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

Ming-Chi Tsai
Joint CMU-Pitt Ph.D. Program
in Computational Biology
Carnegie Mellon University
Pittsburgh, PA, USA
mingchitt@andrew.cmu.edu

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

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

Russell Schwartz
Department of Biological
Science
Carnegie Mellon University
Pittsburgh, PA, USA
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.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms (including Monte Carlo)

General Terms

Algorithms

ACM-BCB'12 October 7-10, 2012, Orlando, FL, USA
Copyright 2012 ACM 978-1-4503-1670-5/12/10 ...$15.00.

Keywords

Population genetics, Metropolis-Hasting, coalescent, population history

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 [21]. Reconstructing historical admixture scenarios also has important practical value in biomedical contexts. For instance, learning the correct time scale when the HIV viruses diverge would be useful for understanding the circumstances surrounding the emergence of the acquired immune deficiency syndrome (AIDS) pandemic and the rate at which HIV-1 diverges [14]. In statistical genetics, studying admixture and population structure can help in identifying and correcting for confounding effects of population structure in disease association tests [8]. Studying admixture can also help in understanding the acquisition of disease-resistance alleles [6].

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 [3, 7]. While such methods provide a powerful tool for visualizing fine substructure and admixture, however, it typically requires 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 [23, 22, 24], 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 [20], and others on different parameters, such as effective population size [17]. We, however, focus here on the more standard three-population scenario and the joint inference of both the admixture proportion and the time of admixture and divergence. 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 [4]. 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 ancient so mutations novel to the admixed populations are no longer negligible. To deal with this issue, several coalescent-based methods have been proposed. *MEAdmix* [26], 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. To address these issues, we have created a novel two-step inference model called Consensus-tree based Likelihood Estimation for AdmiXture (*CLEAX*). Rather than inferring the parameters directly from the molecular data [26, 20, 5], we first learn a set of summary descriptions of the overall population history from the molecular data. 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. 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 a divergence time and admixture fraction for a dataset, our approach first tries to determine a number of subpopulations, potential evolutionary models between the subpopulations, and a summary description that approximates the number of segregating sites (or mutations) that are believed to have occurred after each subpopulation separated from its parental population but before it further divides into additional 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 [1, 26, 5, 4], 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 resulted 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 includes 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). The effective population size, N, and mutation rate, μ will be aggregated with the length of the sequences, l, as a single parameter θ. Under this assumption, the free parameters we must learn are t_1, t_2, α, and θ. Our method is designed to be adaptable for use a subroutine to analyzing multiple admixture events in the context of a more complete reconstruction of population history of a species, although such extensions are not considered in the present work.

Given the admixture model, we would expect different regions of the genome to have genealogies corresponding to the different possible subsamples of the network of ancestral relationships shown in the figure. For example, at some regions of the genome, we would expect to see a genealogy of the three samples derived from Fig. 1(b)(top) while other regions would have genealogies derived from Fig.

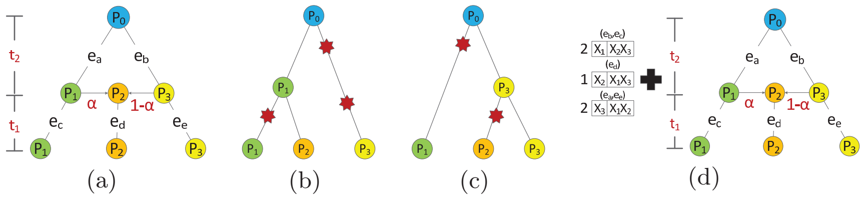

Figure 1: Example of a history of two parental populations (P_1 and P_3) and an admixed population (P_2). Ancestral population P_0 diverged at t_2 to form P_1 and P_3, followed by an admixture event at t_1 to form P_2. (a) The admixture model of the example. (b) Possible history of the example at some 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 divergence and admixture events, admixture fractions, and edge lengths.

1(c)(bottom). If we suppose $\alpha = 0.5$ then we should see these two genealogies with approximately equal frequency.

Given the sequence data derived from admixture scenario, our approach will first learn that there are three subpopulations in the example dataset using an algorithm developed in our previous work [25] for the problem of reconstructing population histories, describing the historical emergence of population subgroups in a broader population, from non-admixed data. At the same time, that prior algorithm will learn the potential evolutionary model shown in Fig. 1(d). The algorithm will also learn a summary description that suggests 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 (branch 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 (branch 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 [25] is conceptually based on the idea of consensus trees [19], which represent inferences as to the robust feature of a family of trees. 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 identifies a robust set of model bipartitions, $B^M = \{b_1^M, b_2^M, ...b_k^M\}$, roughly representing the edges in the consensus tree, and a set of weight values, $W = \{w_0, w_1, w_2, ..., w_k\}$, associated with the model bipartitions. The weights are computed by counting the number of observed bipartitions assigned to a given edge for a minimum description length [9] model of the full population history. Under the assumption that each observed bipartition represents a single mutation that occurred since the sampled individuals diverged from a common ancestor, weights would approximate the 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 in the data. Second, the algorithm should output an inferred evolutionary model, shown in Figure 1(d) and characterized by the model bipartition set $\{b_1^M = P_1|P_2P_3, b_2^M = P_2|P_1P_3, b_3^M = P_3|P_1P_2\}$. Finally, the algorithm should also 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 vs. 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: To make inferences about the parameter set Θ, we will estimate the distribution of a posterior probability of the parameters given the observed weights W associated with branches in the tree. We first 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 length of 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. For each bipartition, if $f(b)$ is a function that computes the optimal index assignment of a bipartition b to the model bipartition set and 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 $\mathcal{G} = \{G_1, G_2, ..., G_n\}$ 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:

$$\begin{aligned} P(W|\Theta) &= \prod_{i=0}^{3} \int_{l=0}^{\infty} \sum_{\mathcal{G}} P(w_i|\Theta, l_{b_i^M}) P(l_{b_i^M}|\mathcal{G}, \Theta) P(\mathcal{G}|\Theta) dl \\ &= \prod_{i=0}^{3} \sum_{\mathcal{G}} P(w_i|l_{b_i^M}, \theta) P(l_{b_i^M}|\mathcal{G}) P(\mathcal{G}|\Theta) \end{aligned}$$

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

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 [5] and [20] 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.

The second approximation that we incorporate into the model is that the total number of distinct genealogies from which lineages evolve (m) is 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, but empirical evidence given in the Results suggests that while specifying $m << n$ independent genealogies allows for a possibility of error, the actual increased error in practice is modest. 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 $m << n$.

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

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

MCMC Sampling: To estimate the posterior probability distribution, we employ the Metropolis-Hastings algorithm. We defined 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 G can only contain of 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:

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

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 when variance is large when difference between the expected and observed number of mutations is large and better when variance is small when the difference between 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

$$\begin{aligned} Q(X_n|X_o) \quad = \quad & P\left(x^n_{t_1}|\mu^o_{t_1}, \sigma_{t_1}\right) P\left(x^n_{t_2}|\mu^o_{t_2}, \sigma_{t_2}\right) P\left(x^n_\alpha|\mu^o_\alpha, \sigma_\alpha\right) \\ & \times P\left(\mathcal{G}|x^n_{t_1}, x^n_{t_2}, x^n_\alpha\right) \end{aligned}$$

yielding a Metropolis-Hastings acceptance ratio r of:

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

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 [12] 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 [11, 18]. 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* [26]. *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 [5, 20, 26], 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 [20, 1], 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 135 simulated datasets and computed the average difference between the true and estimated parameter values for each parameter, ($|\hat{\Theta} - \Theta|$). 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 α=$\{0.05, 0.2, 0.6\}$ 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 difference between the estimated parameters and true parameters. Each MCMC run used 1,000 iterations of burn-in followed by 20,000 MCMC steps.

Real SNP Data: We further evaluated our method by applying it to a bovine SNP dataset [2], chosen due to the lack of publicly available 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-eights 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 for which no appreciable admixture is known to occur. We used the Phase II HapMap data set (phased, release 22) [13] 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 at learning 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 the 50,556 SNPs collected from chromosome 22.

For both datasets, we set the number of genealogies m to be 30 for these tests. We did not evaluate the bovine dataset using *MEAdmix*, as the number of segregating sites on the real dataset exceeded the software's limitations. Like the simulated dataset, we used 1,000 steps in the burn-in period followed by 20,000 MCMC steps.

4. RESULTS

Coalescent Simulated Data: Figure 2 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 increases 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. A similar detailed analysis of accuracy across parameters variations was conducted for t_1 and t_2 but is omitted in the present work due to space constraints.

Aggregate quantitative performance is shown in Table 1(a), which provides the average difference between the estimated parameters and true parameters computed by each algorithm for each group of simulations, ($|\hat{\Theta} - \Theta|$). For datasets with 3.5×10^6-base sequences, *CLEAX* has a worse average difference between estimated and true α when we set the number of genealogies to be 10, but roughly the same average estimated t_1 and t_2 as *MEAdmix.* When we increase the number of genealogies to be 30 or more, *CLEAX* yielded 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 was unable to automatically identify the three major population groups, instead identifying only the divergence into subpopulations P_1 and P_3. We attribute this 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(a),

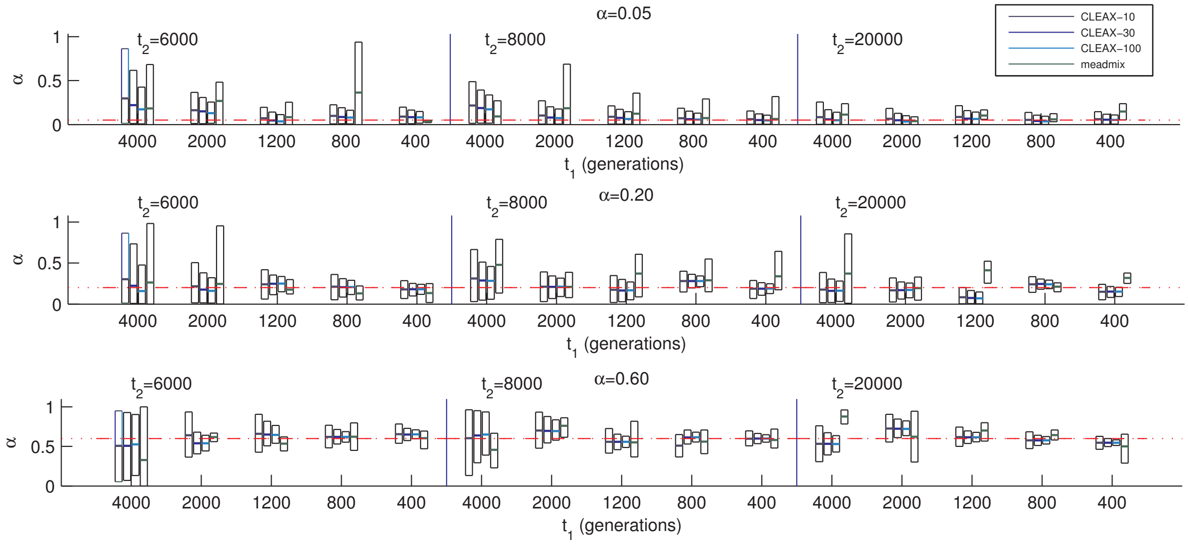

Figure 2: Mean and 95% confidence interval of the estimated α on 3.5×10^6-base sequences. The different shades of blue represents the mean estimated by *CLEAX* using 10, 30, and 100 genealogies (left).

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(a) shows, accuracy of the three estimated parameter is improved relative to the smaller datasets, with roughly 10% to 30% improvements.

We also examined average compute 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.

Real SNP Data: Table 1(b) shows parameter mean and 95% confidence interval estimates on both real datasets. 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 33.2 percent and 50 percent. Mean estimation of divergence time (t_2) is about 29,000 generations. Assuming 7 years per generation for cattle, the divergence time would translate to approximately 200 kya, consistent with the belief that the *indicine* and *taurine* diverged approximately 250 kya [2]. Admixture time (t_1) is estimated to be approximately 6 kya, likely an overestimate of the true value, but ranges between 3.6 kya to 8.0 kya. Mean estimation of $\theta = l \times N \times \mu$ is 36.1. If we assume the effective population size is 2000 based on ancestral effective population size [2] 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 [15, 18]. Using an estimated effective population size of 107 [2], a more consistent estimate of effective population size after a recent population bottleneck derived by averaging the recent effective population size of the three breeds, would yield a more realistic mutation rate of 2.8×10^{-9} [18]. Inaccuracy in the rate might also be due to ascertainment bias or the incomplete detection of the mutations at the sequencing phase.

For the HapMap Phase II data, *CLEAX* estimated α to be less than 1% with a 1% confidence interval. 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 76.5 kya and 89.6 kya. The divergence time (t_1) between Europeans (CEU) and East Asians (CHB+JPT) has a mean estimate of 29.6 kya and a confidence interval between 23.0 kya and 33.6 kya. Mean estimation of θ is $4,320$. Assume the effective population size of human population to be 10,000 [10], the mutation rate would be 2.16×10^{-9} per site per generation, similar to prior estimates [15, 18].

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

Table 1: (a)Means and standard deviations of difference between estimated and true parameter values for 135 simulated data sets. (b) Means and 95% confidence intervals of the estimated parameters from bovine and HapMap datasets. t_1 and t_2 are in units of generations.

(a) Simulated Data

	3.5×10^7		
	$\lvert\hat{\alpha} - \alpha\rvert$	$\lvert\hat{t}_1 - t_1\rvert$	$\lvert\hat{t}_2 - t_2\rvert$
CLEAX, $m = 30$	0.0386 ± 0.03217	230 ± 262	2068 ± 1763
	3.5×10^6		
	$\lvert\hat{\alpha} - \alpha\rvert$	$\lvert\hat{t}_1 - t_1\rvert$	$\lvert\hat{t}_2 - t_2\rvert$
CLEAX, $m = 10$	0.0523 ± 0.0486	483 ± 607	2832 ± 2772
CLEAX, $m = 30$	0.0436 ± 0.0405	365 ± 493	2494 ± 2523
CLEAX, $m = 100$	0.0420 ± 0.0355	328 ± 428	2264 ± 2279
MEAdmix	0.0448 ± 0.0469	485 ± 384	2883 ± 4373
	2.0×10^5		
	$\lvert\hat{\alpha} - \alpha\rvert$	$\lvert\hat{t}_1 - t_1\rvert$	$\lvert\hat{t}_2 - t_2\rvert$
CLEAX, $m = 30$	0.1205 ± 0.0984	$1.22 \pm 1.45 \times 10^4$	$5.98^4 \pm 4.59 \times 10^4$
MEAdmix	0.1341 ± 0.1077	1033 ± 1064	4226 ± 4168

(b) Real Data

	Bovine	HapMap
t_1	841 $(0.559, 1.16) \times 10^3$	1.48×10^3 $(1.15, 1.68)\ \times 10^3$
t_2	2.90×10^4 $(2.43, 3.39) \times 10^4$	4.02×10^3 $(3.83, 4.48) \times 10^3$
α	0.416 $(0.332, 0.500)$	0.007 $(0.007, 0.007)$
θ	36.1 $(33.3, 39.3)$	4.32×10^3 $(4.22, 4.38) \times 10^3$

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(a) 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 providing increased confidence on datasets comparable in size to human chromosomes.

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 [2, 27]. Using the HapMap Phase II dataset, our method's estimation of the YRI-CEU divergence time between 76.5 kya to 89.6 kya was consistent with the STR estimation by [27] (62-133kya) and the HMM estimation by [17] (60-120 kya). Estimation of little or no admixture fraction between the CHB+JPT and CEU was also consistent with the general belief that no admixture occurred between the major human populations. Similarly, using the bovine dataset, estimates of divergence time and admixture fraction were also consistent with the general consensus [2]. One discrepancy in the bovine dataset was an unrealistically high (6,000 year) estimate of admixture time. 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 to a few hundred cattle in recent years [2, 16], 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 may suggest that our method in current form may be a poor choice if one is interested in estimating admixture times on data with significant changes in effective population size over time. However, estimation of admixture fractions remains relatively accurate despite the 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.

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 [5, 20], 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, this unified approach is designed to extend to arbitrarily complicated admixture scenarios by sampling over reticulate histories potentially containing multiple admixture events. While more complex admixture models with additional number of populations were not addressed, the analysis here nonetheless provides a proof of concept for more general admixture models that will be explored in future work.

ACKNOWLEDGEMENT

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.

6. REFERENCES

[1] G. Bertorelle and L. Excoffier. Inferring admixture proportions from molecular data. *Molecular Biology and Evolution*, 15(10):1298–1311, 1998.

[2] T. Bovine HapMap Consortium. Genome-wide survey of snp variation uncovers the genetic structure of cattle breeds. *Science*, 324(5926):528–532, 2009.

[3] K. Bryc, A. Auton, M. R. Nelson, J. R. Oksenberg, S. L. Hauser, S. Williams, A. Froment, J.-M. Bodo, C. Wambebe, S. A. Tishkoff, and C. D. Bustamante. Genome-wide patterns of population structure and admixture in west africans and african americans. *Proceedings of the National Academy of Sciences*, 107(2):786–791, 2010.

[4] R. Chakraborty. Gene admixture in human populations: Models and predictions. *American Journal of Physical Anthropology*, 29(S7):1–43, 1986.

[5] L. Chikhi, M. Bruford, and M. Beaumont. Estimation of admixture proportions: A likelihood-based approach using markov chain monte carlo. *Genetics*, 158(3):1347–1362, 2001.

[6] I. Dupanloup, G. Bertorelle, L. Chikhi, and G. Barbujani. Estimating the impact of prehistoric admixture on the genome of europeans. *Molecular Biology and Evolution*, 21(7):1361–1372, 2004.

[7] O. François, M. Currat, N. Ray, E. Han, L. Excoffier, and J. Novembre. Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, 27(6):1257–1268, 2010.

[8] D. B. Goldstein and L. Chikhi. Human migrations and population structure: What we know and why it matters. *Annual Review of Genomics and Human Genetics*, 3(1):129–152, 2002.

[9] P. Grünwald, I. Myung, and M. Pitt. *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2005.

[10] M. F. Hammer. A recent common ancestry for human y chromosomes. *Nature*, 378(6555):376–378, 1995.

[11] R. C. Hardison, K. M. Roskin, S. Yang, M. Diekhans, W. J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, and et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research*, 13(1):13–26, 2003.

[12] R. Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44, 1990.

[13] International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, October 2007.

[14] B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the ancestor of the hiv-1 pandemic strains. *Science*, 288(5472):1789–1796, 2000.

[15] S. Kumar and S. Subramanian. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences*, 99(2):803–808, 2002.

[16] S. H. Lee, Y. M. Cho, D. Lim, H. C. Kim, B. H. Choi, H. S. Park, O. H. Kim, S. Kim, T. H. Kim, D. Yoon, and S. K. Hong. Linkage disequilibrium and effective population size in hanwoo korean cattle. *Asian-Australasian Journal of Animal Sciences*, 24(12):1660–1665, 2011.

[17] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

[18] G. Liu, L. Matukumalli, T. Sonstegard, L. Shade, and C. Van Tassell. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. *BMC Genomics*, 7(1):140, 2006.

[19] M. Nei and S. Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, 2000.

[20] R. Nielsen and J. Wakeley. Distinguishing migration from isolation: A markov chain monte carlo approach. *Genetics*, 158(2):885–896, 2001.

[21] E. J. Parra, A. Marcini, J. Akey, J. Martinson, M. A. Batzer, R. Cooper, T. Forrester, D. B. Allison, R. Deka, R. E. Ferrell, and M. D. Shriver. Estimating african american admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, 63(6):1839–1851, 1998.

[22] A. L. Price, A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels, I. Ruczinski, T. H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

[23] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[24] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *American journal of human genetics*, 82(2):290–303, 2008.

[25] M.-C. Tsai, G. E. Blelloch, R. Ravi, and R. Schwartz. A consensus tree approach for reconstructing human evolutionary history and detecting population substructure. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 8:918–928, July 2011.

[26] J. Wang. A coalescent-based estimator of admixture from dna sequences. *Genetics*, 173(3):1679–1692, 2006.

[27] L. A. Zhivotovsky. Estimating divergence time with the use of microsatellite genetic distances: Impacts of population growth and gene flow. *Molecular Biology and Evolution*, 18(5):700–709, 2001.