Published in final edited form as:
IEEE/ACM Trans Comput Biol Bioinform. 2013 ; 10(5): 1137–1149.

Coalescent-based Method for Learning Parameters of Admixture Events from Large-Scale Genetic Variation Data

Ming-Chi Tsai,
Joint CMU-Pitt PhD Program in Computational Biology, Pittsburgh, PA, 15213.
mingchit@andrew.cmu.edu

Guy Blelloch,
Department of Computer Science, Carnegie Mellong University, Pittsburgh, PA 15213.
guyb@cs.cmu.edu

R. Ravi, and
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213. ravi@cmu.edu

Russell Schwartz
Department of Biological Science, Carnegie Mellon University, Pittsburgh, PA 15213.
russells@andrew.cmu.edu

Abstract

Detecting and quantifying the timing and the genetic contributions of parental populations to a hybrid population is an important but challenging problem in reconstructing evolutionary histories from genetic variation data. With the advent of high throughput genotyping technologies, new methods suitable for large-scale data are especially needed. Furthermore, existing methods typically assume the assignment of individuals into subpopulations is known, when that itself is a difficult problem often unresolved for real data. Here we propose a novel method that combines prior work for inferring non-reticulate population structures with an MCMC scheme for sampling over admixture scenarios to both identify population assignments and learn divergence times and admixture proportions for those populations using genome-scale admixed genetic variation data. We validated our method using coalescent simulations and a collection of real bovine and human variation data. On simulated sequences, our methods show better accuracy and faster runtime than leading competitive methods in estimating admixture fractions and divergence times. Analysis on the real data further shows our methods to be effective at matching our best current knowledge about the relevant populations.

Index Terms

J.3.a Biology and genetics; E.1.d Graphs and networks; H.1.1.b Information theory; F.2.2.b Computations on discrete structures

1 Introduction

Understanding modern human origins and evolution has long been a central question in anthropology and human genetics. Since our emergence as a species, humans have diverged into numerous subpopulations. In some instances, individuals from different subpopulations

have come into contact, yielding genetically mixed populations. We call this incorporation of genetic materials from one genetically distinct population into another admixture. This process is believed to be common in human populations, where migrations of peoples have repeatedly brought together populations that were historically reproductively isolated from one another. This can be seen, for instance, in the United States where many African Americans contain varying amounts of ancestry from Europe and Africa [1]. Reconstructing historical admixture scenarios also has important practical value in biomedical contexts. For instance, learning the correct time scale on which different strains of the human immunodeficiency virus (HIV) have diverged would be useful for understanding the circumstances surrounding the emergence of the acquired immune deficiency syndrome (AIDS) pandemic as well as its continued genetic divergence [2]. In statistical genetics, studying admixture and population structure can help in identifying and correcting for confounding effects of population structure in disease association tests [3]. Studying admixture can also help in understanding the acquisition of disease-resistance alleles [4].

A recent explosion in available genome-scale variation data has led to considerable prior work on characterizing relationships among admixed populations. One popular approach for qualitatively characterizing such relationships derives from the observation that principal component analysis (PCA) provides a way to visually capture such relationships for complex population mixtures [5], [6]. While such methods provide a powerful tool for visualizing fine substructure and admixture, however, they typically require considerable manual intervention and interpretation to translate these visualizations into concrete models of the population history. Furthermore, these methods provide only limited quantitative data on relationships between admixed populations, providing fractions of admixed data but not complete parameters of an admixture model, such as timing of divergence and admixture events. Other methods focus on the related problem of finding detailed assignments of local genomic regions of admixed individuals to ancestral populations [7], [8], [9], which provides complementary information with important uses in admixture mapping, but similarly provides little direct insight into the history by which these admixtures occurred.

Inferring detailed quantitative models of historical admixture events, especially the timing of these events, remains a difficult problem. It is typically addressed by inferring basic parameters of a single admixture event — the creation of a hybrid population from two ancestral populations. Some methods do examine more complex scenarios, such as the isolation with migration model [10], and others different parameters, such as effective population size [11]. We, however, focus here on the more standard three-population scenario and the joint inference of both the admixture proportion and the times of divergence and admixture. Most methods for this problem use allele frequencies to estimate admixture proportions by assuming that admixed populations will exhibit frequencies that are linear combinations of those of their parental populations and optimizing with respect to some error model [12]. While such methods can be very effective, they generally require substantial simplifying assumptions regarding the admixture process, for example assuming the absence of mutations after admixture events. Such an assumption can be problematic when the mutation rate is high or when the admixture is sufficiently ancient that mutations novel to the admixed populations are no longer negligible.

This issue has been previously addressed by methods utilizing coalescent theory, [13], [14]. a probabilistic model of ancestral relationships that can be used to efficiently sample among possible evolutionary histories of a set of individuals in a population. *MEAdmix* [13], for instance, uses coalescent theory to compute expected numbers of segregating sites (or mutations) between lineages then identifies an optimal admixture proportion by minimizing the squared difference between the expected number and observed number of segregating sites. While such methods were significant advances on the prior art, they have difficulty scaling to large data sets due to long computation time and numerical errors. With genomic-scale data becoming widely available from whole-genome variation studies, new methods are needed to make full use of such data in achieving more accurate and detailed models of population dynamics. The prior methods also assume that we know in advance the population structure and assignment of individuals to that structure, a restriction that is increasingly suspect as we seek ever finer resolution in our population models.

In the present work, we develop a novel approach to reconstructing parameters of admixture events that addresses several limitations of the prior art. Our method is designed to learn, directly from the molecular data, what subpopulations are present in a given data set, the sequence of divergence events and divergence times that produced them, whether admixture exists between these subpopulations, and, if so, with what proportions admixed populations draw their ancestry from each ancestral population.

More formally, we assume the input to the problem is a $n \times m$ [0,1] matrix D where element D_{ij} represents the allele of the jth genetic variation site for the ith taxon. The output is a tuple $T = \{P_1, P_2, P_3, t_1, t_2, \alpha, \theta\}$. P_1, P_2, and P_3 form a tripartition of the rows of D, $t_1 \in \mathbb{R}^+$, $t_2 \in \mathbb{R}^+$, $\alpha \in [0, 1]$. These outputs model a simple history of a population group that arose from an ancestral population, divided into two subpopulations, and then admixed to produce a third subpopulation. P_1, P_2, and P_3 are an assignment of rows of D (taxa) to the three final subpopulations, t_1 is the elapsed time from the admixture event to the present, t_2 is the elapsed time from the divergence event to the admixture event, and α is the fractional contribution of the first population to the admixture. θ is a scaling parameter, explained in more detail in Materials and Methods, that combines effective population size and mutation rate. The problem does not have a simple, standard objective function and the contribution of the present work is in part to define a likelihood-based objective function, explained in detail in Materials and Methods below. We further note that the tripartition is commonly assumed in the literature to be included in the input. A further contribution of the present work is to infer the tripartition as an output together with the real-valued parameters, treating the variation matrix D as the sole input.

We have created a novel two-step inference model called Consensus-tree based Likelihood Estimation for AdmiXture (*CLEAX*). Rather than inferring the population history directly from the molecular data [10], [13], [14], we first learn a set of summary descriptions of the overall population history from the molecular data D corresponding to a inferred set of subpopulations and a set of bipartitions, i.e., partitions of the taxa into two non-empty subsets, with a weight associated with each bipartition. Once the set of summary descriptions is obtained, we then apply a coalescent-based inference model on the summary descriptions to learn divergence times and admixture fractions for the model. A key

advantage of our two-step inference model is substantial reduction in the computational cost for large data sets, making it possible to perform more precise and reliable inferences using genomic-scale variation datasets. In addition, the proposed method has the advantages of learning divergence times and admixture times in a more general framework encompassing simultaneous inference of population groups, their shared ancestry, and potentially other parameters of their history.

2 Materials and Methods

To learn population history for a dataset, our approach first tries to determine a number of subpopulations K and a summary description $H = (B^M, W)$ that approximates the number of segregating sites (or mutations) that separate any given pair of subpopulations. We then use the resulting discrete model of population divergence events to estimate expected times between events and the admixture proportions between subpopulations.

As with much of the prior work [12], [13], [14], [15], we specifically address the problem of accurately reconstructing parameters of a single historical admixture event. As shown in Fig. 1(a), we will assume that there exists a single ancestral population P_0 before time t_2. A divergence event then occurs at time t_2 that results in the formation of two subpopulations P_1 and P_3. Finally, at time t_1, an admixture event occurs between the two parental populations P_1 and P_3 to form a new admixed population P_2. The admixed population P_2 is composed of an α fraction of individuals from P_1 and a $1 - \alpha$ fraction of individuals from P_3. Except for the admixture event itself at t_1, all populations are assumed genetically isolated throughout history. The model can be characterized by the time of the divergence (t_2), the time of admixture (t_1), and the admixture proportion (α). Additional hidden parameters include mutation rate, μ, and the effective population size for the ancestral population (N_0), the two parental populations (N_1 and N_3), and the admixed population (N_2). For simplicity, we will assume that the effective population size stayed constant in each population (e.g., $N_0=N_1=N_2=N_3=N$). While this assumption may not hold for all data, it is supported for non-African human populations, which have been found to have approximately the same effective population sizes [16], [17]. Furthermore, as we demonstrate in Results, the method can still give accurate results when effective population sizes do not vary greatly. Given this assumption, the effective population size, N, and mutation rate, μ will be aggregated with the length of the sequences, l, as a single parameter θ. As a result, the free parameters Θ we must learn are t_1, t_2, α, and θ.

Given the admixture model, we would expect local regions of the genome to each have a tree-like ancestral history, but with different histories in different regions sampled from a network of possible ancestral relationships implied by the divergence and admixture events. A tree-based history corresponding to a local, non-admixed region of the genome is known as a genealogy. For example, at some regions of the genome, we would expect to see a genealogy of the three samples derived from Fig. 1(b) while other regions would have a genealogy derived from Fig. 1(c). If we suppose $\alpha = 0.5$ then we should see these two genealogies with approximately equal frequency across the genome.

Given the sequence data derived from the admixture scenario, our approach will first learn that there are three subpopulations in the example dataset using an algorithm developed in our previous work [18] for the problem of reconstructing population histories, which describe the historical emergence of population subgroups in a broader population, from non-admixed data. The algorithm will also learn a summary description of the data that assigns mutations to biparititions between population subgroups. In the example of Fig. 1, this summary description would suggest that approximately 1 mutation occurred in the genetic region under study after P_2 was formed (branch e_d in Fig. 1(d)), that approximately 2 mutations occurred either in P_1 after P_2 was formed or in P_3 before P_2 was formed (branches e_b and e_c in Fig. 1(d)), and that approximately 2 mutations occurred either in P_3 after formation of P_2 or in P_1 before P_2 (branches e_a and e_e in Fig. 1(d)). Using these inferences, the next step would be to estimate the distribution of the posterior probability of the event times and admixture proportions that best describe the data.

Learning Summary Descriptions

Our previous work on learning population histories from non-admixed variation data [18] is conceptually based on the idea of consensus trees [19], which represent inferences as to the robust features of a family of trees. The algorithm uses the genetic variation dataset to infer a set of local phylogenetic trees from small consecutive regions across the genome. It then breaks each tree into a set of bipartitions, where each bipartition corresponds to one edge in one tree whose removal divides the taxa labeling nodes into two groups (see Fig. 1(f)). From the set of bipartitions, the algorithm then identifies a set of model bipartitions, robust splits between population groups that define an inferred overall population history so as to minimize an information-theoretic minimum description length score [20].

The intuition behind our method is that different regions in the genome should correspond to different genealogies embedded within the overall population structure. By first inferring likely phylogenies on many small regions spanning the genome and learning the robust features of the phylogenies, the algorithm specifically builds a summary description $H = (B^M, W)$ consisting of a set of model bipartitions, $B^M = \{b_1^M, b_2^M, \ldots b_r^M\}$, and a set of weight values, $W = \{w_0, w_1, w_2, \ldots, w_r\}$. Weights $w_1, \ldots, w_r$ are each associated with a model biparition while weight w_0 provides an additional count of observed bipartitions unassigned to any model bipartition. The weights, $w_1, \ldots, w_r$, are computed by counting the number of observed bipartitions optimally assigned to each corresponding model bipartition using an entropy-based scoring function described in our prior work [18] that matches each observed bipartition to its most similar model bipartition or to no bipartition if there is no sufficiently close match. When none of the model bipartitions is a good assignment for the observed bipartition, the bipartition is then assigned to a empty bipartition and attributed to the weight w_0. By matching the observed to model bipartitions, we indirectly estimate the approximate number of mutations that most likely occurred along any given branch in the population history. This set of model bipartitions and its associated weights are then used to reconstruct the evolutionary model.

Under the described admixture scenario, our consensus-tree based algorithm should first identify that there are three subpopulations ($K = 3$) in the data. Second, the algorithm should

output an inferred model bipartition set $B^M=\{b_1^M=P_1|P_2P_3, b_2^M=P_2|P_1P_3, b_3^M=P_3|P_1P_2\}$. Finally, the algorithm should produce a weight vector $W = \{w_0, w_1, w_2, w_3\}$, representing the number of observed bipartitions most likely represented by none of the model bipartitions versus model bipartitions b_1^M, b_2^M, or b_3^M. The method can also predict which of the populations is likely admixed, as the two model bipartitions having the largest weights should represent the two parental populations, P_1 and P_3.

Likelihood Model

Under the two-parental, one-admixed population scenario, learning the directed graph $G = (V, E)$ of ancestry relationships among populations and its label function from the outputs of consensus tree algorithm could be trivially accomplished by associating the model bipartition of highest weight to the divergence between the two non-admixed populations. This would leave us with just the real-valued parameters Θ to infer. To make inferences about the parameter set Θ, we will estimate the distribution of the posterior probability of the parameters given the observed weights W associated with the model bipartitions. We note that in the absence of recombination and assuming an infinite sites model, the number of mutations corresponding to an edge of the genealogy would be Poisson distributed with mean equal to the product of the sum of all branch lengths in the genealogy l_G, the effective population size N, the number of base pairs l in the segment, and the mutation rate μ. We then break down the genealogy into a set of bipartitions corresponding to the edges of the genealogy. For each bipartition b, we determine an assignment $f(b)$ of b either to a model bipartition or to no bipartition so as to optimize the conditional entropy of the assignments. This assignment procedure is described in detail in our prior work [18]. If l_{b_j} is the branch length of the bipartition b_j, then the total branch length $l_{b_i}^M$ that will be assigned to model bipartition b_i^M is given by $l_{b_i^M}=\sum_{b_j\in\{b|f(b)=i\}} l_{b_j}$. This formula gives us an estimated amount of time over which a mutation could have occurred in the genealogy on the ith model bipartition, specifying an independent Poisson distribution for each w_i in that genealogy.

Because of recombination, however, the entire genome is made up of non-recombinant fragments of DNA having different genealogies. Since we do not know the actual genealogy for each fragment of the genome, the likelihood function will have to sum over all possible genealogies. Let $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ be the set of n genealogies each representing a genealogy of a non-recombinant fragment on the genome. Then the likelihood function $\mathcal{L} = P(W|\Theta)$ will be:

$$P(W|\Theta)=\prod_{i=0}^{3}\int_{l_{b_i^M}=0}^{\infty}\sum_{\mathcal{G}}P(w_i|\Theta,l_{b_i^M})P(l_{b_i^M}|\mathcal{G},\Theta)P(\mathcal{G}|\Theta)dl_{b_i^M} \quad (1)$$

where $P(w_i|l_{b_i^M},\theta)=\mathrm{Poisson}(w_i;\theta\times l_{b_i^M})$.

The branch length associated with each model bipartition can be computed exactly given the genealogy set. The integral can then be eliminated, as $P(l_{b_i^M}|\mathcal{G},\Theta)$ becomes zero for any

branch length not consistent with the genealogy and one for any branch length consistent with the genealogy. Hence, the likelihood function simplifies to:

$$P(W|\Theta)=\prod_{i=0}^{3}\sum_{\mathcal{G}}P(w_i|l_{b_i^M},\theta)P(\mathcal{G}|\Theta) \quad (2)$$

As an illustration, suppose the model population history is as shown in Fig. 1(d). If we have a particular parameter set for which we want to evaluate the likelihood function, we would enumerate over all possible genealogies consistent with the specified t_1, t_2, α, and θ. Suppose a genealogy in Fig. 1(e) was one possible genealogy being enumerated. We would evaluate the likelihood by converting the genealogy into a set of bipartitions as shown in Fig. 1(f) and subsequently compute the optimal assignment of each sampled bipartition to the most similar model bipartition by the minimum entropy criterion of [18]. Given the optimal assignment of each bipartition, we can then compute the expected branch lengths $l_{b_i^M}$ associated with the model bipartitions B_1^M, B_2^M, and B_3^M as well as the null bipartition. The optimal assignment in the example should give us expected branch lengths $l_{b_1^M}=l_1+l_2, l_{b_2^M}=l_3+l_4+l_7, l_{b_3^M}=l_5+l_6+l_8+l_9+l_{10}$, and $l_0 = 0$. Using the expected branch lengths and θ, we can then compute the expected number of mutations associated with each model bipartition and with null bipartition and thus the probability $P(w_i|l_{b_i^M})$. The likelihood model assumes that a correct parameter set for a given history will yield a set of bipartition weights that most closely matches the observed weights and thus yields the maximum likelihood score.

We know of no analytical solution to this function and the infinite number of possible genealogies prevents exhaustive enumeration. We therefore employ an MCMC strategy similar to that of [14] and [10] but differing in the details of the likelihood function to better handle large genomic datasets. MCMC sampling may require a large number of steps to accurately estimate the posterior of the likelihood function, so we make two simplifications that drastically reduce the number of steps needed to achieve convergence in exchange for a modest decrease in precision. First, we assume that the coalescence times are fixed at their expected values, rather than being exponentially distributed random variables, yielding a number of genealogies that is finite, although still exponential in *n*. We justify this approximation by noting that, in the limit of large numbers of fragments, the total branch length of the genealogy will converge on the mean implied by the coalescent process, making it a reasonably accurate assumption for a model such as ours designed to work with large genomic datasets.

To prove this, let $L_{tot,G}$ be a random variable representing the total branch length in a genealogy. Suppose we have *k* individuals in the sample, implying $k - 1$ coalescence events needed to reach a common ancestor. $L_{tot,G}$ would then be a function of the $k - 1$ random variables, $L_1, L_2, \ldots, L_{k-1}$, representing the time of each coalescent event relative to the previous coalescent event. Specifically, $L_{tot,G}=\sum_{j=2}^{k} jL_j$.

If we assume that the entire genome is made up of n non-recombinant fragments and that each fragment is relatively independent, then the total branch length of the entire genome $L_{tot,\mathcal{G}}$ would be the sum of n independent random variables $L_{tot,G}$.

$$L_{tot,\mathcal{G}}=\sum_{i=0}^{n} L_{tot,G_i}=n\left(\frac{1}{n}\sum_{i=0}^{n} L_{tot,G_i}\right) \quad (3)$$

Under the weak law of large numbers, the average of a large number of trials should be close to the expected value of each trial. Assuming a genome-wide count of variations represents a sufficiently large sample of an independent per-base mutation rate, we can approximate the above formula as follows:

$$L_{tot,\mathcal{G}} \approx nE(L_{tot,G})=n\left(\sum_{j=2}^{k} jE(L_j)\right) \quad (4)$$

The second approximation that we incorporate into the model is the reduction of the total genealogies from n to m. The intuition is that the total number of distinct genealogies from which lineages evolve (m) should be much less than the number of genetic sites typed (n). This approximation would follow, for example, from the assumption that recombination is sufficiently rare that nearby genetic regions usually have the same genealogy. If we set $m = n$, we would allow for an exact model in which each input genealogy could be distinct. While specifying $m << n$ independent genealogies allows for a possibility of error, we provide empirical evidence in Results to show that the actual increased error in practice is modest and that improvements in accuracy taper off quickly as we increase the number of genealogies. Making this second approximation, however, reduces the number of genealogies we must consider in evaluating the likelihood function to exponential in m rather than n, a much more manageable term when $m << n$.

Letting $\hat{\mathcal{G}}$ be the reduced set of genealogies, we derive the following simplified likelihood function given the two approximations:

$$P(W|\Theta)=\prod_{i=0}^{3}\sum_{\hat{\mathcal{G}}} P(w_i|l_{i}^{b_i^M},\theta)P(\hat{\mathcal{G}}|\Theta) \quad (5)$$

The above assumptions and the constraints on the parameters impose some constraints on the feasible genealogies. From time 0 to t_1, individuals from P_1 P_2, and P_3 can only coalesce with individuals within the same population. Let $m_{x,1}$, $m_{x,2}$, $m_{x,3}$ be the number of lineages that came from populations P_1, P_2, and P_3 respectively at time x. Then the ith coalescence point starting from time 0 to time t_1 going backward will have an expected coalescence time of $4N/((m_{0,y} - i + 1)(m_{0,y} - i))$ from the previous coalescence event. If the next coalescence time point is greater than t_1 then the waiting time until the next coalescence time point beyond that one will be sampled from t_1 rather than from the previous coalescence time point.

MCMC Sampling

To estimate the posterior probability distribution, we employ the Metropolis-Hastings algorithm. We define the state space of the Markov model as the set of all parameters t_1, t_2, α, θ and the set of possible genealogies $\hat{\mathcal{G}}$ spanning the genome, where $|\hat{\mathcal{G}}| = m$. Furthermore, given specific values of t_1 and t_2, the genealogy set $\mathcal{G}$ can only contain genealogies consistent with those values of t_1 and t_2. For any state $X_o=\{x^o_{t_1}, x^o_{t_2}, x^o_{\alpha}, x^o_{\hat{\mathcal{G}}}\}$ the likelihood of that state can be expressed as:

$$P(X_o|W) \propto P(W|X_o)= \left(\prod_{i=0}^{3} P(w_i|l_{b_i^M}) P(l_{b_i^M}|x^o_{\hat{\mathcal{G}}})\right) P(x^o_{\hat{\mathcal{G}}}|x^o_{t_1}, x^o_{t_2}, x^o_{\alpha}) \quad (6)$$

To identify a candidate next state X_n, the algorithm will sample new values of t_1, t_2, α, and θ from independent Gaussian distributions with $\mu^o_{t_1}=x^o_{t_1}$, $\mu^o_{t_2}=x^o_{t_2}$, $\mu^o_{\alpha}=x^o_{\alpha}$, and $\mu^o_{\theta}=x^o_{\theta}$, and σ_{t_1}, σ_{t_2}, σ_{α}, and σ_{θ}, using variances adjusted during the burn-in period by increasing variance when the expected number of mutations is far from the observed number and decreasing variance as the expected and observed numbers of mutations become more similar. We developed this strategy based on the observation that acceptance rate tends to be better for large variances when the difference between the expected and observed number of mutations is large and better for small variances when the difference between the expected and observed numbers of mutations is small.

Once the algorithm selects values of parameters for the new MCMC state X_n, it then samples a new genealogy set through coalescent simulation given the selected new parameters. The resulting new state will thus have a stationary probability

$$Q(X_n|X_o)=P(x^n_{t_1}|\mu^o_{t_1}, \sigma_{t_1}) P(x^n_{t_2}|\mu^o_{t_2}, \sigma_{t_2}) \times P(x^n_{\alpha}|\mu^o_{\alpha}, \sigma_{\alpha}) P(\hat{\mathcal{G}}|x^n_{t_1}, x^n_{t_2}, x^n_{\alpha}) \quad (7)$$

yielding a Metropolis-Hastings acceptance ratio r of:

$$r=\frac{\left(\prod_{i=0}^{3} P(w_i|l_{b_i^M}) P\left(l_{b_i^M}|x^n_{\hat{\mathcal{G}}}\right)\right)}{\left(\prod_{i=0}^{3} P(w_i|l_{b_i^M}) P\left(l_{b_i^M}|x^o_{\hat{\mathcal{G}}}\right)\right)} \quad (8)$$

3 Validation Experiments

Coalescent Simulated Data

We evaluated our method on simulated datasets generated using different t_1, t_2, α, and chromosome lengths. Each simulated dataset consisted of 100 chromosomes from each of the three hypothetical populations (P_1, P_2, and P_3) resulting in a total of 300 chromosomes. We divided the simulated datasets into three groups consisting of chromosomes with 3.5×10^7 base pairs, 3.5×10^6 base pairs, and 2.0×10^5 base pairs. For each group, we generated 45 different datasets from all combinations of t_1 = {400, 800, 1200, 2000, 4000}, t_2 = {6000, 8000, 20000}, and α = {0.05, 0.2, 0.6}. We chose the coalescence simulator MS [21] for generating the simulated datasets. In all of our simulations, we assumed the effective

population size of each population is 10,000. We set the mutation rate to be 10^{-9} per base pair per generation and the recombination rate to be 10^{-8} per generation for simulations, based on estimated human mutation and recombination rates [22], [23]. Using the parameters described above, the simulations generated approximately 50 to 120, 1000 to 2000, and 10,000 to 20,000 SNPs on datasets with 2.0×10^5-, 3.5×10^6-, and 3.5×10^7-base sequences, respectively.

To evaluate the performance of our algorithm, we compared our results obtained from the simulated data with those of another method for learning admixture fractions and divergence times: *MEAdmix* [13]. *MEAdmix* takes as input a set of sequences of genetic variations from individual chromosomes grouped into three different populations and outputs the admixture fraction, divergence time, admixture time, and mutation rates from the input data. While *MEAdmix* produces similar outputs to *CLEAX*, one key difference between *MEAdmix* and *CLEAX* is the specification of populations. In *MEAdmix*, individual sequences must be assigned by the user to one of the three populations. On the other hand, *CLEAX* infers the populations directly from the variation data before estimating the divergence time and admixture fraction. Although there are a number of methods in the literature for learning admixture and divergence times [10], [13], [14], we chose to compare to *MEAdmix* because it estimates similar continuous parameters to *CLEAX* and its software is freely available. The same characteristics apply to *lea*, but it was unsuitable for the present comparison because it is designed for much smaller datasets and proved unable to process even the smallest models of genome-scale data we considered. Other methods were also investigated [10], [15], but we could not directly compare their performance to our own because of different admixture models assumed, different estimated parameters, or lack of availability of the software for comparison.

We ran both *CLEAX* and *MEAdmix* on the $S = 135$ simulated datasets and computed the average absolute relative difference between the true and estimated parameter values for each parameter, $\frac{1}{S}(\sum_i^S \frac{|\hat{\Theta}_i - \Theta_i|}{\Theta_i})$. We terminated a program on a given data set if the analysis took more than 48 hours to complete. When running our method on simulated data, we set the number of genealogies for *CLEAX* to be m=30. For *MEAdmix*, we set the bootstrap iterations to be five, which proved to be a practical limit for the mid-size data sets given the run time bounds.

We also evaluated the accuracy of our algorithm as a function of the number of genealogies, m. Using the same 45 simulated datasets with $t_1 = \{400, 800, 1200, 2000, 4000\}$, $t_2 = \{6000, 8000, 20000\}$, and $\alpha = \{0.05, 0.2, 0.6\}$ obtained from simulations using 3.5×10^6 base pairs, we ran our method with 10, 30, and 100 genealogies. For each genealogy size, we repeated the Markov chain ten times with different starting points and computed the average absolute relative difference between the estimated parameters and true parameters. Each MCMC run used 1,000 iterations of burn-in followed by 20,000 MCMC steps.

In addition to evaluating our algorithm under scenarios in which the effective population size remains fixed, we also examined the performance under scenarios in which this assumption no longer holds in order to explore a possible source of error in the analysis of

real data. To evaluate the performance of the method under scenarios for which effective population size is not constant, we generated four additional sets of simulated data consisting of the same values of admixture time (t_1), divergence time (t_2), and admixture fraction (α) as in previous experiments but with a reduced effective population size for all three populations after the admixture event occurs. Specifically, prior to time t_1, the effective population size is assumed to be 10,000 as in our other simulated data sets. From t_1 to the present time, though, the effective population size of all three populations is reduced to 2,000, 4,000, 6,000, or 8,000. Using the original data and the additional four groups of 45 simulated datasets, we evaluated the performance of the algorithm by the average absolute difference between the true and estimated parameter values within each group. Additionally, we computed the ratio of t_1 to t_2 across all 45 datasets in order to test whether one could get accurate estimates of both times if a single "anchor" time was already known.

Real SNP Data

We further evaluated our method by applying it to a bovine SNP dataset [24], chosen due to the limited availability of large-scale human genetic variation data containing known admixed individuals. The bovine data consists of 497 cattle from 19 breeds. Of the 19 different breeds of cattle, 3 of them are indicine (humped), 13 of them are taurine (humpless), and the rest are hybrids of indicine and taurine. Because the dataset has more breeds than the supported admixture model, we filtered the dataset until only one hybrid population and two non-admixed populations remained. In particular, we selected a total of 76 cattle as our input dataset: 25 Brahman, 27 Hereford, and 24 Santa Gertrudis. The Brahman are a breed of taurine, the Hereford a breed of indicine, and the Santa Gertrudis a cross between Shorthorn and Brahman with an approximate mixture proportion of five-eighths Shorthorn and three-eighths Brahman. Because the dataset did not include the Shorthorn cattle, we used the Hereford as a representative of the Shorthorn since they are closely related to the Shorthorn breeds. Given the filtered bovine data, we tested our algorithm on 2,587 SNP sites genotyped from chromosome 6.

We then tested our method on a human data set from 1,000 Genomes Project Phase I release version 3 in NCBI build 37 [25]. The dataset consisted of 1,092 individuals from a number of different ethnic backgrounds that can largely be grouped into four different continents of origin: Africa, Europe, Asia, and America. Of the 1,092 individuals sequenced, 246 have African ancestry from Kenya, Nigeria, and Southwest US. 379 individuals have European ancestry from Finland, England, Scotland, Spain, Italy, and Utah. 286 individuals have Asian ancestry from China and Japan. The remaining 181 individuals from America consist mainly of admixed individuals from Mexico, Puerto Rico, and Columbia. Similarly to the bovine dataset, we filtered the dataset until only one admixed population and two parental populations remained by removing the 246 individuals having African ancestry. Due to computational limitations, we ran our algorithm on a uniformly selected subsample of 150,000 variant sites across the whole genome.

In addition to positive validations, we also performed a negative control for our method on a human data set for which no appreciable admixture is known to occur. We used the Phase II HapMap data set (phased, release 22) [26] which consists of over 3.1 million SNP sites

genotyped for 270 individuals from four populations: 90 Utah residents with Northern and Western Europe ancestry (CEU); 90 individuals with African ancestry (YRI); 45 Han Chinese (CHB); and 45 Japanese (JPT). For the CEU and YRI groups, which consist of trio data (parents and a child), we used only the 60 unrelated parents. Although the HapMap dataset does not contain known admixed populations, the dataset allows us to evaluate the method's ability to learn the divergence time between populations. In addition, it serves as a useful negative control for detecting admixture. For the HapMap dataset, we tested our algorithm on all 50,556 SNPs collected from chromosome 22.

For all three datasets, we set the number of genealogies m to be 30 for these tests. We did not evaluate the real datasets using *MEAdmix*, as the number of segregating sites in the real dataset exceeded the software's limitations. As with the simulated datasets, we used 1,000 steps in the burn-in period followed by 20,000 MCMC steps. We ran 10 independent copies of each chain for bovine and HapMap data and 50 for 1,000 Genomes data to minimize the risk of poor sampling due to a chain becoming stuck in local optima.

4 Results

Coalescent Simulated Data

Figure 2(a) shows the estimated α computed by *CLEAX* using 10, 30, and 100 genealogies and by *MEAdmix* on the 3.5×10^6-base sequences. Estimations of α by *CLEAX* tend to improve as we increase the number of genealogies. When comparing results to *MEAdmix*, estimations of α by *CLEAX* generally have a slight edge over *MEAdmix* using 30 and 100 genealogies. The major exceptions are data with large t_1 (4000 generations) and small t_2 (6000 generations). The advantage of *CLEAX* is less consistent when using only 10 genealogies. Mean and 95% confidence interval estimations of α by *CLEAX* also tend to improve as we increase the number of genealogies. The two methods are about equally likely to cover the true α within the confidence interval, but *CLEAX* tends to have a smaller confidence interval, especially when run with 30 or 100 genealogies. While *MEAdmix* does not show any obvious trend as we vary parameters, *CLEAX* tends to do better on sequences with small t_1 and large t_2.

Estimates of t_1 (Figure 2(b)) and t_2 (Figure 2(c)) show similar trends to α. As with α, mean estimations by *CLEAX* tend to be closer to the true values than those of *MEAdmix* in the majority of cases. Mean and 95% confidence interval estimations of t_1 and t_2 again improve for *CLEAX* as we increase the number of genealogies. Confidence intervals estimated by *CLEAX* are wider than those for *MEAdmix* for these parameters, but more often covered the true parameters.

Aggregate quantitative performance is shown in Table 1, which provides the average absolute relative difference between the estimated parameters and true parameters computed by the algorithm for different lengths of simulations, $(\frac{|\hat{\Theta} - \Theta|}{\Theta})$. For datasets with 3.5×10^6-base sequences, *CLEAX* has a worse average relative difference between estimated and true t_2 and α parameters when we set the number of genealogies to be 10, but better average

relative difference for t_1. When we increase the number of genealogies to 30 or more, *CLEAX* yields more accurate estimates for all three parameters than did *MEAdmix*.

We next examined performance on smaller sequences of 2.0×10^5 bases (approximately 50 to 120 SNPs), to test scaling of the methods to sub-genomic scale data. For these sequences, our program is unable to automatically identify the three major population groups. Instead, it identifies only the divergence into subpopulations P_1 and P_3. We attribute this failure to the small number of SNPs providing insufficient evidence for the existence of a separate admixed subpopulation P_2. Since *MEAdmix* depends on the user to perform this assignment of population groups, we manually performed the comparable assignment for our program in order to test just assignment of continuous parameters in this low-data scenario. For these data, both methods again perform comparably to one another at estimating α, with *MEAdmix* showing slightly lower mean and standard deviation in errors. Compared to the 3.5×10^6-base data, both methods show substantially worse α estimations, with approximately a three-fold increase in mean error. Estimates of t_1 and t_2 on the smaller dataset also show substantially worse performance for both methods. As seen in Table 1, *CLEAX* is worse in estimating t_1 and t_2 under these conditions, likely because the assumptions of our simplified likelihood model are valid only in the limit of large numbers of segregating sites and thus yield more pronounced inaccuracy on short sequences. Both programs, however, do worse on this small dataset than on the larger ones.

We next examined scaling to larger (genomic-scale) data sets by testing on simulated data of 3.5×10^7 bases. *MEAdmix* did not report any progress on any of these data sets after 48 hours of run time, and so results are reported only for *CLEAX*. As Table 1 shows, accuracy of the three estimated parameter is improved relative to the smaller datasets, with roughly 35%, 1%, and 6.5% improvements for t_1, t_2, and α for $m = 30$.

We also examined the average running times for these data sets. *CLEAX* with $|\hat{\mathcal{G}}| = 30$ required 1.27 hours, 1.94 hours, and 7.61 hours, respectively, for the 2.0×10^5-, 3.5×10^6-, and 3.5×10^7-base data sets. *MEAdmix* required 2.8 hours for the 2.0×10^5-base data set and 6.2 hours for the 3.5×10^6-base data set, while making no apparent progress in 48 hours on the 3.5×10^7-base data set.

To understand the effect of varying effective population size on the performance of the algorithm, we evaluated our method on datasets with reduced effective population size after admixture events. Figure 3 shows the average absolute difference between the estimated and the true parameter values across different reduced effective population sizes after admixture. Across all parameters, the average absolute difference between the estimated and true parameter values increases as the effective population size decreases. For α, we observe a modest change in the absolute difference between the estimated and true parameter values from 0.04 when the effective population remains constant to 0.10 when the effective population size is reduced to 20% of the original size. Estimates for t_1 and t_2, on the other hand, are significantly affected as we decrease the effective population size. For both t_1 and t_2, average absolute difference increases roughly 100-fold as we decrease the effective population to 20% of the original size after admixture. This suggests that estimation of α

would be less likely to be affected by fluctuation of effective population size throughout history.

We next examine the performance of the method under varying effective population sizes by plotting the estimated t_1/t_2 ratio against true t_1/t_2 ratio. This allows us to determine if the estimation of the time can be corrected when effective population size is drastically changed by anchoring one time point using external information. Figure 4 shows the t_1/t_2 ratio for different effective population sizes. Aside from the datasets where the effective population size drops to 20% of the original size, most of the estimates maintain ratios close to one, suggesting that errors induced by changes in effective population size can be effectively corrected if additional partial data is available fixing one of the two times.

Real SNP Data

Figure 5(a) shows the smoothed probability density distribution, the mean, and 95% confidence interval of each parameter value for the bovine dataset. Each gray line in the figure represents the smoothed probability distribution from one independent run of the Markov chain. All ten runs of the chain on the bovine data yielded consistent probability distributions. The estimated mean admixture proportion for the bovine dataset is 41.6 percent Brahma and 58.4 percent Hereford. The 95% confidence interval for admixture proportion α is between 32.2 percent and 50.6 percent. The mean estimate of divergence time (t_2) is about 28,000 generations. Assuming 7 years per generation for cattle, the divergence time would translate to approximately 195 kya (thousand years ago), consistent with the belief that the *indicine* and *taurine* diverged approximately 250 kya [24]. Admixture time (t_1) is estimated to be approximately 6 kya with ranges between 3.5 kya to 8.5 kya. This range is likely an overestimate of the true value since artificial breading of the hybrid did not become common until the past 100 years [24]. The mean estimate of $\theta = l \times N \times \mu$ is 36.1. If we assume the effective population size is 2000 based on the estimated ancestral effective population size [24] then the mutation rate would be approximately 2.0×10^{-10} base per site per generation, a much lower estimate than is supported by the prior literature [23], [27]. Using an estimated effective population size of 107 [24], a more consistent estimate of effective population sizes after a recent population bottleneck derived by averaging the recent effective population sizes of the three breeds, would yield a more realistic mutation rate of 2.8×10^{-9} [23]. Inaccuracy in the rate might also be due to ascertainment bias or the incomplete detection of the mutations at the sequencing phase.

Figure 5(b) shows distributions of *CLEAX* estimates for the 1,000 Genomes Project data. The method interprets the American group, consisting of individuals from Mexico and Puerto Rico, as admixed from the Asian and European groups. *CLEAX* inferred an average of 9% admixture from the Asian group and 89% from the European group. The admixture fractions α from different chains are most concentrated around 0.05. Six out of 50 chains, however, appear likely to have become stuck in local optima, with values of approximately 0.3 for five chains and 0.6 for another. While the mean estimate is slightly lower than is found in prior work [28], [29], [30], the 95% confidence interval overlaps estimates from the prior literature. The mean estimate of the admixture time t_1 was 48 generations with a 95% confidence interval of 17 to 150 generations. Assuming 20 years per generations, this would

translate to approximately 960 years ago with a 95% confidence interval ranging from 340 years ago to 3,000 years ago. This range is somewhat higher than the 200–500 year ago estimate by Tang *et al.* [29] but with some overlap. The mean divergence time t_2 was estimated to be 161 generations ago with a 95% confidence interval of 74 to 447 generations ago. Using the same assumption of 20 years per generations, this would translate to approximately 4,800 years ago and a 95% confidence interval of 1,500 years to 9,500 years ago, a range consistent with that of Garrigan *et al.* [31] although more recent than that of Zhivotvosky *et al.* [32].

Figure 5(c) shows the probability distribution for the HapMap Phase II data. As with the bovine dataset, there is a generally high consistency across the ten runs in the parameter estimates. For the HapMap Phase II data, *CLEAX* estimated α to be less than 1% with a 0% to 6% confidence interval. The mean divergence time (t_2) was estimated to be about 4,000 generations. Assuming 20 years per generation, the estimated divergence time of Europeans (CEU) and Africans (YRI) would be around 80 kya with a confidence interval between 57.6 kya and 106 kya. The divergence time (t_1) between Europeans (CEU) and East Asians (CHB +JPT) has a mean estimate of 26.1 kya and a confidence interval between 18.9 kya and 33.6 kya. The mean estimate of θ is 4, 320. Assuming the effective population size of human population to be 10,000 [33], the implied mutation rate would be 2.16×10^{-9} per site per generation, similar to prior estimates [23], [27].

5 Discussion

In this paper, we propose a method to learn admixture proportions and divergence times of admixture events from large-scale genetic variation data. Prior coalescent-based methods for estimating such parameters have been proposed in recent years, but such methods tend to be computationally costly and poorly suited to handling genomic-scale data. Our new method provides comparable estimates of admixture proportions to the prior art on smaller datasets while scaling to much larger data sets with increasing accuracy. Although the average errors for t_1 and t_2 were worse than those of *MEAdmix* for datasets with 2.0×10^5-base long sequences, we observed a general improvement in *CLEAX* estimates over *MEAdmix* as we increased the length of the input datasets. Our method also provides much better time estimates than *MEAdmix* on larger datasets, yielding average t_1 and t_2 estimation errors roughly two-thirds of those of *MEAdmix* for chromosome-scale data. The poor performance on short sequences may be due to the assumption that coalescence times in the genealogies are fixed, an assumption whose validity breaks down in the limit of small numbers of variant sites.

Variance between true and estimated parameter tends to be high for datasets with shorter sequences, as evident in Table 1, but decreases as we increase the length of the sequences. We expect the variance to continue to reduce further as we use longer sequences. Our method thus appears to be a poorer choice on older, gene-scale data than prior methods, but a clear improvement on datasets comparable in size to human chromosomes.

The performance of *CLEAX* also tends to improve as we increase the number of genealogies, $|\hat{\mathcal{G}}|$, used to estimate the expected branch length. While the estimates of α by *CLEAX* are

worse than those of *MEAdmix* when $|\hat{\mathcal{G}}|$ is set to 10, the results are better than those of *MEAdmix* for $|\hat{\mathcal{G}}| = 30$. Results showed little improvement upon further increase of $|\hat{\mathcal{G}}|$ to 100, suggesting that a relatively small number of genealogies is adequate to closely approximate the true likelihood function.

Results on the real datasets provide further confidence in the method, yielding estimates of divergence times and admixture fractions generally consistent with the current literature [24], [32], [34]. Using the HapMap Phase II dataset, our method's estimation of the YRI-CEU divergence time between 76.5 kya to 89.6 kya is consistent with the STR estimation of [32] (62–133kya) and the HMM estimation of [11] (60–120 kya). Estimation of little or no admixture fraction between the CHB+JPT and CEU is also consistent with the general belief that negligible admixture has occurred between the major human populations. Our estimates of the divergence time between Asians and Europeans of 23.0 kya to 33.6 kya for HapMap are similar to estimates by Gutenkunst *et al.* [34]. Estimates of the divergence time between Asians and Europeans from the 1,000 Genomes data are also similar to the estimates from HapMap and consistent with estimates from Garrigan (7–13kya) *et al.* [31], albeit with a slightly more recent range. While the mean estimate of admixture time for the American group was somewhat higher than expected (980 years), the lower bound of 340 years ago is reasonable. The admixture fraction estimate for the American group is also consistent with existing literature [29], [30].

Similarly, using the bovine dataset, estimates of divergence time and admixture fraction were also consistent with the general consensus [24]. One discrepancy in the bovine dataset was an unrealistically high estimate of admixture time (6,000 years). One plausible source of error is the algorithm's assumption of fixed effective population size. Because there is believed to have been a drop of effective population size to a few hundred cattle in recent years [24], [35], the decrease in effective population size would increase the chance that cattle share a most recent common ancestor at a much earlier time. As a result, more mutations that occurred before the admixture time will be miscategorized as mutations that occurred after the admixture time, resulting in a bias in estimated admixture time. This observation may suggest that our method in current form is poorly suited to estimating admixture times on data with significant changes in effective population size over time. Our analysis of simulated data, however, suggests that estimates of admixture fractions should remain accurate despite changes in effective population size. The discrepancy could also be attributed to the difference between the Hereford and Shorthorn breeds, where the mutations over-represented in the hybrid population that led to the long estimates of time since admixture could actually have been misattributed mutations between the Hereford and Shorthorn breeds.

When we examine the results of our method on simulated data, we observe generally worse performance with increasing admixture time, especially for simulations with low admixture proportions. Such a phenomenon is likely caused by the fact that there are fewer lineages at the admixture time as we increase the admixture time. For example, for simulations with admixture time t_1 of 4,000, we would expect roughly 10 lineages left by the time the admixture event occurred, preventing the method from inferring admixture proportions at a resolution of better than 10%. Consequently, fewer lineages at the admixture time would

increase the variance of the admixture fraction estimate. This observation suggests that our method will work better at analyzing more recent admixture.

The effects of varying effective population size on inference accuracy suggest that estimates of times of divergence and admixture is sensitive to changes in effective population size but that such changes have only a modest effect on the admixture fraction estimation. This observation suggests that estimates of the admixture fraction should be considered more reliable than estimates of divergence and admixture time when one suspects effective population has changed drastically over time. Time estimates were within an order of magnitude when the change in effective population size was up to 40%, suggesting estimates could still be trusted if changes in effective population size are modest. Furthermore, estimates of the ratio between t_1 and t_2 seem to be accurate even when effective population size changes significantly. Despite poor estimates of time when effective population size changes drastically, we could potentially correct time estimates using the ratios if we could anchor at least one time point using external data sources or prior knowledge.

Despite some of the shortcomings of the algorithm, our method nonetheless has demonstrated its capability in estimating accurate parameters on long sequence datasets. While our MCMC strategy is similar to a number of prior approaches [10], [14], our algorithm is distinguished by novel strategies for simplifying the likelihood model in ways especially suited to genomic-scale variation data sets, trading off increases in performance that are substantial for long sequences with decreases in accuracy that are modest under the same circumstances. Our method also has the unique feature of automatically inferring the population substructure, history of formation of that structure, and likely admixture model in a single unified inference, allowing it to take advantage of the fact that each aspect of that inference is dependent on the answers to the other two. Although our method currently only estimates divergence times and admixture fractions for a standard three-population single-admixture scenario, the approach establishes a method for assigning likelihoods to admixture events and sampling over parameters for these events that could in principle be used as a module for considering more complicated scenarios potentially involving larger numbers of populations or multiple admixture events.

Acknowledgments

This work was supported by US NSF award IIS-0612099. M.-C.T. was additionally supported by training grant T32 EB009403 and R.S. by US NIH awards 1R01CA140214 and 1R01AI076318.

REFERENCES

1. Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, Shriver MD. Estimating african american admixture proportions by use of population-specific alleles. American Journal of Human Genetics. 1998; 63(no. 6):1839–1851. [PubMed: 9837836]
2. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. Timing the ancestor of the hiv-1 pandemic strains. Science. 2000; 288(no. 5472):1789–1796. [PubMed: 10846155]

3. Goldstein DB, Chikhi L. Human migrations and population structure: What we know and why it matters. Annual Review of Genomics and Human Genetics. 2002; Vol. 3(no. 1):129–152.
4. Dupanloup I, Bertorelle G, Chikhi L, Barbujani G. Estimating the impact of prehistoric admixture on the genome of europeans. Molecular Biology and Evolution. 2004; Vol. 21(no. 7):1361–1372. [PubMed: 15044595]
5. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo J-M, Wambebe C, Tishkoff SA, Bustamante CD. Genome-wide patterns of population structure and admixture in west africans and african americans. Proceedings of the National Academy of Sciences. 2010; Vol. 107(no. 2):786–791.
6. Franois O, Currat M, Ray N, Han E, Excoffier L, Novembre J. Principal component analysis under population genetic models of range expansion and admixture. Molecular Biology and Evolution. 2010; Vol. 27(no. 6):1257–1268.
7. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; Vol. 155(no. 2):945–959. [PubMed: 10835412]
8. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics. 2009; Vol. 5(no. 6):e1000519. [PubMed: 19543370]
9. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. American Journal of Human Genetics. 2008; Vol. 82(no. 2):290–303. [PubMed: 18252211]
10. Nielsen R, Wakeley J. Distinguishing migration from isolation: A markov chain monte carlo approach. Genetics. 2001; Vol. 158(no. 2):885–896. [PubMed: 11404349]
11. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011; Vol. 475(no. 7357):493–496. [PubMed: 21753753]
12. Chakraborty R. Gene admixture in human populations: Models and predictions. American Journal of Physical Anthropology. 1986; Vol. 29(no. S7):1–43.
13. Wang J. A coalescent-based estimator of admixture from dna sequences. Genetics. 2006; Vol. 173(no. 3):1679–1692. [PubMed: 16624918]
14. Chikhi L, Bruford M, Beaumont M. Estimation of admixture proportions: A likelihood-based approach using markov chain monte carlo. Genetics. 2001; Vol. 158(no. 3):1347–1362. [PubMed: 11454781]
15. Bertorelle G, Excoffier L. Inferring admixture proportions from molecular data. Molecular Biology and Evolution. 1998; Vol. 15(no. 10):1298–1311. [PubMed: 9787436]
16. Tenesa A, Hayes PNBJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. Recent human effective population size estimated from linkage disequilibrium. Genome Research. 2007; Vol. 17(no. 4):520526.
17. Mel M, Javed A, Pybus M, Zalloua P, Haber M, Comas D, Netea MG, Balanovsky O, Balanovska E, Jin L, Yang Y, Pitchappan R, Arunkumar G, Parida L, Calafell F, Bertranpetit J, Consortium TG. Recombination gives a new insight in the effective population size and the history of the old world human populations. Molecular Biology and Evolution. 2011
18. Tsai M-C, Blelloch GE, Ravi R, Schwartz R. A consensus tree approach for reconstructing human evolutionary history and detecting population substructure. IEEE/ACM Trans. Comput. Biol. Bioinformatics. 2011 Jul.Vol. 8:918–928.
19. Nei, M.; Kumar, S. Molecular Evolution and Phylogenetics. Oxford University Press; 2000.
20. Grünwald, P.; Myung, I.; Pitt, M. Advances in Minimum Description Length: Theory and Applications. The MIT Press; 2005.
21. Hudson R. Gene genealogies and the coalescent process. Oxford Surveys in Evolutionary Biology. 1990; Vol. 7:1–44.
22. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. Genome Research. 2003; Vol. 13(no. 1):13–26. [PubMed: 12529302]

23. Liu G, Matukumalli L, Sonstegard T, Shade L, Van Tassell C. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. BMC Genomics. 2006; Vol. 7(no. 1):140. [PubMed: 16759380]
24. The Bovine HapMap Consortium. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 2009; Vol. 324(no. 5926):528–532.
25. A map of human genome variation from population-scale sequencing. Nature. 2010; Vol. 467(no. 7319):1061–1073.
26. International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. Nature. 2007 Oct; Vol. 449(no. 7164):851–861.
27. Kumar S, Subramanian S. Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences. 2002; Vol. 99(no. 2):803–808.
28. Rangel-Villalobos H, Muoz-Valle JF, Gonzlez-Martn A, Gorostiza A, Magaa MT, Pez-Riberos LA. Genetic admixture, relatedness, and structure patterns among mexican populations revealed by the y-chromosome. American Journal of Physical Anthropology. 2008; Vol. 135(no. 4):448–461. [PubMed: 18161845]
29. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. Recent genetic selection in the ancestral admixture of puerto ricans. American Journal of Human Genetics. 2007; Vol. 81(no. 3):626–633. [PubMed: 17701908]
30. Martinez-Cortes G, Salazar-Flores J, Gabriela Fernandez-Rodriguez L, Rubi-Castellanos R, Rodriguez-Loya C, Velarde-Felix JS, Franciso Munoz-Valle J, Parra-Rojas I, Rangel-Villalobos H. Admixture and population structure in mexican-mestizos based on paternal lineages. Journal of Human Genetics. 2012
31. Garrigan D, Kingan SB, Pilkington MM, Wilder JA, Cox MP, Soodyall H, Strassmann B, Destro-Bisol G, de Knijff P, Novelletto A, Friedlaender J, Hammer MF. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, x and y chromosome resequencing data. Genetics. 2007; Vol. 177(no. 4):2195–2207. [PubMed: 18073427]
32. Zhivotovsky LA. Estimating divergence time with the use of microsatellite genetic distances: Impacts of population growth and gene flow. Molecular Biology and Evolution. 2001; Vol. 18(no. 5):700–709. [PubMed: 11319254]
33. Hammer MF. A recent common ancestry for human Y chromosomes. Nature. 1995; Vol. 378(no. 6555):376–378. [PubMed: 7477371]
34. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics. 2009; Vol. 5(no. 10):e1000695. [PubMed: 19851460]
35. Lee SH, Cho YM, Lim D, Kim HC, Choi BH, Park HS, Kim OH, Kim S, Kim TH, Yoon D, Hong SK. Linkage disequilibrium and effective population size in hanwoo korean cattle. Asian-Australasian Journal of Animal Sciences. 2011; Vol. 24(no. 12):1660–1665.

Biographies

Ming-Chi Tsai Ming-Chi Tsai received his BA in Computer Science and Molecular and Cell Biology from the University of California, Berkeley in 2003. Since 2007, he has been a PhD student at the Joint CMU-Pitt PhD Program in Computational Biology. His primary area of research is computational biology.

Guy Blelloch Guy Blelloch received a BA degree in Physics and a BS degree in Engineering in 1983 from Swarthmore College, and MS and PhD degrees in Computer Science from the Massachusetts Institute of Technology in 1986 and 1988, respectively. He is currently a Professor of Computer Science at Carnegie Mellon University and co-director of the ALADDIN center for the study of algorithms. His research interests are in programming languages and applied algorithms.

R. Ravi R. Ravi received his B. Tech. in Computer Science and Engineering from the Indian Institute of Technology, Madras in 1989 and a PhD in Computer Science from Brown University in 1993. After post-doctoral fellowships at UC Davis and DIMACS, Princeton University, he joined the Operations Research faculty at the Tepper School of Business at Carnegie Mellon University in 1995, where he is currently Carnegie Bosch Professor of Operations Research and Computer Science.

Russell Schwartz Russell Schwartz received his BS, MEng, and PhD degrees from the Department of Electrical Engineering and Computer Science at the Massachusetts of Technology, the last in 2000. He later worked in the Informatics Research group at Celera Genomics. He joined the faculty of Carnegie Mellon University in 2002, where he is currently a Professor of Biological Sciences.

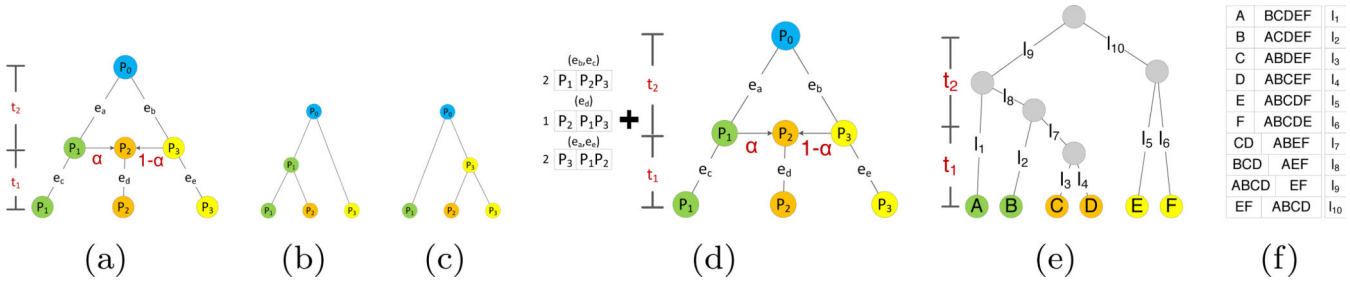

Fig. 1.

Example of a history of two parental populations (P_1 and P_3) and an admixed population (P_2). Ancestral population P_0 diverged at t_2 to form P_1 and P_3, followed by an admixture event at t_1 to form P_2. (a) The admixture model of the example. (b) Possible history of the example at some non-recombinant region of the genome with mutations occurring at various branches of the tree. (c) Alternative history of the example at other non-recombinant region of the genome with mutations occurring at various branches of the tree. (d) The desired output of the consensus tree algorithm applied to the genetic variation data, inferring the set of model bipartitions and its associated weights as well as a crude model of population history without the actual parameters. (e) Genealogy generated from parameters t_1, t_2, and α showing a possible ancestry of all taxa, including branch lengths. Here, AB is in P_1, CD is in P_2, and EF is in P_3. (f) The corresponding bipartitions and associated branch lengths obtained from the genealogy in (e).

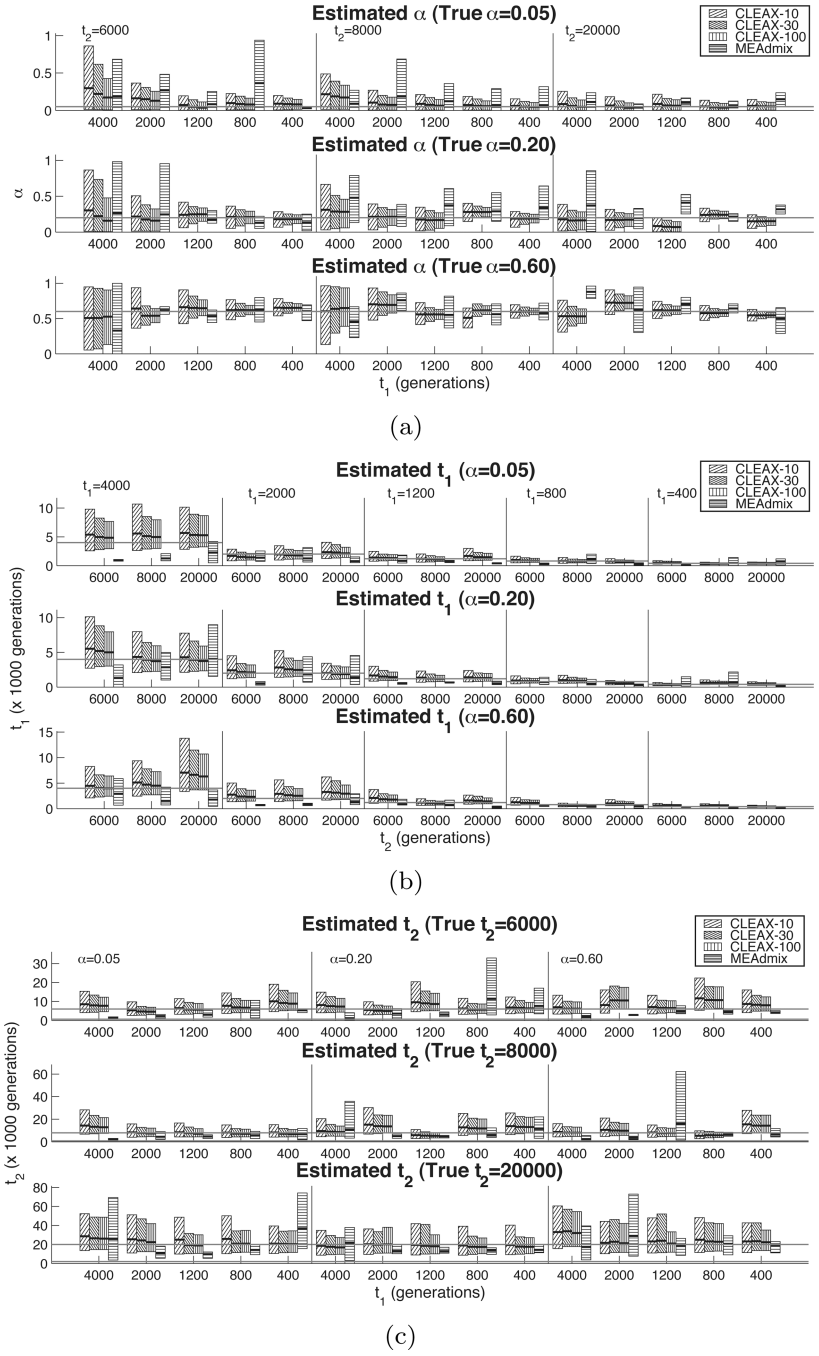

Fig. 2.

Mean and 95% confidence interval of the estimated parameters on 3.5×10^6-base sequences. The different bars represent the means estimated by *CLEAX* using 10, 30, and 100 genealogies (left) and by *MEAdmix* (right). Solid gray horizontal bars represent true parameter values used for the simulated data. (a) Estimated α organized into three rows of distinct true α values and grouped vertically by true t_2. (b) Estimated t_1 in generations organized into three rows of true α and grouped by true t_1. (c) Estimated t_2 in generations organized into three rows of true t_2 and grouped by true α.

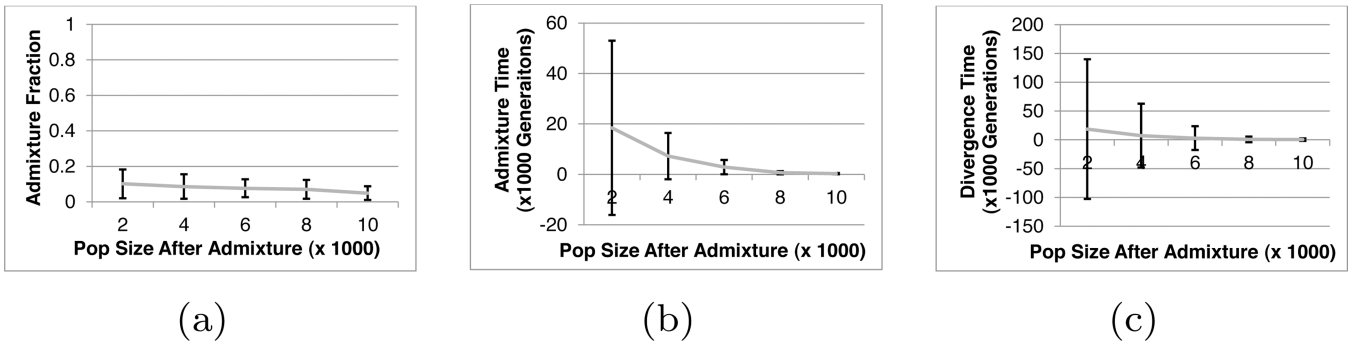

Fig. 3.

Plot of the mean and standard deviation of the average absolute difference between the estimated and true parameter values when the effective population size changes from 10000 to 2000, 4000, 6000, 8000, and 10000. (a) Plot of the average absolute difference between the estimated α and the true α. (b) Plot of the average absolute difference between the estimated t_1 and the true t_1. (c) Plot of the average absolute difference between the estimated t_2 and true t_2.

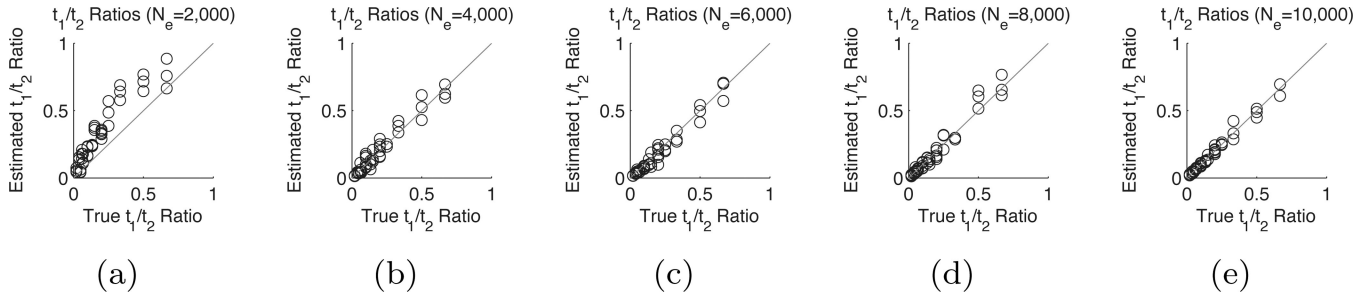

Fig. 4.

Plot estimated t_1/t_2 ratio against true t_1/t_2 ratio from datasets when the effective population size changes from 10000 to 2000, 4000, 6000, 8000, and 10000 (a–e).

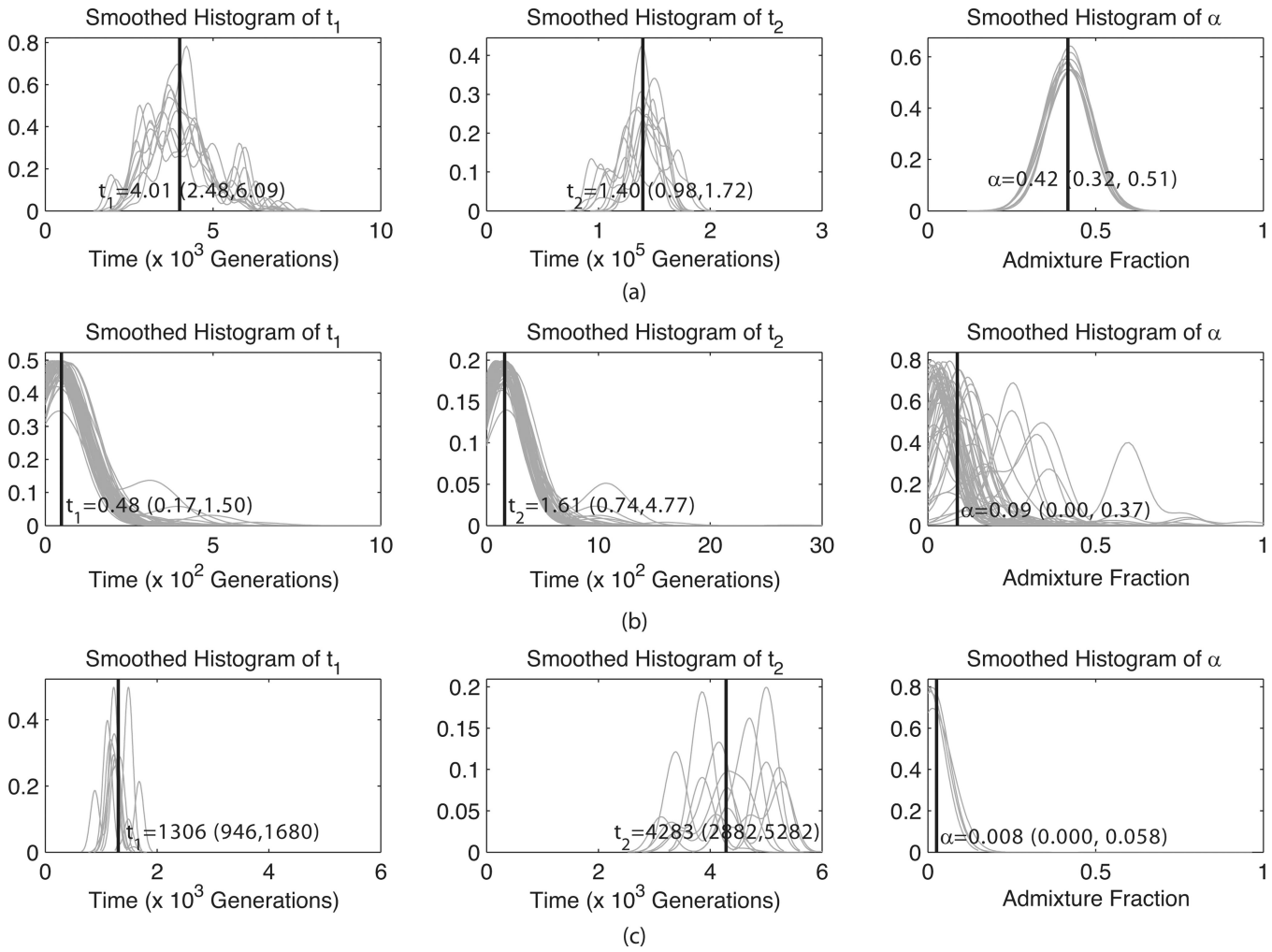

Fig. 5.

Probability density of the estimated parameter values, t_1, t_2, and α (left to right) for the bovine, HapMap, and 1,000 Genomes datasets. The dark vertical lines represent the means of the parameter values. The 95% confidence intervals are shown in parentheses. (a) 10 MCMC chains run on 76 cattle from the Bovine HapMap dataset on each of the 10 independent runs [24]. (b) 50 MCMC chains run on 1092 individuals from the 1,000 Genomes dataset [25]. (c) 10 MCMC chains run on 210 individuals from HapMap Phase II dataset [26].

TABLE 1

The three quartiles (25%,50%,75%) of the relative difference between estimated and true parameter values for 135 simulated data sets. t_1 and t_2 are in units of generations.

	$\frac{\lvert\hat{t}_1 - t_1\rvert}{t_1}$	$\frac{\lvert\hat{t}_2 - t_2\rvert}{t_2}$	$\frac{\lvert\hat{\alpha} - \alpha\rvert}{\alpha}$
2.0×10^5			
CLEAX-30	[2.200 4.535 12.819]	[2.077 5.584 8.922]	[0.223 0.441 1.272]
MEAdmix	[0.317 0.512 0.666]	[0.226 0.479 0.698]	[0.290 0.470 1.337]
3.5×10^6			
CLEAX-10	[0.082 0.216 0.397]	[0.069 0.193 0.420]	[0.078 0.168 0.523]
CLEAX-30	[0.087 0.179 0.289]	[0.068 0.125 0.335]	[0.071 0.156 0.267]
CLEAX-100	[0.079 0.165 0.254]	[0.063 0.121 0.321]	[0.062 0.153 0.264]
MEAdmix	[0.114 0.356 0.592]	[0.069 0.127 0.329]	[0.069 0.165 0.299]
3.5×10^7			
CLEAX-30	[0.061 0.116 0.199]	[0.064 0.124 0.268]	[0.062 0.146 0.248]