are derived from the population depicted in Figure 3B that underwent a bottleneck from size $N_0 = 10{,}000$ 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 predict analytically for each simulated history.
doi:10.1371/journal.pgen.1003521.g003

sequenced at 20–60x coverage and phased with the help of the children by the 1000 Genomes consortium [51]. As expected, we observe 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 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 [51]. 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 5). In section 3.2 of Text S1 and Figure S4, 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 create 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 [52,53]. Our theoretical framework would make it possible 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 [53] to calculate the average recombination rate across all sites that are part of L-base IBS

Table 1. Inferring the parameters of a simple admixture scenario.

	τ_a (gens)	τ_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 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 τ_a, τ_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.
doi:10.1371/journal.pgen.1003521.t001

tracts. The results, plotted in Figure 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 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 accurately infer past population sizes from data with simulated recombination hotspots.

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 are dispersed fairly evenly throughout the genome, but human-chimp fixed differences cover more than 10% of the sites they span (see Figure 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 [52]. 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).

Human demography and the migration out of Africa

Previously published models of human demography. After generating spectra of empirical IBS tract sharing in the 1000 Genomes trios, we simulated IBS tract data under several

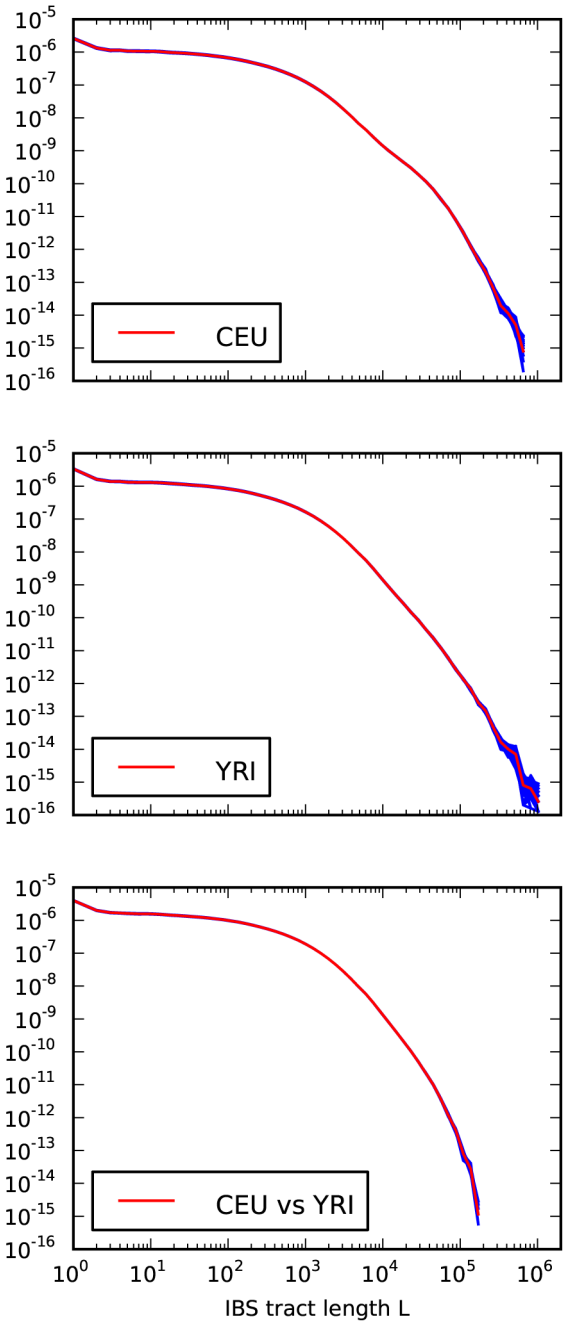

Figure 4. Frequencies of IBS tracts shared between the 1000 Genomes trio parental haplotypes. Each plot records the number of L-base IBS tracts observed per base pair of sequence alignment. The red spectrum records tract frequencies compiled from the entire alignment, while the blue spectra result from 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 Europeans.
doi:10.1371/journal.pgen.1003521.g004

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.*; 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 [24]; the other was fit by

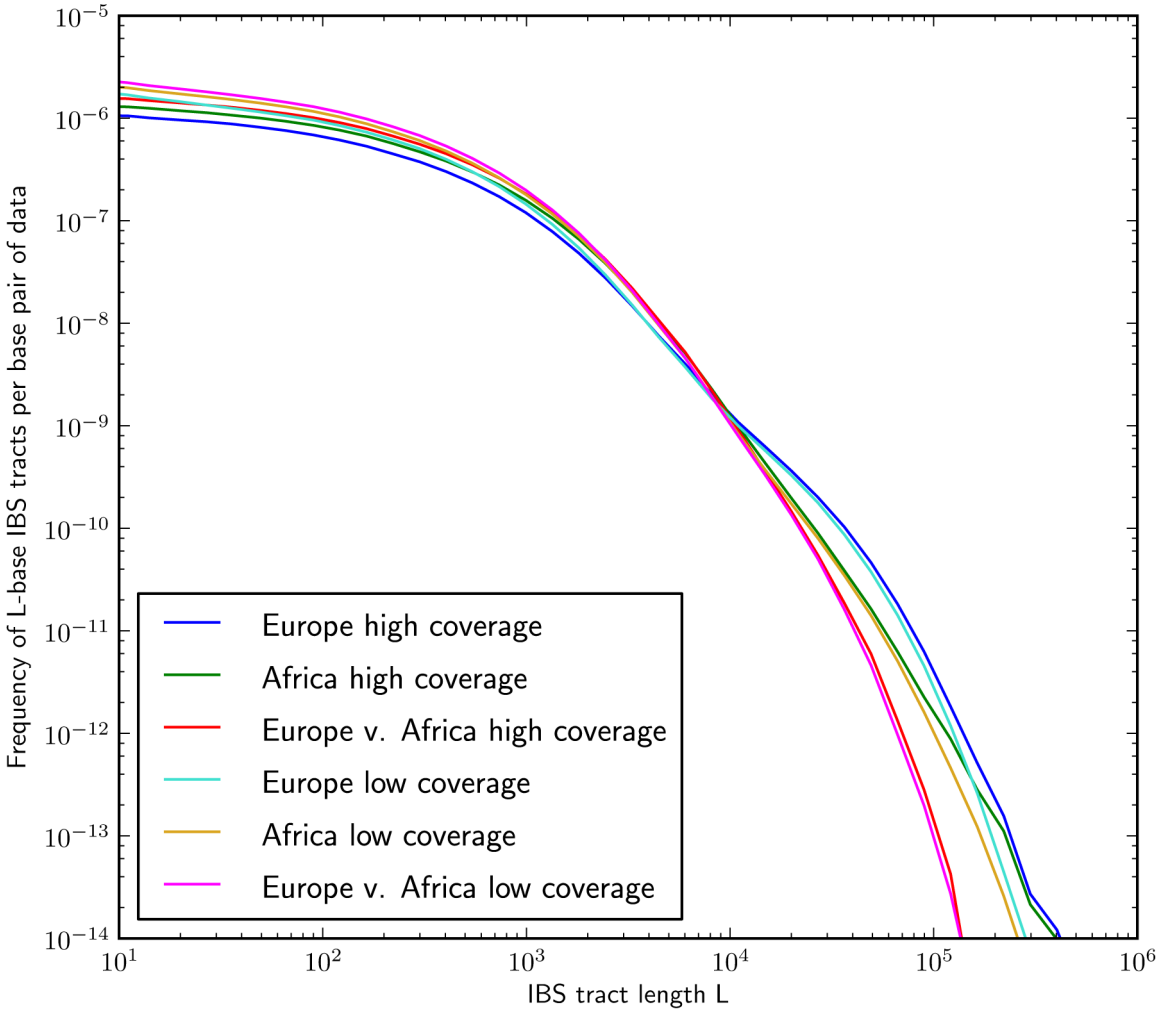

Figure 5. IBS tract lengths in the 1000 Genomes pilot data: trios v. low coverage. These IBS tract spectra were generated from pairwise alignments of the 1000 Genomes high coverage trio parental haplotypes and the CEU (European) and YRI (Yoruban) low coverage haplotypes, aligning samples within each 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.
doi:10.1371/journal.pgen.1003521.g005

Gravel, *et al.* to a SFS of the 1000 Genomes low coverage data that was corrected for low coverage sampling bias [9]. 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 S5). 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 S6), 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 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.

We fit our model to the data using a Poisson composite likelihood approach; maximum likelihood parameters are listed in

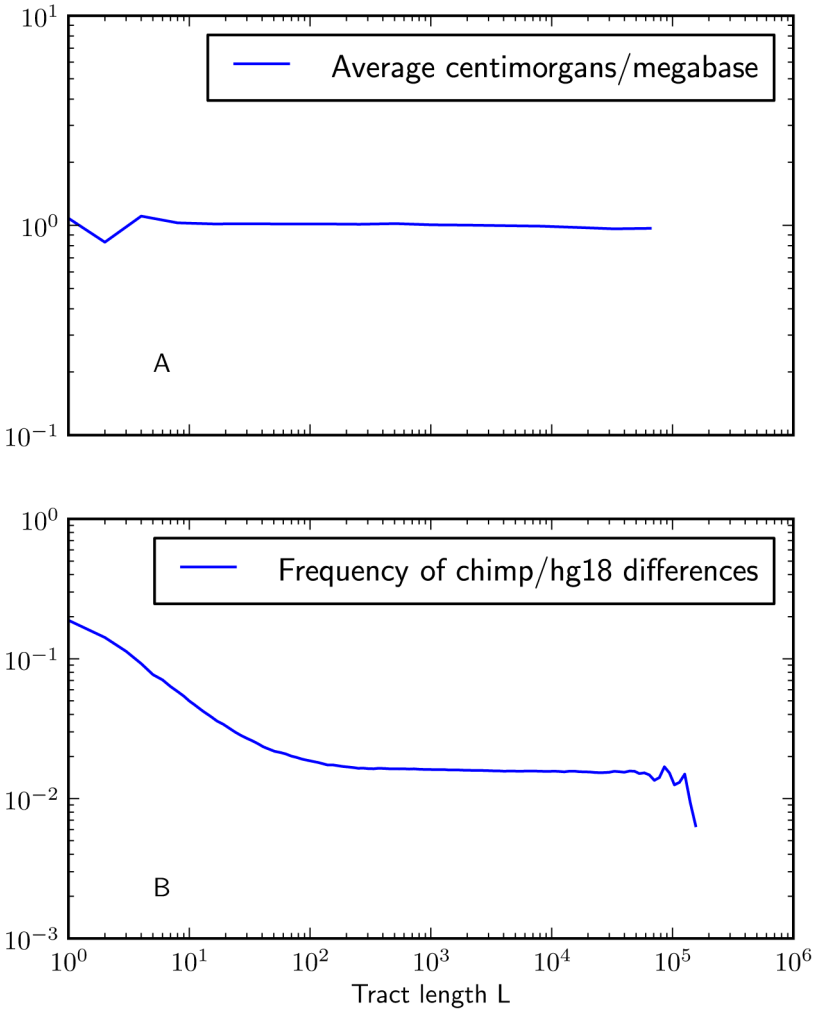

Figure 6. Mutation and recombination rates within L-base IBS tracts. Figure 6A shows that there is no length class of IBS tracts with a significantly higher or lower mutation rate than the genome-wide average (recombination rates are taken from the deCODE genetic map [53]). In contrast, Figure 6B shows 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 Europeans and Africans, but the results are similar for IBS sharing within each of the populations.
doi:10.1371/journal.pgen.1003521.g006

Table 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 8).

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 Figure 4). These inference results were consistent and low-variance. In addition, we used Hudson's MS to simulate 30 datasets under the inferred demographic history, then estimated demographic parameters from each simulated dataset (see Section 3.3 of Text S1 for the command line used to generate the data). This parametric bootstrapping revealed some modest parameter estimate biases, though there were no qualitative differences between the histories inferred from replicate simulations and the histories inferred from real data (see Section 3.4 of Text S1 and Figures S7, S8 and S9 for the parameter distributions inferred from simulated data). Supplementary Figure S10 compares the schematic 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

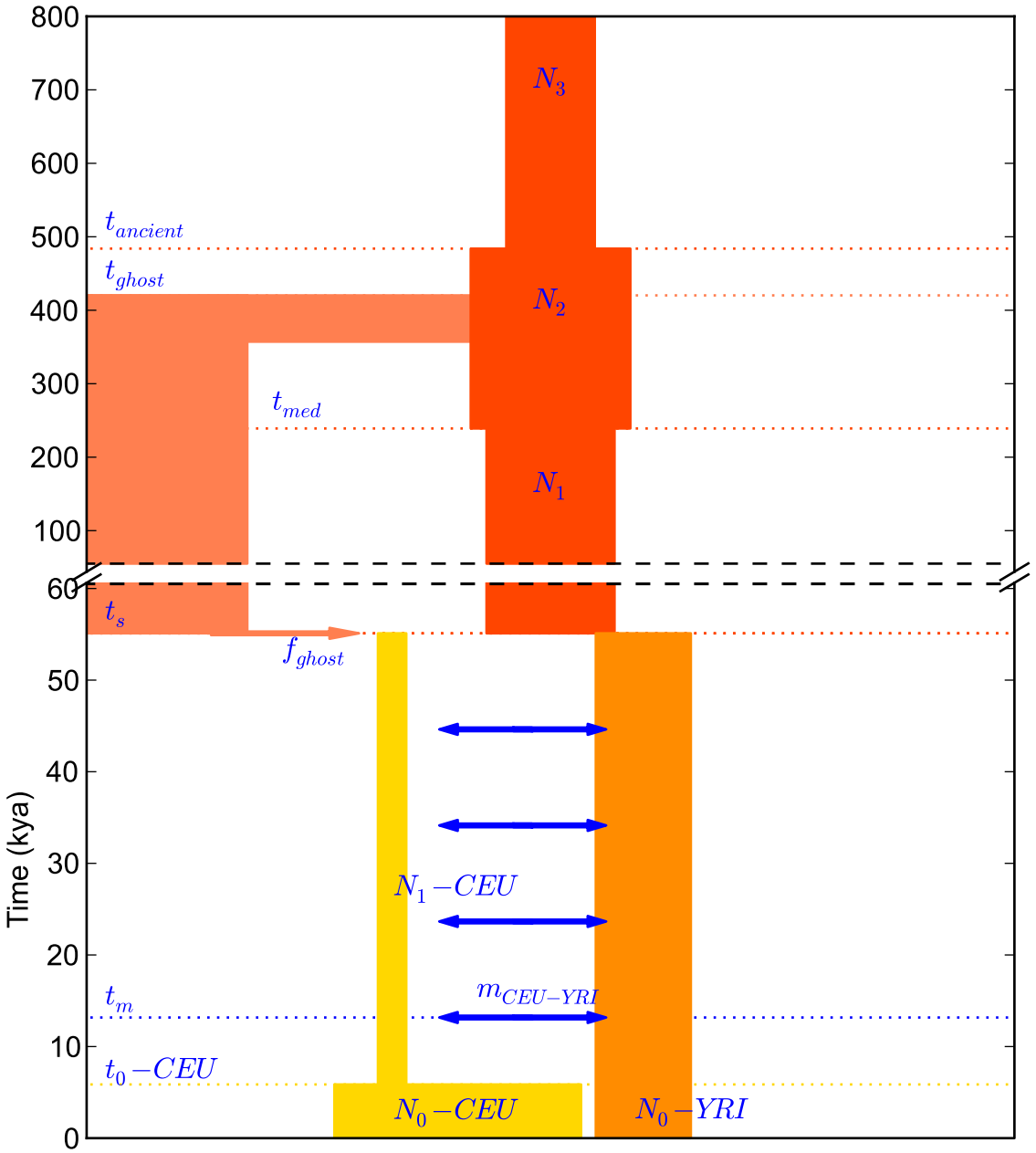

Figure 7. A history inferred from IBS sharing in Europeans and Yorubans. This is the simplest history we found to satisfactorily explain IBS tract sharing in the 1000 Genomes trio data. It includes ancient ancestral population size changes, an out-of-African bottleneck in Europeans, ghost admixture into Europe from an ancestral hominid, and a long period of gene flow between the diverging populations.
doi:10.1371/journal.pgen.1003521.g007

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 4.2 of Text S1 and Supplementary Figure S11). The discrepancy might result from biased population size estimates or from differences in the effects of errors on IBS tract and SFS data.

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

Table 2. Demographic parameters estimated from trio data.

Parameter	Estimate (kya)	Mean est. from simul.	Parameter	Estimate	Mean est. from simul.
$t_0 - CEU$	5.86	5.0	$N_0 - CEU$	13,298	106,036
t_m	13.17	15.03	$N_1 - CEU$	1,531	1,695
$t_0 - YRI$	55.11	47.13	$N_0 - YRI$	5,125	5,117
t_{med}	239.06	145.19	N_1	6,900	6,312
t_s	55.11	57.10	N_2	8,606	7,898
t_{ghost}	365.12	280.26	N_3	4,772	5,609
$t_{ancient}$	483.89	426.11	$m_{CEU-YRI}$	$1.52 \times 10^{-4}\ gen^{-1}$	1.79×10^{-4}
			f_{ghost}	0.0589	0.0393

These times, population sizes and migration rates parameterize the history depicted in Figure 7. The migration rate $m_{CEU-YRI}$ is the fraction of the European 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 African population made up of new European immigrants each generation during the same time period.

doi:10.1371/journal.pgen.1003521.t002

distinguish short, extreme population bottlenecks from longer, less extreme ones that produce similar reductions in total genetic diversity.

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 [27,28,54], losing only the information about which IBS tracts are adjacent to each other in the data. In exchange for this modest information loss, our 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.* can [24], whereas Li and Durbin are limited to a *post hoc* argument linking large or infinite population size to periods of divergence.

All parameter estimates, including admixture times, were found to be approximately unbiased in the context of a simple four-parameter model, but we observed a weak estimation bias for some parameters in the context of a complex history with 14 total 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 are the European-African 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. To minimize bias, it is crucial that we derive the IBS tract length distribution from Marjoram and Wall's SMC' [46], which provides a more accurate approximation to the correlation structure of sequential coalescence times than the earlier SMC [32] (see Methods and Supplementary Figure S12). 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. 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.* and Gronau, *et al.*, 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 [51,55,56]. 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 $T_{split}(\theta + \rho)$.

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.* [24,49] and by a analysis of human X chromosome diversity that employed the IM method of Hey and Nielsen [57]. 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

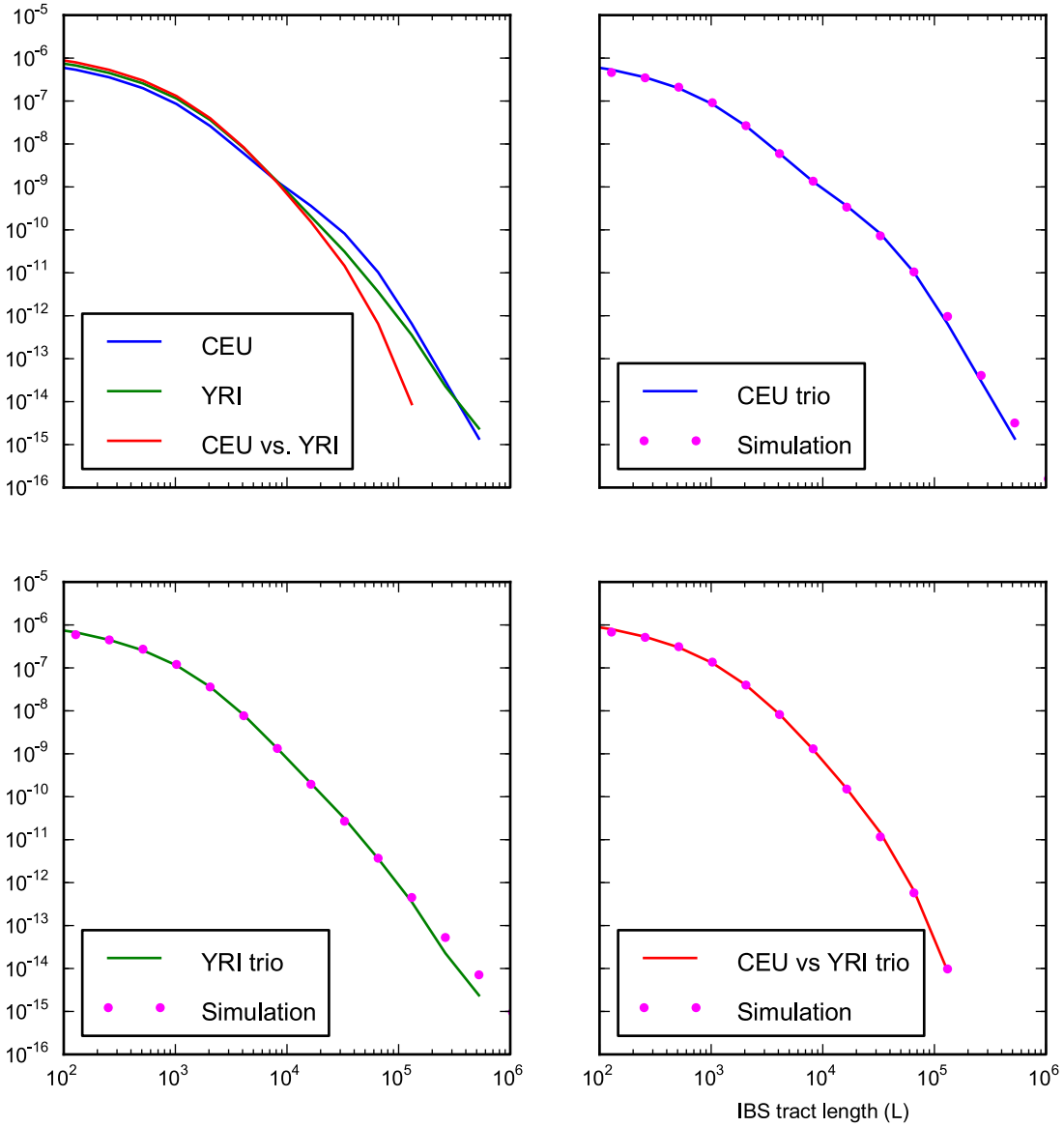

Figure 8. Accurate prediction of IBS sharing in the trio data. The upper left hand panel summarizes IBS tracts shared within the European and Yoruban 1000 Genomes trio parents, as well as IBS tract sharing between the two groups. The remaining three panels compare these real data to data simulated according to the history from Figure 7 with the maximum likelihood parameters from Table 2.
doi:10.1371/journal.pgen.1003521.g008

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 [17,58]. 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.* recently estimated that the Neanderthal gene flow event happened at least 47,000 years ago [59], 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 shared between inferred haplotypes [60,61], 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 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 [62,63]. Histories with high SFS-based likelihoods can be quite inconsistent with genomic LD [24]; to accurately describe neutral linkage in the genome, it is essential to harness linkage information as we have done here. Schaffner, *et al.* addressed this need with their 2005 demographic model that reproduces r^2 correlations between pairs of common SNPs [48], but our model explains genome-wide LD on a finer scale.

The empirical IBS tract length distributions studied here are highly similar among bootstrap subsamples, making it unlikely that they are influenced by isolated loci under strong selection or other regional peculiarities. However, the data and results could easily be influenced by background selection [64,65]. Background selection reduces diversity in a way that has been compared to a simple reduction in effective population size [64,66], 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 a much smaller collection of 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.

Methods

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) [32] 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 [31]. Because inference and computation under this model are so challenging, McVean and Cardin [32] introduced a simpler coalescent process (the SMC) for which

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

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 $\rho e^{-\rho t_r} 1(t_r < t) dt_r$, 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 $\zeta(t)$:

$$\begin{aligned} p(t_n|t_{n-1}) &= \exp(-\rho t_{n-1}) \delta_{t_{n-1}, t_n} \\ &+ \int_{t_{(r)}=0}^{\min(t_{n-1}, t_n)} \rho \exp(-\rho t_{(r)}) \left(\int_{\tau=0}^{t_{(r)}} \zeta(\tau) d\tau \right)^{-1} dt_{(r)} \cdot \zeta(t_n) \end{aligned} \tag{2}$$

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 $n_{L_{tot}}(L)$, 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 $L_{tot} - L - 1$ base pairs could be the leftmost polymorphic endpoint of an L-base IBS tract. Moreover, each of these $L_{tot} - L - 1$ 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 $\theta = 2N\mu$ and $\rho = 2Nr$, it follows that

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

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 $1_{hom}(b_L)$ 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

$$H_{\tau_s}(L,t) = \int_{t_{L-1}=\tau_s}^{\infty} H_{\tau_s}(L-1,t_{L-1}) \cdot P(t_L = t | t_{L-1}) \cdot P(1_{hom}(b_L) | t_L = t) dt_{L-1} \quad (3)$$

$$= e^{-t\theta} \int_{t_{L-1}=\tau_s}^{\infty} H_{\tau_s}(L-1,t_{L-1}) \cdot P(t_L = t | t_{L-1}) dt_{L-1}. \quad (4)$$

When $t = t_{L-1}$, the quantity $P(t_L = t | t_{L-1})$ 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 $P(t_L = t | t_{L-1})$ 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 9). Paul, Steinrücken, and Song used a similar computation to motivate the transition probabilities of their sequentially Markov conditional sampling HMM [54]:

$$p(t_L | t_{L-1}) = \int_{t_{(r)}=0}^{\min(t_{L-1},t)} \mathbb{P}(\text{an ancestor of } b_{L-1} \text{ recombined at time } t_{(r)}) \quad (5)$$

$$\cdot \mathbb{P}(t_L = t | t_L > t_{(r)}) dt_{(r)} + e^{-\rho t} \delta_{t_{L-1},t} \quad (6)$$

$$= \int_{t_{(r)}=t_{L-1}}^{\min(t_{L-1},t)} \rho e^{-\rho t_{(r)}} \cdot \frac{e^{-(t-\tau_s)}}{e^{\min(0,-t_{(r)}+\tau_s)}} dt_{(r)} + e^{-\rho t} \delta_{t_{L-1},t} \quad (7)$$

$$= \frac{\rho e^{-(t-\tau_s)-\rho\tau_s}}{1-\rho} (e^{\min(t-\tau_s, t_0-\tau_s)(1-\rho)} - 1) + e^{-(t-\tau_s)}(1 - e^{-\rho\tau_s}) + e^{-\rho t} \delta_t(t_{L-1}) \quad (8)$$

This yields that

$$H_{\tau_s}(L,t) = \int_{t_{L-1}=\tau_s}^{\infty} H_{\tau_s}(L-1,t_{L-1}) \cdot e^{-t\theta} \left(\frac{\rho e^{-(t-\tau_s)-\rho\tau_s}}{1-\rho} (e^{\min(t-\tau_s, t_{L-1}-\tau_s)(1-\rho)} - 1) + e^{-(t-\tau_s)}(1 - e^{-\rho\tau_s}) \right) \quad (9)$$
$$+ H_{t_s}(L-1,t) e^{-t(\rho+\theta)}.$$

To find $H_{\tau_s}(L,t)$, all we need to do is apply the integral operator (9) $L-1$ times to the base case

$$H_{\tau_s}(0,t) = e^{-(t-\tau_s)} - e^{-(t-\tau_s)-t\theta}. \quad (10)$$

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

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

$$H_{\tau_s}(L, t+\tau_s) = \sum_{i=0}^{L} A_i(L) e^{-t(1+i(\rho+\theta))} + B_i(L) e^{-t(1+i(\rho+\theta)+\theta)}, \quad (11)$$

with coefficients that satisfy the following recursions and base cases:

$$A_{i+1}(L+1) = A_i(L) e^{-\tau_s(\rho+\theta)} \left(1 - \frac{\rho}{(i(\rho+\theta)+\rho)(1+i(\rho+\theta))} \right) \quad (12)$$

$$B_{i+1}(L+1) = B_i(L) e^{-\tau_s(\rho+\theta)} \left(1 - \frac{\rho}{(i+1)(\rho+\theta)(1+i(\rho+\theta)+\theta)} \right) \quad (13)$$

$$B_0(L+1) = \sum_{i=0}^{L} \frac{\rho A_i(L) e^{-\tau_s(\rho+\theta)}}{(i(\rho+\theta)+\rho)(1+i(\rho+\theta))} + \frac{\rho B_i(L) e^{-\tau_s(\rho+\theta)}}{(i+1)(\rho+\theta)(1+i(\rho+\theta)+\theta)} + \left(e^{-\tau_s\theta} - e^{-\tau_s(\theta+\rho)}\right) \sum_{i=0}^{L} \left(\frac{A_i(L)}{1+i(\rho+\theta)} + \frac{B_i(L)}{1+\theta+i(\rho+\theta)} \right) \quad (14)$$

$$A_0(L+1) = 0 \quad (15)$$

$$A_0(0) = 1, \quad (16)$$

$$B_0(0) = -e^{-\theta\tau_s} \quad (17)$$

It is straightforward to prove Claim 1 by applying the integral operator (9) to expression (11). The upshot is that

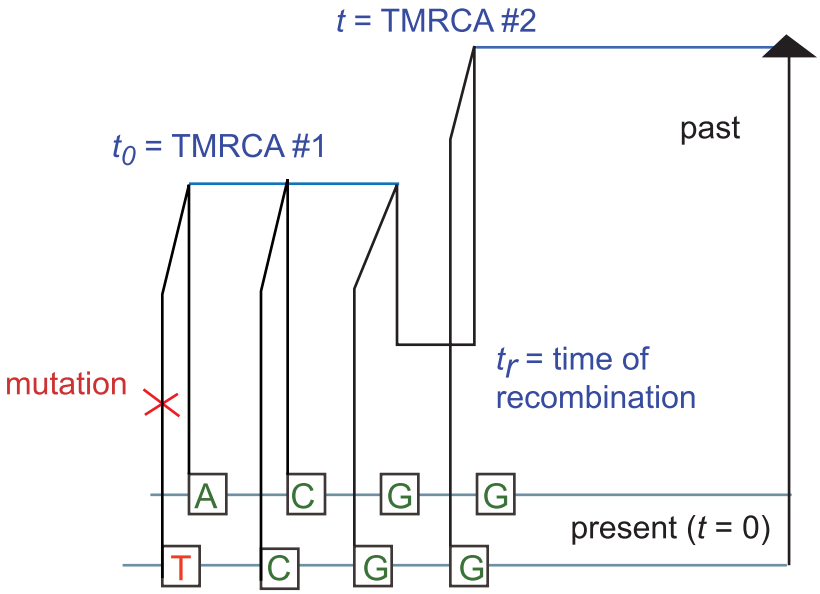

Figure 9. The coalescent with recombination and the sequentially Markov coalescent associate an observed pair of DNA sequences with a history that specifies a time to most recent common ancestry for each base pair. Polymorphisms are caused by mutation events, while changes in TMRCA are caused by recombination events.
doi:10.1371/journal.pgen.1003521.g009

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

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 (9), 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. Figure 10 depicts an IBS tract with three distinct TMRCAs separated by 2 internal recombinations. 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 $\mathbb{E}[t_L] \gg \mathbb{E}[t_{L-1}]$ [67].

To speed the computation, we assume that an L-base IBS tract contains at most two internal recombinations. To make this precise, we let $H_{\tau_s}(L) \approx H^{(0)}_{\tau_s}(L) + H^{(1)}_{\tau_s}(L) + H^{(2)}_{\tau_s}(L)$, where $H^{(i)}_{\tau_s}$ 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:

$$H^{(0)}_{\tau_s}(L) = \int_{t=\tau_s}^{\infty} e^{-(t-\tau_s)} \cdot (1-e^{-t\theta}) \cdot e^{-tL(\rho+\theta)} dt \qquad (18)$$

$$= \frac{e^{-\tau_s L(\rho+\theta)}}{1+L(\rho+\theta)} - \frac{e^{-\tau_s (L(\rho+\theta)+\theta)}}{1+L(\rho+\theta)+\theta} \qquad (19)$$

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

$$H^{(1)}_{\tau_s}(L) = \sum_{L_1=1}^{L-1} \int_{t_0=\tau_s}^{\infty} \int_{t=\tau_s}^{\infty} \int_{t_{(r)}=0}^{\min(t_0,t)} e^{-t_0}(1-e^{-t_0\theta}) \qquad (20)$$
$$\cdot e^{-t_0 L_1(\rho+\theta)} \cdot \rho e^{-\rho t_{(r)}}$$

$$\cdot e^{-(t-t_{(r)})} \cdot e^{-t(L-L_1)(\rho+\theta)} dt_{(r)} dt dt_0 \qquad (21)$$

In the supplementary section S??, we evaluate this expression in closed form after approximating the sum by an integral. In the same way, we compute $H^{(2)}_{\tau_s}(L)$ (see Section 1.2 in Text S1).

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

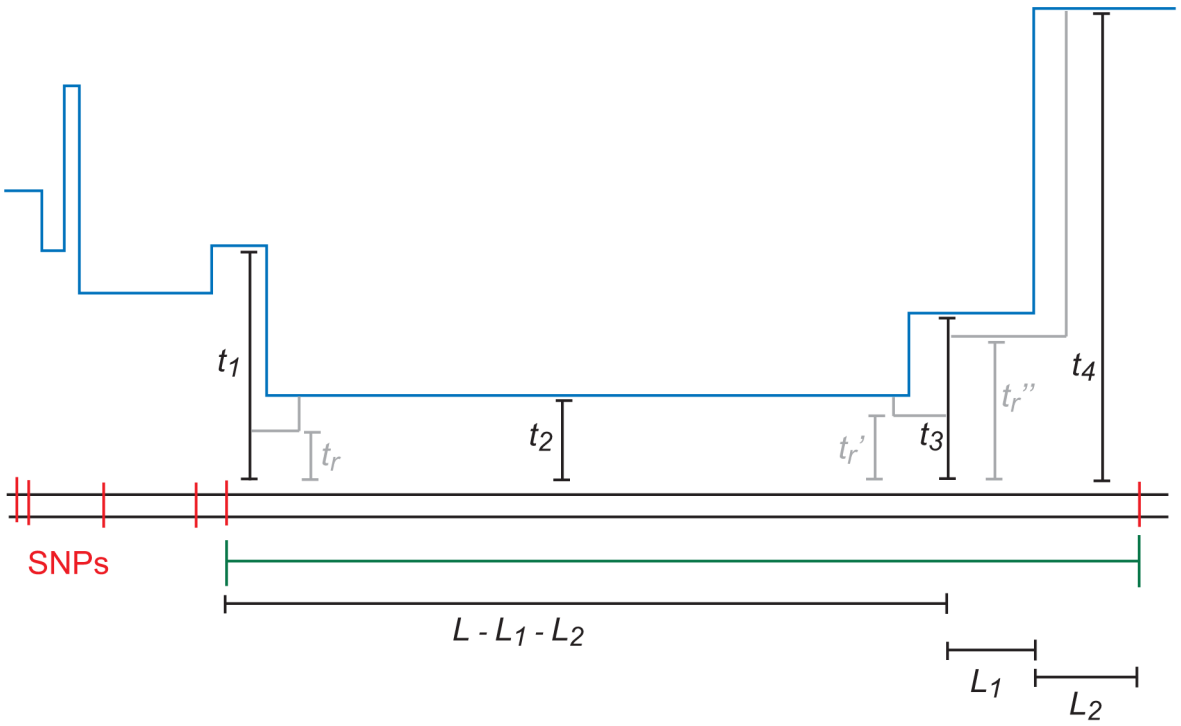

Figure 10. An L-base IBS tract with three recombination events in its history. A blue skyline profile represents the hidden coalescence history of this idealized IBS tract. In order to predict the frequency of these tracts in a sequence alignment, we must integrate over the coalescence times t_1,t_2,t_3 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.
doi:10.1371/journal.pgen.1003521.g010

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 $H^{(0)}_{\tau_a,\tau_s,f}(L), H^{(1)}_{\tau_a,\tau_s,f}(L), \ldots$ 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 $H^{(0)}_{\tau_a}(L)$:

$$H^{(0)}_{\tau_a,\tau_s,f}(L) = \frac{fe^{-\tau_a L(\rho+\theta)} + (1-f)e^{-\tau_s L(\rho+\theta)}}{1+L(\rho+\theta)} - \frac{fe^{-\tau_a(L(\rho+\theta)+\theta} + (1-f)e^{-\tau_s(L(\rho+\theta)+\theta)}}{1+L(\rho+\theta)+\theta} \quad (22)$$

$$= fH^{(0)}_{\tau_a} + (1-f)H^{(0)}_{\tau_s}. \quad (23)$$

The next term $H^{(1)}_{\tau_a,\tau_s,f}(L)$ is much more challenging to compute exactly; this is done in supplementary section S1.3. 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 (see Supplementary Figure S13). 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 $\mathbf{n} = (v_0 = 0, v_1, v_2, v_3 = \infty)$ with $v_1 < v_2$ and the vector of sizes $N = (N_1, N_2, N_3)$. 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 $H_{\mathbf{n},N}(L)$ 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 $N(t) = \sum_i N_i 1(v_i \le t < v_{i+1})$. This population has a coalescence density function that is a linear combination of exponentials, which implies that $H^{(0)}_{\mathbf{n},N}(L)$ is a linear combination of the quantities $H^{(0)}_{v_i,N_i}$:

$$H^{(0)}_{v,\mathbf{N}}(L) = \int_{t=0}^{\infty} \left(1(t < v_0)\frac{1}{N_0}e^{-t/N_0} + 1(v_1 \le t < v_2)\frac{1}{N_1}e^{-v_1/N_0-(t-v_1)/N_1} \right. \quad (24)$$

$$\left. + 1(t \ge v_2)\frac{1}{N_2}e^{-v_1/N_0-(v_2-v_1)/N_1-(t-v_2)/N_2} \right) e^{-tL(\rho+\theta)}dt$$

$$= H^{(0)}_{0,N_0}(L) + e^{-v_1/N_0}\left(H^{(0)}_{v_1,N_1}(L) - H^{(0)}_{v_1,N_0}(L)\right) \quad (25)$$

$$+ e^{-v_1/N_0-(v_2-v_1)/N_1}\left(H^{(0)}_{v_2,N_2}(L) - H^{(0)}_{v_2,N_1}(L)\right).$$

As in the case of an admixed population, the next term $H^{(1)}_{\mathbf{n},\mathbf{N}}(L)$ 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 v_1 and the right hand part coalesces later than v_1 during a time period of smaller effective population size). The higher terms ($H^{(2)}_{\mathbf{n},\mathbf{N}}$, etc.) are more complicated still. Rather than attempt to compute these terms for

a simple bottleneck history, we have developed an approximation for $H_{\mathbf{n},\mathbf{N}}(L)$ 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 v_i and v_{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 $\sum_j \frac{1}{N_j}e^{-(t-t_{(r)})/N_j}1(v_j<t\le v_{j+1})$. If we let $\hat{\boldsymbol{H}}_{\mathbf{n},\mathbf{N}}(L)$ be the IBS tract spectrum under this assumption, we have that

$$\hat{\boldsymbol{H}}_{v,\mathbf{N}}(L)=H_{0,N_0}(L)+e^{-v_1/N_0}\left(H_{v_1,N_1}(L)-H_{v_1,N_0}(L)\right) \quad (26)$$

$$+e^{-v_1/N_0-(v_2-v_1)/N_1}\left(H_{v_2,N_2}(L)-H_{v_2,N_1}(L)\right)$$

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 3 shows a close agreement between $\hat{\boldsymbol{H}}_{\mathbf{n},\mathbf{N}}(L)$ and the IBS tracts in data simulated under bottleneck histories with MS.

Improving 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 underestimation can bias our estimates of population size and other demographic parameters (see Supporting Figure S22), but this bias can be substantially 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 [46]. 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 (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 $\mathbb{P}_{SMC}(t_n=t|t_{n-1}=t)=e^{-\rho t}$. 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,

$$\mathbb{P}_{SMC',\tau_S}(t_n=t\mid t_{n-1}=t)=e^{-\rho t} + \int_{t_r=0}^{\tau_S}\rho e^{-\rho t_r}\left(\int_{t'=t_r}^{\tau_S}e^{-(t'-t_r)}dt'\right. \quad (27)$$

$$\left.+\int_{t'=\tau_S}^{t}\frac{1}{2}\cdot 2e^{-(\tau_S-t_r)-2(t'-\tau_S)}dt'\right)dt_r \quad (28)$$

$$+\int_{t_r=\tau_S}^{t}\rho e^{-\rho t_r}\int_{t'=t_r}^{t}\frac{1}{2}\cdot 2e^{-2(t'-t_r)}dt'dt_r \quad (29)$$

$$=1-\frac{\rho}{4}\left(3-2e^{-\tau_S}-2e^{-t-(t-\tau_S)}+e^{-2(t-\tau_S)}+2t-2\tau_S\right)+O(\rho^2) \quad (30)$$

Motivated by Equation (30), 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 (18) to compute the probability of observing L base pairs that are IBS with no internal recombinations that change the coalescence time. We obtain that

$$H^{(0')}_{\tau_S}(L)=\int_{t=\tau_S}^{\infty}e^{-(t-\tau_S)}\cdot(1-e^{-t\theta})\cdot e^{-tL\theta} \quad (31)$$

$$\cdot\exp(-\rho L(3-2e^{-\tau_S}-2e^{-t-(t-\tau_S)}+e^{-2(t-\tau_S)}+2t-2\tau_s)/4)dt \quad (32)$$

$$=2e^{-L((1-e^{-\tau_S}-e^{\tau_S}+e^{2\tau_S})\rho/2+\tau_S\theta)}. \quad (33)$$

$$\left(\frac{1}{2+L\rho+2L\theta}\cdot{}_1F_1\left(1,\frac{6+L\rho+2L\theta}{4},-\frac{L\rho}{4}\left(1+2e^{\tau_S}-2e^{2\tau_S}\right)\right)\right. \quad (34)$$

$$\left.-\frac{e^{-\tau_S\theta}}{2+L\rho+2(L+1)\theta}\cdot{}_1F_1\left(1,\frac{6+L\rho+2(L+1)\theta}{4},\right.\right. \quad (35)$$
$$\left.\left.-\frac{L\rho}{4}\left(1+2e^{\tau_S}-2e^{2\tau_S}\right)\right)\right).$$

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.

Inference strategy

The previous section described how we compute $\mathbb{E}[n_{L_{tot}}(L)]$, 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 $n_{L_{tot}}(L)$ 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:

$$\log\mathcal{L}(\Theta)= \sum_{L=1}^{\infty}-\mathbb{E}_{\Theta}[n_{L_{tot}}(L)]+n_{L_{tot}}(L)\log\mathbb{E}_{\Theta}[n_{L_{tot}}(L)]-\log n_{L_{tot}}(L)! \quad (36)$$

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 $n_{L_{tot}}(b_i,b_{i+1}) = \sum_{L=b_i}^{b_{i+1}-1} n_{L_{tot}}(L)$, we define

$$\log \mathcal{L}_b(\Theta) = \sum_i -\mathbb{E}_\Theta[n_{L_{tot}}(b_i,b_{i+1})] + n_{L_{tot}}(b_i,b_{i+1}) \log \mathbb{E}_\Theta[n_{L_{tot}}(b_i,b_{i+1})] - \log n_{L_{tot}}(b_i,b_{i+1})! \quad (37)$$

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). 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.

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 [52,53]. 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.*, among others, have shown that sites of elevated mutation rate are not always grouped together in the human genome [52]. 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 $P'_{IBS}(L)$ 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 $P_{IBS}{}''(L)$, the difference between $P_{IBS}{}'(L)$ and $P_{IBS}{}''(L)$ would be on the order of the miniscule difference between $P_{IBS}(L)$ and $P_{IBS}(1.038 \times L)$. Neglecting this second effect, we replace $P_{IBS}(L)$ with $0.999 \times P_{IBS}(L) + 0.001 \times P_{IBS}{}'(L)$ for the purpose of inferring demography from human data.

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 [51]. 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 Section 3.1 of Text S1 for details). 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 and Durbin [28], 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).

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 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.

Supporting Information

Figure S1 Each of these histograms was generated from 100 simple admixture history datasets that were simulated with gene flow time $\tau_a = 400$. The true parameter value is shown in red. All parameter estimates have low variance and appear consistent.
(EPS)

Figure S2 These histograms record the distribution of parameter estimates for 100 simple admixture histories with gene flow time $\tau_a = 200$. True parameter values are shown in red.
(EPS)

Figure S3 These IBS tract likelihood surfaces were generated from two of the 200 simulated data sets that were analyzed to produce Table 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 of the main text.
(EPS)

Figure S4 This figure compares IBS tract length frequencies in the 1000 Genomes low coverage pilot data to the frequencies of IBS tracts that remain in the high-coverage trio data after the addition of Poisson-distributed false SNPs to simulate an error rate of 5×10^{-6} per base pair. In the notation of section S?? of Supplementary Text S1, it plots the low coverage IBS tract frequencies $f_{lc}(L)$ along with the error-degraded high coverage trio frequencies $f_{hc}(L)e^{-L\varepsilon_{IBS}}$. For $L>1000$, we can see that $f_{lc}^{CEU}(L)\approx f_{hc}^{CEU}(L)e^{-L\varepsilon_{IBS}^{CEU}}$, $f_{lc}^{YRI}(L)\approx f_{hc}^{YRI}(L)e^{-L\varepsilon_{IBS}^{YRI}}$, and $f_{lc}^{CEU-YRI}(L)\approx f_{hc}^{CEU-YRI}(L)e^{-L\varepsilon_{IBS}^{CEU}}$ when we let $\varepsilon_{IBS}^{CEU}=\varepsilon_{IBS}^{YRI}=5\times10^{-6}$.
(EPS)

Figure S5 IBS tract sharing in data simulated under the Gutenkunst, *et al.* demographic model. Each panel here shows the length distribution of IBS tracts in the 1000 Genomes trios compared to the length distributions that are obtained by simulating data under the model of Gutenkunst, *et al.* 2009 [15]. Compared to real human data, this demographic model predicts too few long IBS tracts shared between Europeans and Africans, as well as too few long IBS tracts shared within Europe.
(EPS)

Figure S6 IBS tract sharing in data simulated under the Li and Durbin demographic model. Li and Durbin's PSMC does not measure the extent of gene flow between species, but implicitly uses IBS tracts to estimate past population sizes. We simulated data under each of the European and African histories published in Li and Durbin, 2011 [23] and plotted their IBS tract length frequencies against frequencies from the 1000 Genomes trios. Long tracts have similar frequencies between the real and simulated data, though short tracts are less accurately predicted by the PSMC.
(EPS)

Figure S7 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 S?? and the maximum likelihood parameter values shown in green.
(EPS)

Figure S8 Results of block bootstrapping and parametric bootstrapping: Part II of III.
(EPS)

Figure S9 Results of block bootstrapping and parametric bootstrapping: Part III of III.
(EPS)

Figure S10 The solid colored blocks in this figure depict the phases of the demographic history that was inferred from the 1000 genomes trio data. The overlaid black lines depict the mean history inferred from replicate MS simulations.
(EPS)

Figure S11 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.
(EPS)

Figure S12 Parameter inference without the SMC'. These histograms were generated by optimizing the parameters of an IBS tract length distribution derived purely from the SMC, not the SMC'. The simulated datasets are the same ones used to generate Figure S12, but in this case, model inaccuracy leads us to underestimate the effective population size and the admixture fraction.
(EPS)

Figure S13 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.
(EPS)

Table S1 Numerical performance of ∂a∂i optimization vs. IBS tract inference. The left 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. The right table contains the result of 20 analogous optimizations that use an equivalent amount of IBS tract data (one of the 100 replicates used to generate Table 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.
(PDF)

Text S1 Supporting information.
(PDF)

Acknowledgments

We thank R. Gutenkunst for sharing his NIEHS site frequency spectrum data with us, E. Huerta-Sanchez and A. Ferrer-Admetlla for computational and server support, and R. Durbin for the initial idea to compute the IBS tract distribution under the SMC. We also thank Y. Song, P. Ralph, G. Coop, N. Barton, K. Lohse, J. Schraiber, and members of the Nielsen lab for helpful discussions, as well as three anonymous reviewers for their work on this paper.

Author Contributions

Conceived and designed the experiments: KH RN. Performed the experiments: KH. Analyzed the data: KH. Contributed reagents/materials/analysis tools: RN. Wrote the paper: KH RN. Derived mathematical results: KH.

References

1. Slatkin M, Madison W (1989) A cladistic measure of gene ow inferred form the phylogenies of alleles. Genetics 123: 603–613.
2. Templeton A (2002) Out of Africa again and again. Nature 416: 45–51.
3. Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.
4. Slatkin M, Hudson R (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129: 555–562.
5. Wakeley J, Hey J (1997) Estimating ancestral population parameters. Genetics 145: 847–855.
6. Griffths R, Tavaré S (1994) Ancestral inference in population genetics. Stat Sci 9: 307–319.
7. Griffths R, Tavaré S (1994) Simulating probability distributions in the coalescent. Theor Pop Biol 46: 131–159.
8. Kuhner M, Yamato J, Felsenstein J (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140: 1421–1430.
9. Nielsen R (1998) Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor Pop Biol 53: 143–151.
10. Nielsen R (1997) A likelihood approach to populations samples of microsatellite alleles. Genetics 146: 711–716.
11. Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci USA 98: 4563–4568.
12. Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov Chain Monte Carlo approach. Genetics 158: 885–896.
13. Yang Z, Rannala B (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol Biol Evol 14: 717–724.
14. Gronau I, Hubisz M, Gulko B, Danko C, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43: 1031–1034.
15. Schierup M, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879–891.
16. Strasburg J, Rieseberg L (2010) How robust are "isolation with migration" analyses to violations of the IM model? A simulation study. Mol Biol Evol 27: 297–310.
17. Green R, Krause J, Briggs A, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neanderthal genome. Science 328: 710–722.
18. Rasmussen M, Guo X,Wang Y, Lohmueller K, Rasmussen S, et al. (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334: 94–98.
19. Tavaré S, Balding D, Griffths R, Donnelly P (1997) Inferring coalescence times from DNA sequence data. Genetics 145: 505–518.
20. Pritchard J, Seielstad M, Perez-Lezun A, Feldman M (1999) Population growth of human Y chromosomes: a study of Y chromosome microsattelites. Mol Biol Evol 16: 1791–1798.
21. Beaumont M, Zhang W, Balding D (2002) Approximate Bayesian computation in population genetics. Genetics 192: 2025–2035.
22. Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154: 931–942.
23. Williamson S, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, et al. (2005) Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci USA 102: 7882–7887.
24. Gutenkunst R, Hernandez R, Williamson S, Bustamante C (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5: e1000695.
25. Nielsen R, Wiuf C (2005) Composite likelihood estimation applied to single nucleotide polymorphism (SNP) data. In: ISI Conference Proceedings. 5–12 April 2005. Sydney, Australia. URL http://www.math.ku.dk/pbx512/journalWiuf/ISI2005.pdf.
26. Wiuf C (2006) Consistency of estimators of population scaled parameters using composite likelihood. Math Biol 53: 821–841.
27. Hobolth A, Christensen O, Mailund T, Schierup M (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genetics 3: e7.
28. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475: 493–496.
29. Steinrücken M, Paul J, Song Y (2012) A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theor Popul Biol. Epub ahead of print. doi:10.1016/j.tpb.2012.08.004.
30. Sheehan S, Harris K, Song Y (2013) Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics. Epub ahead of print. Doi: 10.1534/genetics.112.149096
31. Wiuf C, Hein J (1999) Recombination as a point process along sequences. Theor Popul Biol 55: 248–259.
32. McVean G, Cardin N (2005) Approximating the coalescent with recombination. Phil Trans Royal Soc B 360: 1387–1393.
33. Mailund T, Halager A, Westergaard M, Dutheil J, Munch K, et al. (2012) A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. PLoS Genetics 8: e1003125.
34. Miller W, Schuster S, Welch A, Ratan A, Bedoya-Reina O, et al. (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. Proc Natl Acad Sci USA 109: E2382–E2390.
35. Browning B, Browning S (2011) A fast, powerful method for detecting identity by descent. Am J Hum Gen 88: 173–182.
36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Gen 81: 559–575.
37. Moltke I, Albrechtsen A, Hansen T, Nielsen F, Nielsen R (2011) A method for detecting IBD regions simultaneously in multiple individuals- with applications to disease genetics. Genome Res 21: 1168–1180.
38. Gusev A, Lowe J, Stoffel M, Daly M, Altshuler D, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19: 318–326.
39. Hayes B, Visscher P, McPartlan H, Goddard M (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res 13: 635–643.
40. MacLeod I, Meuwissen T, Hayes B, Goddard M (2009) A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genet Res 91: 413–426.
41. Palamara P, Lencz T, Darvasi A, Pe'er I (2012) Length distributions of identity by descent reveal fine-scale demographic history. Am J Hum Gen 91: 809–822.
42. Ralph P, Coop G (2013) The geography of recent genetic ancestry across Europe. PLoS Biology 11: e1001555
43. Pool J, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711–719.
44. Gravel S (2012) Population genetics models of local ancestry. Genetics 191: 607–619.
45. Moorjani P, Patterson N, Hirschhorn J, Keinan A, Hao L, et al. (2011) The history of African gene flow into southern Europeans, Levantines, and Jews. PLoS Genetics 7: e1001373.
46. Marjoram P, Wall J (2006) Fast "coalescent" simulation. BMC Genetics 7: 16.
47. Pritchard J (2011) Whole-genome sequencing data offfer insights into human demography. Nature Genetics 43: 923–925.
48. Schaffner S, Foo C, Gabriel S, Reich D, Daly M, et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15: 1576–1583.
49. Gravel S, Henn B, Gutenkunst R, Indap A, Marth G, et al. (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA 108: 11983–11988.
50. Press W, Teukolsky S, Vetterling W, Flannery B (2007) Numerical Recipes: The Art of Scientific Computing. 3rd edition. Cambridge University Press.
51. The 1000 Genomes Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
52. Hodgkinson A, Ladoukakis E, Eyre-Walker A (2009) Cryptic variation in the human mutation rate. PLoS Biology 7: e1000027.
53. Kong A, Gudbjartsson D, Sainz J, Jonsdottir G, Gudjonsson S, et al. (2002) A high-resolution recombination map of the human genome. Nature 31: 241–247.
54. Paul J, Steinrücken M, Song Y (2011) An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics 187: 1115–1128.
55. Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. Nature Rev Gen 13: 745–753.
56. Kong A, Frigge M, Masson G, Besenbacher S, Sulem P, et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. Nature 488: 471–475.
57. Cox M, Woerner A, Wall J, Hammer M (2008) Intergenic DNA sequences from the human X chromosome reveal high rates of global gene ow. BMC Genetics 9: 1471–2156.
58. Noonan J, Coop G, Kudarvalli S, Smith D, Krause J, et al. (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113–1118.
59. Sankararaman S, Patterson N, Li H, Pääblo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. PLoS Genetics 8: e1002947.
60. Browning S, Browning B (2009) A uni_ed approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Gen 84: 210–223.
61. Li Y, Willer C, Ding J, Scheet P, Abecasis G (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Gen Epidem 34: 816–834.
62. Sabeti P, Reich D, Higgins J, Levine H, Richter D, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
63. Pickerell J, Coop G, Novembre J, Kudarvalli S, Li J, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826–837.

64. Charlesworth D, Charlesworth B, Morgan M (1995) The pattern of neutral molecular variation under the background selection model. Genetics 141: 1619–1632.
65. McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genetics 5: e1000471.
66. Lohmueller K, Albrechtsen A, Li Y, Kim S, Korneliussen T, et al. (2011) Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS Genetics 7: e1002326.
67. Barton N (June 28, 2012). Personal communication.
68. Hudson R (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18: 337–338.